
**Working with Collinearity in Epidemiology:
Development of Collinearity Diagnostics,
Identifying Latent Constructs in Exploratory
Research and Dealing with Perfectly Collinear
Variables in Regression**

Andrew Stephen Woolston

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Medicine

June, 2012

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

1. Woolston, A., Tu, Y. K., Baxter, P. D., & Gilthorpe, M. S. (2012). A Comparison of Different Approaches to Unravel the Latent Structure within Metabolic Syndrome. *PLoS One*, 7(4), e34410. doi:10.1371/journal.pone.0034410.

Conceived and designed the experiments: AW PDB. Performed the experiments: AW. Analysed the data: AW. Contributed reagents/materials/analysis tools: AW. Draft the paper: AW. Revision of paper: YKT PDB MSG.

2. Tu, Y. K., Woolston, A., Baxter, P. D., & Gilthorpe, M. S. (2010). Assessing the impact of body size in childhood and adolescence on blood pressure: an application of partial least squares regression. *Epidemiology*, 21(4), 440-448. doi:10.1097/EDE.0b013e3181d62123

Conceived and designed the experiments: YKT AW. Performed the experiments: YKT AW. Analysed the data: YKT AW. Contributed reagents/materials/analysis tools: YKT. Draft the paper: YKT. Revision of paper: AW PDB MSG.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Acknowledgements

Foremost, I would like to thank Professor Mark Gilthorpe, Dr. Yu-Kang Tu and Dr. Paul Baxter for their excellent advice and supervision throughout my four years of research. They have always given me great belief that I would be able to complete this work, even when I have doubted it myself. The time spent with me discussing ideas has been invaluable and is very much appreciated.

Thanks to all of my colleagues, past and present, in the biostatistics division at the University of Leeds. They have all made time for me and given me a great insight into academic life. I have been given opportunities to work on projects outside of my PhD research which has given me a taste of the field and made me even more determined to seek a future in academic research

I would like to thank my girlfriend Jia-Ying for her amazing support and patience in the most difficult years. This work would have been a much harder task without her. Thanks also to my mother, father, brother and grandmother. They have never doubted that I would be able to complete my work and have always been there for me when I have needed them most. I would like to thank my close friends, Andy, Anneka, Clare, Joe, Jennie, Nicola and Nikki. I very much appreciate the confidence and support they have shown throughout my time in Leeds in achieving this goal.

This research would not have been possible without the funding of the medical research council. It has given me the means to visit conferences, courses and seminars around the world in places such as Belfast, Paris and Taipei. I am very grateful for these opportunities that I have been given and I hope that my finished work reflects the hard work put in over these years. This has been a great experience and I very much appreciate the help and guidance I have received from all of the people around me.

Abstract

Collinearity plays an integral role in regression studies involving epidemiological data. Variables often form part of a common biological mechanism or measure the same element of a latent structure. It is a natural feature of most data and as such it is rarely possible to physically control for collinearity in data collection. A focus is placed on the analytical assessment of the data. Departures from independence can severely distort the interpretation of a model and the role of each covariate. This leads to increased inaccuracy as expressed through the regression coefficients and increased uncertainty as expressed through coefficient standard errors. Such a feature has the potential to impact on the clinical conclusions formed from regression studies.

The work in this thesis first considers an assessment of the impact of collinearity on model parameters and the conclusions formed. A new collinearity index is developed which incorporates the role of the response in moderating the impact of collinearity. The idea for the new index is developed using vector geometry and extended to a general measure. The work in collinearity is later extended to consider the formation of a dependency structure from a collection of collinear variables. A novel methodology, labelled the matroid approach, is coded and implemented on a metabolic syndrome dataset to extract a latent structure that could represent this clinical construct. Comparisons are subsequently made to existing exploratory factor analysis and clustering methods in the literature. Finally, the unique problem of perfect collinearity is considered in a lifecourse and age-period-cohort setting. The justification of constraint and non-constraint regression methods is considered in an attempt to provide 'solutions' to the identification problem generated by collinearity.

Table of Contents

1.	Introduction.....	1
1.1	Background and Motivation	1
1.2	Aims of the thesis	4
1.3	Thesis Structure	5
1.4	Publications	7
1.5	Notation.....	8
1.6	Statistical Software	9
2.	Regression Analysis in Applied Clinical and Epidemiological Research	10
2.1	The Origins of Regression in Application.....	11
2.2	The Regression Model	15
2.2.1	Applied Regression Analysis	16
2.2.2	The Least Squares Estimator	17
2.2.3	Quality of an Estimator.....	19
2.3	Inference in Epidemiology	21
2.3.1	Study Design	22
2.3.2	Causation and Causal Inference	24
2.3.3	Modelling Strategy	29
2.4	Vector Geometry of Linear Regression.....	31
2.4.1	Variable Space and Subject Space	32
2.4.2	Converting Analytic Ideas to Geometry	34
2.4.3	The Least Squares Estimate.....	36
2.5	Conclusions.....	38
3.	Measuring the Impact of Collinearity	39
3.1	What is Collinearity?.....	40
3.2	Collinearity in Epidemiology	42
3.2.1	Causal Modelling	43
3.2.2	Understanding Third Variable Effects.....	47
3.2.3	Development of a Causal Hypothesis.....	52

3.3	Collinearity Diagnosis	54
3.3.1	Developing a Collinearity Index.....	55
3.3.2	Approaches to Diagnosing Collinearity	55
3.3.3	Application of Collinearity Indices	61
3.3.4	The Impact of Collinearity on Variance	66
3.3.5	Remedies of Collinearity.....	68
3.4	Investigating the Role of the Response	69
3.5	Conclusions.....	77
4.	Vector Geometry of Exploratory Analysis and Advanced Regression Methods ...	78
4.1	Geometry of Multivariable Regression.....	79
4.1.1	Incorporating Error into Vector Geometry.....	79
4.1.2	The Error Space	80
4.1.3	Least Squares Regression in Higher Dimensions	81
4.2	Exploratory Analysis	83
4.2.1	Principal Component Analysis	83
4.2.2	Exploratory Factor Analysis	86
4.2.3	The Role of Confirmatory Factor Analysis	87
4.3	Cluster Analysis.....	88
4.3.1	Hierarchical Clustering.....	88
4.3.2	Visualizing the Results	89
4.3.3	Variable Clustering	90
4.4	Shrinkage Regression.....	91
4.4.1	Principal Component Regression	92
4.4.2	Partial Least Squares Regression	94
4.5	Conclusions.....	97
5.	Developing an Index to Measure the Impact of Collinearity	98
5.1	Example Application	99
5.2	Potential Applications of the Index	100
5.3	Incorporating the Response into Existing Diagnostics	101
5.4	Assessment of Deviation	104
5.4.1	Calculating the Index	107
5.4.2	Interpretation of the Index.....	108
5.4.3	Comparison to a Variance Based Index.....	109

5.5	Deconstructing the Impact of Collinearity.....	112
5.5.1	Simulation Study.....	113
5.5.2	Extending the Bivariate Index.....	116
5.6	Measurement Error and Sampling Variation.....	120
5.6.1	The Construction of a Confidence Interval.....	121
5.6.2	Confidence Interval of the D-Index	124
5.7	Diagnosis in Higher Dimensions	126
5.7.1	Extending the Multivariate Index	130
5.8	Regression Study: Body Fat Analysis	131
5.9	Discussion	134
5.10	Conclusions.....	135
6.	Approaches to Unravel the Latent Structure within Metabolic Syndrome	136
6.1	Example Application	137
6.2	Defining Metabolic Syndrome	138
6.3	Factor Analysis and Clustering Procedures	139
6.4	PCA vs. EFA in Applied Research	141
6.4.1	Decision Making in an EFA Study.....	142
6.4.2	Limitations of Factor Analysis.....	147
6.5	The Relationship of Factor Analysis and Metabolic Syndrome	149
6.5.1	Factor Analysis and MetS	150
6.5.2	Why is Factor Analysis Chosen for MetS?	153
6.6	Cluster Analysis.....	155
6.6.1	The VARCLUS Procedure	156
6.6.2	Application of VARCLUS.....	158
6.7	Variable Clustering using Matroids	163
6.7.1	Matroids to Identify Latent Structures.....	164
6.7.2	Matroid Analysis in MetS Research.....	170
6.8	Identifying Consistency across MetS Studies	171
6.9	Conclusions.....	183
7.	Identifying the Critical Phase of Growth in the Lifecourse	184
7.1	Example Application	185
7.2	The Foetal and Developmental Origins of Disease.....	186
7.3	Adjustment for Lifecourse Variables	186

7.4	The identification Problem	189
7.4.1	SVD of the Ill-Defined Matrix.....	189
7.4.2	Vector Geometry of the Lifecourse	190
7.4.3	Should we Standardize?	191
7.5	Tracing the Lifecourse	192
7.6	Constraint Solutions of the Ill Defined Matrix.....	195
7.6.1	Linear Constraints on the Estimation	196
7.6.2	The Moore-Penrose Generalized Inverse.....	198
7.7	The Application of Shrinkage Methods to the Lifecourse	201
7.7.1	Principal Component Regression	201
7.7.2	Partial Least Squares Regression	206
7.8	Identifying the Critical Phase of Growth	210
7.9	Discussion	217
7.10	Conclusions.....	219
8.	Perfect Collinearity in the Age-Period-Cohort Model	220
8.1	Definition of the Variables and APC Data.....	221
8.2	Example Application	223
8.3	Graphical Analysis.....	224
8.4	Regression Analysis.....	227
8.4.1	Geometry of the Solutions	231
8.4.2	The Impact of a Linear Constraint	234
8.5	Selected Solutions from the Literature	239
8.5.1	Curvature & Drift	239
8.5.2	Minimal Differences	242
8.5.3	Individual Records & Natural Weights	243
8.6	Latent Variable Methods in APC Analysis.....	247
8.6.1	Justification for the Application of Latent Variable Methods	247
8.6.2	Application of Latent Variable Methods	249
8.7	Regression Study	251
8.8	Conclusions.....	260
9.	Conclusions and Future Developments.....	261
	Abbreviations.....	265
	References	266

List of Figures

Figure 2.1: A plot of the mean diameters of Galton’s sweet pea plants.....	11
Figure 2.2: Average heights when grouped on (a) parental and (b) adult offspring.	13
Figure 2.3: Galton’s distribution of heritage.	14
Figure 2.4: An example Hierarchy of Evidence.....	22
Figure 2.5: Path diagram representation of a multiple regression model.	24
Figure 2.6: DAG demonstrating a causal relation between BMI and CVD.	26
Figure 2.7: SMK as a confounder to the relationship between BMI and CVD.	26
Figure 2.8: Potential confounders brought about by the inclusion of SMK.....	27
Figure 2.9: SMK considered as a competing exposure to BMI.	28
Figure 2.10: Hypothetical presence of a preceding unobserved variable.....	28
Figure 2.11: DAG demonstrating a sample correlation between two exposures.	30
Figure 2.12: Observations (p_1, p_2) in (a) variable and (b) subject space.	32
Figure 2.13: Simple linear regression in (a) variable space and (b) subject space.....	33
Figure 2.14: An illustration of the power of multivariate geometry.	33
Figure 2.15: An example of axis rotation.....	34
Figure 2.16: Orthogonal OLS projection of y onto the regression space spanned by X	35
Figure 3.1: Illustrative effects of collinearity on a linear regression model.	40
Figure 3.2: (a) A true causal link and (b) an association generated by confounding.	44
Figure 3.3: A causal pie illustration of MetS based on the IDF definition.	46
Figure 3.4: Third variable effects in which Z is (a) a mediator and (b) a confounder.	47
Figure 3.5: Statistical interpretation of (a) a confounding and (b) suppression effect.	50
Figure 3.6: An example path diagram.....	51
Figure 3.7: Classical birth Control example from Pearl (2000).	52
Figure 3.8: Presence of an additional mediator to the example.	53
Figure 3.9: An illustration of the construction of R_{x_j} using vector geometry.	56
Figure 3.10: Trend of the VIF function in the bivariable regression case.	57
Figure 3.11: Illustration of the SVD decomposition of the VC matrix.....	59
Figure 3.12: Results of a simulation from the theoretical correlation matrix.....	62
Figure 3.13: The effects of external factors on the impact of collinearity.	66

Figure 3.14: Models investigated in (a) simulation 3.1 and (b) simulation 3.2.	69
Figure 3.15: Parameter values from Simulation 3.1.	70
Figure 3.16: Parameter values from Simulation 3.2.	72
Figure 3.17: Effect of an unequal correlation between covariate and response	73
Figure 3.18: Change in b_1 and b_2 with a rotation of the response ($R_y^2 = 1$).	74
Figure 3.19: Simulation 3.4 – Varying R_y^2 on unequal correlations.	75
Figure 3.20: Results of simulation 3.4 for (a) $r_{12} = 0.9$ and (b) $r_{12} = 0.3$	75
Figure 4.1: An illustration of error in vector geometry.	79
Figure 4.2: An example of a trivariable regression model.....	81
Figure 4.3: Construction of the projected response in 2D regression space.	82
Figure 4.4: PCA analysis in subject space ($r_{12}=0.71$).	85
Figure 4.5: The decomposition of predictors into common and shared variation.....	86
Figure 4.6: Geometrical Illustration of a 3-dimensional factor space.	87
Figure 4.7: (a) Hierarchical and (b) k -means clustering on the $X^T X$ in section 3.3.3.....	88
Figure 4.8: (a) PCR on two covariates with the first PC retained and (b) on three covariates demonstrating the y_{PCR}^m projections onto regression surfaces.....	94
Figure 4.9: Graphical presentation of the NIPALS PLS algorithm.	96
Figure 4.10: (a) The covariance maximizing procedure of PLS on the first component of a two predictor model and (b) the PLS axes of a three predictor model.	97
Figure 5.1: Geometrical projection of (a) X onto y and (b) y onto X	101
Figure 5.2: Covariance maximizing procedure for two covariates.	103
Figure 5.3: The projection of y to generate univariable and multivariable estimates.....	104
Figure 5.4: Placing uncorrelated univariable estimates onto collinear axes.....	104
Figure 5.5: Deviation of point estimates in different collinearity conditions.....	105
Figure 5.6: Change in D_2 under varying collinearity conditions.	106
Figure 5.7: Computation of the D-index.	107
Figure 5.8: D_{22} and VIF for two pairs of predictors in the body fat data.	109
Figure 5.9: The role of the response in the impact of collinearity.	110
Figure 5.10: Vector Geometry of the Weight + Abdomen model.	112
Figure 5.11: A high vs. low association between response and predictors.....	112
Figure 5.12: Simulation 5.1 - (a) Change in point estimate and (b) variance for b_1	113
Figure 5.13: Simulation 5.2 – Variance, point estimates and index results.	114
Figure 5.14: An alternative Illustration of the computation of α_2 and β_2	116
Figure 5.15: Components of the correlation between the D and x_1	117

Figure 5.16: Correlations between predictor and D_2 for simulations 5.2.....	118
Figure 5.17: r_{D_2} (a) minimized and (b) maximized for simulation 5.2.	118
Figure 5.18: Equal correlations between covariate and D under differing r_{12}	119
Figure 5.19: Confidence intervals of orthogonal predictors.....	121
Figure 5.20: Confidence intervals for two sets of highly correlated predictors.	122
Figure 5.21: Demonstration of the construction of the confidence region.	122
Figure 5.22: An illustration of the construction of the ellipsoid confidence region.....	123
Figure 5.23: (a) Change of sign regions and (b) uncertainty region for a 'guesstimate'. ...	124
Figure 5.24: The translation of the confidence region from the bivariable models.	125
Figure 5.25: D_2 from the model including biceps, hip and abdomen.	126
Figure 5.26: Illustration of the extension to D_2 with x_3 included.	127
Figure 5.27: Orthogonal projection of D_3 onto x_1	131
Figure 6.1: Model assumed by applying (a) EFA and (b) PCA.....	141
Figure 6.2: (a) Scree plot and bi-plot's of (b) raw, (c) varimax, (d) promax solutions.....	145
Figure 6.3: Cluster analysis using Pearson r^2 with (a) single and (b) average linkage.....	155
Figure 6.4: A dendrogram summary of the VARCLUS procedure with $\text{maxeigen}=1$	159
Figure 6.5: Dendrogram of the cluster components for the second VARCLUS analysis....	161
Figure 6.6: Plot of PCIns, Ins and PCGlu, indicating the smallest eigenvalue.	165
Figure 6.7: An example of an LHD depiction.	166
Figure 6.8: Problematic situation using the eigenvalue selection criteria.	167
Figure 6.9: Matroid analysis of the MetS data using an inverse VIF criterion.	169
Figure 6.10: Scree plots of (a) the study population and (b) the untreated subgroup.	172
Figure 6.11: Scree plot for the eigenvalues of the factors in each group.	174
Figure 6.12: Matroid analysis using minimum VIF criteria for the complete sample.....	177
Figure 6.13: Matroid analysis of the female subset cohort.....	180
Figure 6.14: Potential factor structure from complete data	181
Figure 6.15: Potential factor structure from untreated sample.....	182
Figure 7.1: A path diagram of a hypothesized system in the lifecourse model.	186
Figure 7.2: Perfectly collinear standardized predictors in the Cebu study.	190
Figure 7.3: Vector Geometry of the Standardized Predictors.	191
Figure 7.4: Growth trajectory plot of (a) the raw data and (b) z-scores.	192
Figure 7.5: Illustration of the planes in which the three PC's must lie.	199
Figure 7.6: (a) Unstandardized and (b) standardized coefficients from PCR.....	203
Figure 7.7: Illustration of PCA for (a) a bivariable and (b) full model.	204

Figure 7.8: Illustration of PCA for the standardized predictors in the full model.....	205
Figure 7.9: Plot of the (a) unstandardized and (b) standardized coefficients from PLS....	208
Figure 7.10: PLS components from (a) standardized and (b) unstandardized data.	209
Figure 7.11: Plots of (a) unstandardized and (b) standardized coefficient estimates.....	211
Figure 8.1: Diagnosis rates of males as (a) a response surface and (b) contour plot.....	224
Figure 8.2: Age specific plots stratified by period and cohort.....	225
Figure 8.3: Path diagram of the sociological interpretation.....	227
Figure 8.4: Solution space geometry for the Cebu lifecourse example.....	231
Figure 8.5: Solutions planes for the linear constraints used in the Cebu example.	232
Figure 8.6: Point estimates from simulation with constraint $\beta_{A1} = \beta_{P1} = \beta_{C5} = 0$	235
Figure 8.7: Point estimates with constraints - (a) $\beta_{Ai} = \beta_{Pj} = \beta_{Ck} = 0$ and (b) $\beta_{A1} = \beta_{P3} = \beta_{C5} = 0$	236
Figure 8.8: The impact of sampling error when the correct constraint is employed.....	238
Figure 8.9: The Standard Structure of APC data.	243
Figure 8.10: A single cell of the Lexis diagram.....	244
Figure 8.11: Lexis table on a triangular mesh, with linear effect directions A, P, C.....	245
Figure 8.12: Cell weightings obtained directly from the lexis diagram.	245
Figure 8.13: Anti-log least squares curvature estimates for males in the study.	255
Figure 8.14: Anti-log least squares curvature estimates for females in the study.....	256
Figure 8.15: Coefficients from a full component PLS regression analysis.	258

List of Tables

Table 3.1: The VDP of the $X^T X$ matrix in fig.3.9.....	64
Table 3.2: The VDP of the counter-example used by Belsely (1991) in eqn(3.29).....	65
Table 5.1: Pearson correlations for the Penrose data.....	99
Table 5.2: D_2^2 in the lower triangle with VIF's in the upper triangle.....	109
Table 5.3: D_2 from the model including biceps, hip and abdomen.	126
Table 5.4: Results from the D-index for the four predictor body fat example.....	132
Table 6.1: Pearson correlations for the 10 metabolic risk factors.	137
Table 6.2: PAF extraction (Loadings $> .3$ are in bold to indicate significance).	151
Table 6.3: A Varimax rotated factor pattern for the PAF extraction.	151
Table 6.4: Promax factor solution (Loadings $> .3$ are in bold to indicate significance).	152
Table 6.5: Correlations between factors in the 'promax' solution.....	152
Table 6.6: A summary of the cluster components generated.	158
Table 6.7: R^2 measures demonstrating the quality of each cluster component.	158
Table 6.8: A cluster summary of variable loadings on each cluster component.....	159
Table 6.9: The inter-cluster correlations.	159
Table 6.10: Cluster components generated for the four cluster solution.....	160
Table 6.11: R^2 measures demonstrating the quality of each cluster component.	160
Table 6.12: Cluster loadings of the 4 cluster model.	160
Table 6.13: Inter-correlations of the cluster components.	161
Table 6.14: Female study population - coefficients for the entire population of the Ely study are displayed in regular typeface, coefficients in bold are for subgroup patients.....	171
Table 6.15: Factor loadings for females using the original methodological decisions.....	173
Table 6.16: PAF extraction with promax rotation (loadings $> .3$ highlighted in bold).	174
Table 6.17: Inter-factor correlations.	174
Table 6.18: A summary of the cluster components from a four cluster solution.	175
Table 6.19: Cluster structure with R^2 measures to indicate the stability of each cluster..	175
Table 6.20: Loadings on each cluster.....	176
Table 6.21: Inter-correlations of the clusters.	176

Table 6.22: Five factors - (a) cluster summary, (b) R^2 measures and (c) loadings.....	178
Table 6.23: Inter-correlations of the clusters from the five cluster solution.	179
Table 7.1: Pearson correlations of the BMI predictors and SBP in Metro Cebu.	185
Table 7.2: Unstandardized and standardized OLS coefficients of the Cebu data.....	187
Table 7.3: PCR results for one and two component models on the Metro Cebu data.....	202
Table 7.4: PLS results for one and two component models on the Metro Cebu data.	207
Table 7.5: Correlations of the Cebu data including BMI change measures.	210
Table 7.6: Coefficient estimates from two attainable models using OLS regression.	211
Table 7.7: Results of the PCR analysis with 1-6 components retained.....	213
Table 7.8: Results of the PLS analysis with 1-6 components retained.	214
Table 8.1: An example lexis diagram.	221
Table 8.2: Absolute numbers of cases in Sweden during 1983-2007 with cumulative incidence at 35 years in brackets for males (standard font) and females (bold font).	223
Table 8.3: Coefficients from an Age + Period regression of the male subjects.....	251
Table 8.4: Coefficients from an Age + Period regression of the female subjects.....	252
Table 8.5: Coefficients from an Age + Cohort regression of the male subjects.	252
Table 8.6: Coefficients from an Age + Cohort regression of the female subjects.	253
Table 8.7: Deviance and AIC statistics from nested APC models.	253
Table 8.8: Default constraints along with least squares and mean difference curvatures.	255
Table 8.9: Default constraints along with least squares and mean difference curvatures.	256
Table 8.10: PLS Coefficients from a maximum components model for Males.....	257
Table 8.11: PLS Coefficients from a maximum components model for Females.	257

1. Introduction

1.1 Background and Motivation

Covariates in epidemiological and clinical research are almost guaranteed to be collinear. The variables may form part of a common biological mechanism or measure the same element of a latent structure. Whilst many researchers accept collinearity as an intrinsic feature of the data, there is often a lack of understanding regarding the impact on statistical/clinical inference and the means to work with such dependencies. Whilst a great number of estimators have been developed since the 1960's, the least squares approach provides the baseline to understanding this impact. Entering highly collinear variables (although not perfectly correlated) will not directly violate the traditional assumptions of least squares regression and so the estimates will remain unbiased with minimum variance in their class. However, in a single sample case, the unbiased property will have limited use if the estimator is not precise. This provides the motivation to delete variables, combine covariates into a single index or employ alternative estimators. In practice, these decisions are often based on common features of collinearity such as insignificant p-values and unexpected changes of sign on the coefficient point estimates. Such "symptoms" are neither necessary nor sufficient for the presence of collinearity and so this remedial action may lack justification. Along with external biological knowledge, the use of collinearity diagnostics can be vital to understanding the potential impact of collinearity. However, an extension to the most popular diagnostics, such as the variance inflation factor or the condition index, can only provide part of the answer. A review of the current methodology demonstrates that the role of factors external to the covariate correlations (such as the response, sample size and sampling variation) in moderating the impact of collinearity is almost entirely disregarded in the literature.

Study design plays a crucial role in dictating the presence and form of collinearity amongst covariates. Whilst the focus of a researcher may be on studying a hypothetical causal relationship between a main exposure and a response, observational research

designs (for instance) do not share the same control over measured variables as an experimental study. Variables that are external to the main relationship are allowed to vary which presents a decision for the analyst. If the role of these variables is ignored they can influence the cause-effect relationship between the main exposure and the disease, thus potentially influencing the interpretation. However, should the answer be to adjust for these variables in the model? If the variables are expected to have a correlation based on external biological knowledge, then adjustment for such a variable would increase the precision of the estimate. In comparison, if a causal relationship is not hypothesized, the correlation is a nuisance to the estimation and will add bias to the coefficient. The need to develop new diagnostic tools to evaluate the extent of collinearity in statistical models, both beneficial (i.e. adjustment for confounding) and adverse (i.e. bias) is crucial to understanding the impact of collinearity on the modelling process. This also demonstrates why the use of the term collinearity ‘diagnostic’ is an uncomfortable one in applied research. To have a *diagnostic*, we must first have a *disease*. This is not always an accurate description of collinearity.

The notion that variables may share some common latent biological mechanism should motivate us to better understand the form and structure of the system through the analysis of observed collinear variables. As with collinearity diagnosis, ‘blanket’ approaches in the form of exploratory and confirmatory factor analysis have been applied in a wide range of subject fields, but the use of these methods and the means of application often experience little justification in the clinical literature. On the surface, the general intentions of employing such approaches are likely to be well understood, but the required theoretical understanding is complex. An exploratory factor analysis requires the user to specify the factor extraction, the number of factors to retain, the method of rotation and the interpretation of the factor loadings. These decisions are not arbitrary and can lead to a wide range of subjective interpretations. For instance, the choice of employing a principal components analysis or factor analysis is often made with little concern, however the methods are based on different underlying causal models that reflect a specific causal structure. For statistical inference to aid with clinical understanding, such methods must not be readily interchanged and the decisions must reflect the users biological or clinical understanding of the problem. In practice, these key decisions are often disregarded and confused. For example, principal components analysis is commonly listed under the heading of “factor analysis” in many statistical packages and simply seen as the ‘default’ option of an exploratory approach. It is essential that these approaches are met with a

greater concern and an understanding of the consequences of such decision making. If this can be achieved, there is the potential to attain reliable and consistent statistical evidence from observed collinearity that reflects an unobserved population model.

Although some degree of collinearity is present in almost all forms of biological data, the presence of perfect collinearity somewhat re-defines the problem. This is a violation of the classical assumptions of least squares regression. A perfect linear relationship amongst the covariates will prevent the researcher from obtaining regression estimates for all of the predictors in the model. The estimator is unable to partition the variance explained in the response into that attributable to each of the predictors involved in the linear dependency. This feature can sometimes be remedied if, for instance, the researcher has entered predictors measuring the same quantity simply by removing one of the predictors in the model. However, in some applications, entering the full complement of predictors could be justified conceptually as essential to the problem. Therefore, simply removing one of the covariates may no longer be an appropriate solution as this would add bias to the estimation if the constraint is not reflective of the population model. Some researchers have turned to the use of more complex regression methods. Unfortunately, there is often a lack of justification in employing such methods. This is particularly crucial in the perfect collinearity problem to the interpretation of the results. Whilst some estimators will be able to bypass the terminal problem of least squares, the justification of these methods is often loose or simply not presented.

This thesis is focussed on epidemiological and clinical applications. However, due to the popularity of regression methods, many of the ideas will be applicable to a wider range of subject areas if their application can be correctly justified. Whilst statistical approaches to problems encountered in epidemiological research will be considered in this work, it is important to highlight the role of the clinical context and external knowledge in determining what the methods indicate in practice. This adds an extra level of complexity to the problem. A major feature of this work is the use of vector geometry to construct statistical methodology and visual diagrams to simplify the problem for the benefit of both statistical and non-statistical thinkers. For instance, least squares, principal components regression, partial least squares can be compared as different geometrical projections in the same space. The vector geometry performs a useful role to demonstrate the importance of the response to the assessment of collinearity and provides the fundamental ideas to incorporate a response into a new collinearity index.

1.2 Aims of the Project

The work in this thesis has focused on the broad topic of collinearity in epidemiology. This involved understanding the impact of collinearity on statistical estimates and subsequently any potential impact on the clinical conclusions formed. The original research proposal outlined specific objectives to generate a new collinearity index that would incorporate the important role of the response in dictating the impact of collinearity. It was intended that geometry would play a key role in the formation of a novel index and also communicating the methodology to a primarily non-statistical audience. The project began with simulation work to understand the statistical impact of collinearity – identifying common features from the analysis of collinear variables such as coefficient changes of sign, deflation (or inflation) of point estimates and the impact on the precision of the estimate. The notion of causality becomes an important concept in understanding the meaning of these features in epidemiology. This includes examples of confounders, mediators and competing exposures which may be statistically indistinguishable but conceptually very important. This had to be considered for any index intended for use in model building.

The first attempt at forming a collinearity index was the development of the ‘matroid approach’. Matroids can be used to generate a dependency structure using an existing collinearity index. However, as current indices focus on the covariance amongst the predictors only, this methodology did not provide an answer to the original goal of incorporating the response. Instead the matroids work generated a new project and an interesting statistical discussion regarding the identification of the metabolic syndrome construct (see chapter 6). The work progressed onto the traditional geometry of multiple regression. A simple adaptation of this geometry generated the idea for an entirely novel collinearity index - labelled the D-index. The D-index would offer a ‘global’ measure of the impact between regression models, which importantly incorporates the role of the response (see chapter 5). Partial least squares further provided an insight into the potential of the D-index and a justification for its use in application. The geometry also highlighted a further advancement of the index to identify the role of each predictor in contributing to this global impact. With the guaranteed presence of collinearity in epidemiological research, the D-index has a huge potential to impact greatly on statistical modelling in epidemiology and related fields. An extension project was also undertaken in the final year of the project to consider a special case of collinearity in epidemiology encountered in lifecourse and age-period-cohort studies (see chapters 7-8).

1.3 Thesis Structure

The thesis consists of nine chapters in total. Chapters 2, 3 and 4 are intended to provide a background and explore the motivation for the methodology developed in later chapters. Chapters 5 to 8 consider three key areas in statistical epidemiology: the role and development of collinearity indices (chapter 5), the formation of a latent structure using exploratory analysis (chapter 6), and the analysis of perfectly collinear variables (chapters 7 & 8). Each of these chapters includes a worked example to demonstrate an applied use of the techniques developed.

Chapter 2 provides a background to regression in epidemiology. A portion of this chapter is dedicated to discussing the ordinary least squares estimator (OLS). The concepts of causality and confounding are then introduced through path diagrams and the directed acyclic graph. This is intended to highlight the complex relationship between statistical and clinical inference. Finally, a vector geometry approach to regression models is introduced that will be utilized throughout the thesis to illustrate and develop methodology.

Chapter 3 focuses on the impact of entering linearly dependent covariates on the OLS estimate. A selection of collinearity indices are presented with their merits of application discussed in relation to applied research. These techniques will often measure the degree of collinearity amongst a set of covariates. This information is limited if we wish to truly understand the role of collinearity in regression studies. Simulation work will build upon the geometrical construction of chapter 2 to investigate how changes to the response, the predictors and the error in the model can impact on the OLS estimates and any subsequent clinical interpretations.

Chapter 4 introduces a selection of multivariate methods for use in epidemiological application. These include principal components analysis, factor analysis, partial least squares and cluster analysis. They each are presented along with geometrical illustrations of the methods. The projection and rotation techniques that were introduced in chapter 2 are utilized to illustrate links between techniques. A particular focus is placed on the model assumptions to justify their use in application. This is crucial to later chapters that discuss the justification of applying such methods, which will ultimately promote or place caution on their use.

Chapter 5 focuses on developing a new index to assess the impact of collinearity in epidemiological research. The limitations of existing techniques are first considered along with possible extensions to correlation based indices. The intention is to incorporate the covariance structure between predictors and the response into this new measure. The index that is formed considers the deviation in model coefficients, both from a 'global' measure of the impact on the model and also the role of individual predictors in producing this impact. The idea is developed from a geometrical understanding of the regression model. It is first illustrated using the bivariable model in 3-dimensional space. This is extended to the three predictor model and later to a general index. Including the response in an index also requires some measure of uncertainty in the estimates. This is illustrated in the geometry and a confidence interval produced for the bivariable case.

Chapter 6 analyzes the structure of collinear risk factors in the study of metabolic syndrome (MetS). Traditional approaches in the study of MetS are considered, such as principal components analysis and exploratory factor analysis. In psychological research, the use of such exploratory methods is often scrutinized, however in the epidemiological and clinical literature there appears to be less guidance. Prominent researchers in psychology have argued for caution to be taken regarding the subjectivity of the decision making in these methods and the use of (potentially) default decisions. These arguments are considered in an epidemiological context and potential alternatives are identified that could improve the consistency of methodological decision making and interpretations in applied studies. An existing technique in the SAS library, labelled PROC VARCLUS, is proposed, whilst a novel methodology based on matroid theory is coded and developed.

Chapters 7 & 8 consider the problem of entering perfectly collinear variables in the regression model. Chapter 7 considers an example in a lifecourse setting involving continuous predictors and chapter 8 looks at the age-period-cohort model using categorical predictors. Although the concept remains the same in entering perfectly collinear predictors, the problem is somewhat redefined by the variable types. The categorical predictors encounter perfect collinearity in the full model, but also a further perfect collinearity in each of the categories added. The identification problem is considered in the least squares model with a focus placed on the generalized inverse. This discussion is extended to the novel use of shrinkage regression methods. The final chapter provides a framework for comparison of various 'solutions' suggested in the literature.

1.4 Publications

Chapter 5

3. Woolston A, Tu YK, Baxter PD, Gilthorpe MS. (2010) A New Index to Assess the Impact of Collinearity in Epidemiological Research. *Journal of Epidemiology and Community Health* 2010; 64:A59.
4. Woolston A, Tu YK, Baxter PD, Gilthorpe MS. Measuring the Impact of Collinearity in Epidemiological Research (under review).

Chapter 6

5. Woolston, A., Baxter, PD., Gilthorpe, MS., Tu, Y-K., Goodman, E. (2009) An Analysis of the Structure of the Components of Metabolic Syndrome using Matroids; *Journal of Epidemiology and community health*, 2009; 63 (Suppl II): A17.
6. Woolston, A., Tu, Y. K., Baxter, P. D., & Gilthorpe, M. S. (2012). A Comparison of Different Approaches to Unravel the Latent Structure within Metabolic Syndrome. *PLoS One*, 7(4), e34410. doi:10.1371/journal.pone.0034410.

Chapter 7

7. Tu, Y. K., Woolston, A., Baxter, P. D., & Gilthorpe, M. S. (2010). Assessing the impact of body size in childhood and adolescence on blood pressure: an application of partial least squares regression. *Epidemiology*, 21(4), 440-448. doi:10.1097/EDE.0b013e3181d62123
8. Tu, Y. K., Woolston, A., Baxter, P. D., & Gilthorpe, M. S. (2010). On Separating the Effects of Body Size and Growth on Later Blood Pressure. *Epidemiology*, 21(4), 452-453. doi:10.1097/EDE.0b013e3181e08d4d

Related Work

9. Tu, Y. K., Gilthorpe, M. S., D' Aiuto, F., Woolston, A., & Clerehugh, V. (2009). Partial least squares path modelling for relations between baseline factors and treatment outcomes in periodontal regeneration. *Journal of Clinical Periodontology*, 36(11), 984-995. doi:10.1111/j.1600-051X.2009.01475.x
 10. Gale, C. P., Cattle, B. A., Woolston, A., Baxter, P. D., West, T. H., Simms, A. D., . . . West, R. M. (2012). Resolving inequalities in care? Reduced mortality in the elderly after acute coronary syndromes. The Myocardial Ischaemia National Audit Project 2003-2010. *European Heart Journal*, 33(5), 630-639. doi:10.1093/eurheartj/ehr381
-

1.5 Notation

In this thesis, matrices are denoted by boldface uppercase letters (e.g. \mathbf{X}), column vectors by boldface lowercase letters (e.g. \mathbf{x}_j) and scalar values by lower case italic letters (e.g. a). The transpose of a matrix is denoted by a T in the superscript of the matrix or vector (e.g. \mathbf{X}^T) and a matrix inverse by a -1 in the superscript (e.g. \mathbf{X}^{-1}). A vector shall be denoted with an arrow (e.g. $\vec{\mathbf{x}}_j$). A generalized inverse is denoted by a $-$ in the superscript (e.g. \mathbf{X}^-). A complete list of notation is provided below.

\mathbf{X}	An $n \times k$ data matrix of observations on the independent variables
\mathbf{y}	An n -vector of observations on the response variable
$\boldsymbol{\beta}$	A p -vector of population model parameters
\mathbf{b}	A p -vector of estimated model parameters
$\boldsymbol{\varepsilon}$	An n -vector of residual errors
$\hat{\mathbf{y}}$	Fitted (projected) response variable
p	Number of model parameters
k	Number of predictors entered into the model
n	Number of observations in the sample
f	Link function for the regression model
N	Normal Distribution
$\sigma_{\boldsymbol{\varepsilon}}^2$	Variance of the residuals
$s_{\boldsymbol{\varepsilon}}^2$	Standard error of the residuals
R_y^2	Coefficient of determination
$\mathbf{A}_{n \times k}$	An $n \times k$ sample matrix used to demonstrate general matrix operation
$E[\mathbf{A}]$	Expectation of \mathbf{A}
$VC[\mathbf{A}]$	Variance-Covariance Matrix of the predictors
$\text{Rank}(\mathbf{A})$	Number of linearly independent column vectors of \mathbf{A}
r_{ij}	Correlation between covariates i and j
sr_{ij_k}	Semi partial correlation between i and j , holding k constant
$pr_{ij_{kl}}$	Partial correlation between i and j , holding k and l constant

θ_{xy}	Angle between vectors \vec{x} and \vec{y}
$\ \vec{x}\ $	Length of vector \vec{x} (i.e. the standard deviation of \mathbf{x})
sd_x	Standard deviation of \mathbf{x}
R	Rotation Matrix
P	Projection Matrix
$\mathbf{I}_{n \times k}$	An $n \times k$ identity matrix
P_{AB}	A path coefficient from A to B
λ_j	The j^{th} eigenvalue
Λ	A $k \times k$ diagonal matrix of eigenvalues
V	A $k \times k$ orthogonal matrix with eigenvectors as columns (Weights in a PCA)
Z	Principal Components (Loadings in a PCA)
P	Weights in a Partial Least Squares
C	Partial Least Squares Components (Loadings in a PLS)
m	Number of factors/components retained
F	Residual matrix of the bilinear decomposition
$\text{Cos}(\theta)$	Cosine of θ (representing correlation in vector geometry)
$\text{Sin}(\theta)$	Sine of θ
$\text{Tan}(\theta)$	Tangent of θ

A list of abbreviations for methods, clinical terms etc. are provided in chapter 9.

1.6 Statistical Software

Most of the statistical analyses in this thesis were performed in R (R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.) SAS (version 9.1) is used in chapter 6 to perform VARCLUS analyses and STATA (Version 12) is used in chapter 8 to perform constrained regression. Vector geometry illustrations were produced using Archimedes Geo3D (version 1.2.11).

2. Regression Analysis in Applied Clinical and Epidemiological Research

Regression analysis describes a range of statistical methods that are designed to explore and forecast relationships amongst two or more variables. In its simplest form, it involves fitting a straight line to a scatter plot to explain the variation in a response variable attributed to changes in an explanatory variable. However, explaining the response is rarely restricted to a single predictor and a linear relationship. For this reason more advanced regression methods have been developed to handle a range of variable types and model structures. As such, the potential applications are huge - epidemiology alone can generate an almost infinite number of questions that look to investigate the natural and physical processes that surround us. These are often very complicated structures that are too complex to model in their entirety. Instead, the researcher will look to capture the essential features of the process, and along with sound external knowledge the models can be used to build and develop our understanding of a 'true' population structure.

In this chapter the concept of linear regression in epidemiology is introduced, both from an algebraic and geometrical perspective. Ronald Fisher (1915), one of the early exponents of geometry in regression, used the combination of techniques to great effect to construct a theory that had puzzled statisticians for years (Herr 1980). The two disciplines require very different thought processes. Whilst the thinking behind geometry can be uncomfortable in some situations, it can simplify complex mathematical theory in others. By following a similar path, a fresh perspective can be gained on some of the issues faced by applying regression methods in clinical research and the development of methodology in these areas. This chapter is intended to provide a background in the statistical understanding of regression methodology. Study design is discussed along with the application of regression methods in epidemiology. The chapter concludes with the construction of the ordinary least squares (OLS) estimate using only vector geometry, which will be valuable to understanding the behaviour of the estimator in later work.

2.1 The Origins of Regression in Application

Examples of regression analysis can be found in the work of Adrien-Marie Legendre and Carl Friedrich Gauss as far back as the early 19th century. However, it was the combined efforts of Francis Galton, Francis Ysiro Edgeworth and Karl Pearson in the late 19th century that defined the practice and demonstrated its applicability to a wide range of research areas (Stigler 1990). Galton (1822-1911), an esteemed meteorologist, psychologist and statistician, has often been considered the catalyst behind the concept of regression and correlation. He was the half cousin of Charles Darwin. Along with his own experiences from exploration, Darwin's first publication of evolution entitled "The origin of species" in 1859 sparked a fascination in Galton into the study of genetics. Much of his later work from 1865 onwards became focussed on the topic of heredity. He was particularly interested in how traits and characteristics were passed from one generation to the next.

Galton's work in heredity was generally of a statistical nature. However, the statistical analyses deviated from the standard procedures of the time. In 1875 Galton conducted an experiment in which he provided ten sweet pea seeds of seven uniform weights to seven of his friends and asked each to grow the plants and return them to him. The sweet pea provided an ideal test subject because they appeared self-fertilizing and generally hardy to conditions (seemingly lowering the impact of external factors). He plotted diameters of the offspring against those of the parent plant. Galton observed that although the means differed between the parent and progeny seeds, the variability about the mean remained approximately constant, regardless of the original parent diameter. In addition, Galton reported that the mean's of the progeny seeds (calculated from the sample of ten seeds for each weight), "reverted" toward a population mean.

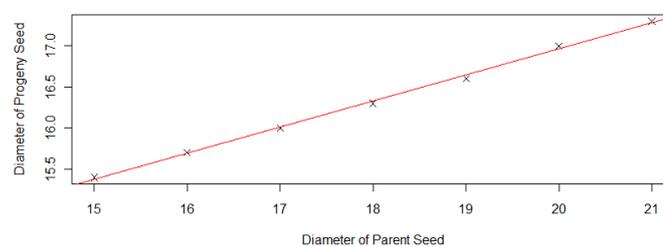


Figure 2.1: A plot of the mean diameters of Galton's sweet pea plants.

The plot of the parent vs. offspring seeds (Figure 2.1) was later described by Karl Pearson as "the first regression line" (Gillham 2002). Whilst ground breaking, this experiment was

hampered by two substantial flaws that ultimately impacted on Galton's conclusions. Galton labelled the sweet pea plant as self-fertilizing, which is not necessarily true. They can be cross fertilized, meaning that if the offspring from a large plant had been the product of a cross with a smaller plant, the seeds would on average be smaller. Also, the seeds were separated by their own size, not on that of their predecessors. Therefore, a large seed could have originated from a plant of any size. The size of the parent plant may have been influenced by environmental factors, which may not have been present in the environment Galton's friends grew them in. Through the nature of the experiment the plants were inclined to "regress" toward mediocrity.

Although the sweet pea experiment was flawed, it provided the first mathematical evidence that his theories on the passing of characteristics across generations were valid. It was also the first application of what we now realise to be regression analysis. These results encouraged Galton to progress onto what he considered his main problem in the natural inheritance of humans. Due to a lack of data, this extension was not straight forward. Galton set up an anthropometric laboratory in which he collected data from 1,567 visitors based on various non-standard measures such as "keenness of sight", "force of blow" and "breathing power" to which he had developed a range of devices to measure. Along with many other statistical analyses, Galton analyzed the heights of the children and their parents included in this sample. This added an extra dimension (and an extra level of difficulty) to the sweet pea plant study in that children are the offspring of two parents (although, unknown to Galton this was partially true in the plant study also). Therefore, Galton had to adjust his analysis to compensate for this.

The heights of the women were multiplied by a factor of 1.08 to adjust for general differences between men and women in height. Galton wished to compare the height of the adult offspring to a single height of the parents. This, he believed was justified and supported it by presenting a sample of 525 adults grouped by the difference in heights of their parents, observing no 'significant' pattern in the heights. The parental groups consisted of those with heights greater than or equal to the overall average (68 inches) being labelled "tall parents" and those below the average as "small parents". The heights again appeared to 'revert' from the extreme heights of the parent toward "mediocrity". This conclusion would appear to counter the theory of evolution, suggesting that everyone would reach the same attributes after a number of generations. This is a result of 'regression to the mean' (Bland and Altman 1994;Stigler 1990;Tu et al. 2004;Tu and Gilthorpe 2007). Galton's results are illustrated in Figure 2.2a,

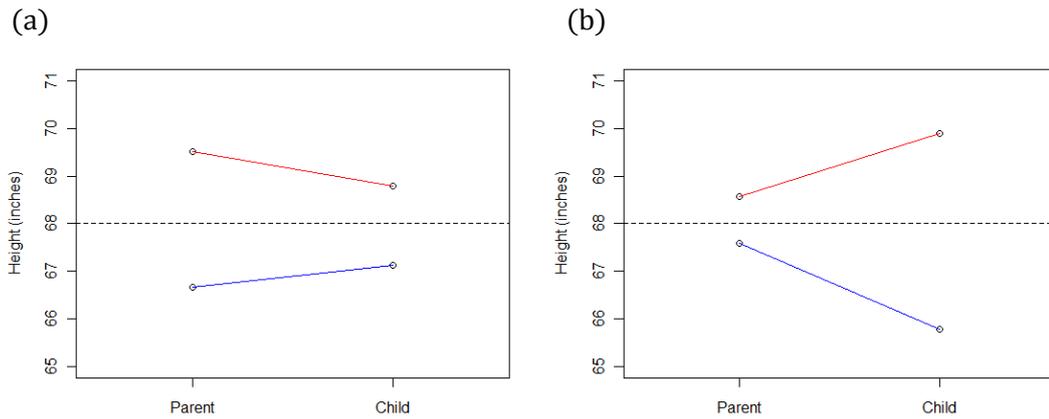


Figure 2.2: Average heights when grouped on (a) parental and (b) adult offspring.

In Figure 2.2b the heights of the adult offspring are separated by their mean height. Those with a height greater than or equal to the average (68 inches) are labelled “tall children”, whilst those shorter than the average are labelled “small children”. The average heights of the two groups now appear to diverge. Although both plots are based on the same set of data, they appear to be providing contradictory statements about the patterns of human height. In the first, an analyst may conclude that across many generations there will become fewer ‘tall’ and ‘short’ humans and eventually they will likely reach some common height. In the second, the interpretation is that there will be many more ‘tall’ and ‘short’ humans in the future. Of course, neither interpretation represents any true biological significance, they are instead a statistical result of ‘regression to the mean’.

This result occurred because the correlation between the parent and adult offspring heights in Galton’s data was not perfect (i.e. $r_{12} = 1$ - see section 1.5 for a list of notation). If this were the case then the lines would be parallel as the heights would not change. The correlation is in fact between 0 and 1, therefore the lines appear to diverge from the variable in which the subjects were grouped. These plots only serve to exaggerate the correlation found in the sample. They can easily be misinterpreted to suggest that the heights were diverging, when in fact the correlation could still be close to unity. Depending on which variables were grouped, the following measurements would always appear to diverge. If the heights were traced back further, then the variation in the grandparents would appear much larger. This is because the association between the child and the parent will be stronger than that between the child and the grandparent. Therefore, the regression plots from Galton’s data illustrate only the coefficients from a ‘simple’ (i.e. single predictor) regression (Tu and Law 2010).

The results of the sweet pea experiment along with that of the human heights, were presented in Galton's book "Natural Inheritance". This text is a culmination of all that Galton had worked on from 1877 to 1888 and is considered one of the most influential works in modern statistics and by many to be the birthplace of biometrics (Gillham 2002). Galton suggested that rather than height being a "simple element" it was in fact determined by "over a hundred bodily parts". Although, much of what we know today in genetics was unknown to Galton, his ideas (although often vague and using metaphor) touched upon some of the fundamental concepts that provide the corner stone to which the work is currently built – such as environmental influences, dominant/recessive traits and particulate inheritance. He also hypothesized that each parent contributes half of their own latent and personal attributes. In a later paper entitled 'A diagram of heredity' (Galton 2003), Galton presented his "law of ancestral heritage" through the division of a square which highlights the very essence of multiple regression (see Figure 2.3).

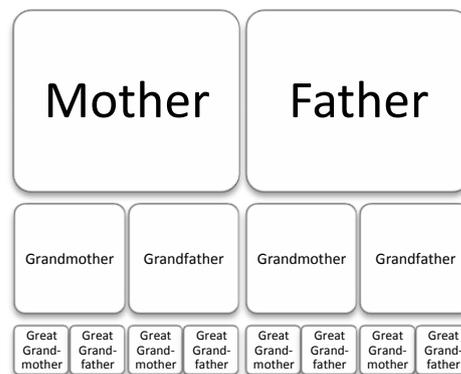


Figure 2.3: Galton's distribution of heritage.

Each characteristic or trait has been influenced by a multitude of factors from previous generations, with more recent carrying a greater weight than others. Whilst Galton's work generated the ideas of regression, much of it was graphical and described by example. It was Edgeworth (1845 - 1926) who managed to generalize the mathematics to be applied in a much broader scope and Pearson (1857-1936) who provided the mathematical rigour to the techniques (Stigler 1990). The development of regression methods has a complicated history and is clouded by conflicting views and disagreements. However, many of the ideas inspired by Galton have since been extensively studied and developed. This has made regression a viable option in a wide range of disciplines and when used with appropriate caution, can provide a valuable insight for researchers into the relationships of variables in an applied setting.

2.2 The Regression Model

Regression analysis describes a collection of statistical methods that are intended to draw inferences and forecast relationships amongst variables in a scientific system (Myers 1990). The relationships are expressed in the form of a model that describes the behaviour of a response variable in terms of one or more covariates. The response is denoted y and the set of predictors by a group of fixed variables \mathbf{x}_j (where $j = 1, \dots, k$ for k covariates with $i = 1, \dots, n$ observations in the sample). The general regression model is built using a link function f of \mathbf{x}_j and a residual error term ϵ (see eqn(2.1)).

$$y = f(\mathbf{x}_j) + \epsilon \quad (2.1)$$

The link function describes the relationship between the mean response and the predictors. This is the deterministic portion of the model (Freund and Wilson 1998). The model also contains a stochastic component (ϵ), necessary to account for any uncertainty when observations deviate from the population mean (e.g. a result of model misspecification, measurement error or biological variation due to external influences such as the environment).

The complexity of the model function can vary greatly and is dependent on the application and how much is known about the process being studied (Rawlings et al. 1998). Most preliminary work will begin with the use of a linear additive model,

$$y = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \dots + \beta_k \mathbf{x}_k + \epsilon \quad (2.2)$$

where β_j ($j = 0, \dots, k$) are unknown parameters to be determined from the data. These are labelled partial regression coefficients. They are defined as the change in the response, brought about by a unit change in the associated predictor whilst holding all other covariates in the model constant. The set of covariates \mathbf{x}_j are assumed to be non-random and measured without error (Myers 1990). The linear model is just one of a number available to the researcher, but it is often the first to be utilised due to its interpretability and developed inference methods. When f is not the identity, eqn(2.1) is termed a 'non-linear regression model', though the expression may still be a linear combination of \mathbf{x}_j 's (e.g. a polynomial model). This includes 'probit' and 'logit' models. The non-linear class of models may not follow the distributional assumptions of the linear form and so more complex iterative estimation and non-parametric inference methods may be required.

2.2.1 Applied Regression Analysis

The aims of the researcher in a regression study are generally split into two categories - those who wish to forecast future values of a response, and those who look to understand the system. In terms of prediction, the realism of the model is of little importance as long as it adequately predicts future values (Rawlings et al. 1998). In comparison, if we wish to make inferences about the population structure, it is vital that the model reflects the form of the system. In some situations, a linear assumption on the parameters may not be adequately satisfied and a non-linear alternative may be considered more realistic (e.g. modelling crop growth). However, the complexity of these models makes interpretation of the coefficients difficult. When distributional assumptions are relaxed, non-parametric tests must be employed – thus further adding to the complexity.

Naturally, the complexity of regression techniques has built with the advancement of computers and software. Methods that were considered too computationally heavy for practical use a decade ago have become viable options in mainstream statistical packages today. Unfortunately, this has brought its own danger. The growth in complexity should only heighten the caution taken by the researcher. This makes it vital for the user to understand and be able to justify reasons for employing a particular method and be clear on the aims of their research. Myers (1990) describes many of these methods as a 'black box'. The data is "thrown in one end and results come out the other". There is a risk that without truly understanding the process or how the estimates have been formed, we can easily misuse or misinterpret the results gained from such methods (Rawlings et al. 1998).

The approach employed by Galton and Pearson to estimate the unknown parameters of the regression model is labelled the least squares estimator. It is the simplest and most intuitive available to the researcher. The estimates also have several desirable properties under certain assumptions. These advantages are reflected in the popularity of the methodology, with this estimator often providing the baseline approach. It is important to note that the least squares estimator is just one of a number of methods available. In the late 1960's alternative estimators began to appear in regression studies (Myers 1990). These methods are often better suited to the data when the assumptions are violated in application (which is often the case). However, to recognize the 'failings' of this 'simple' technique, an understanding is gained of what the problem is and when an alternative estimator should be considered.

2.2.2 The Least Squares Estimator

Least squares regression is generally credited to Carl Friedrich Gauss, with a reported first application in 1795. However, this work was never published and so the credit is based primarily on indirect evidence and the word of Gauss (Stigler 1990). Adrien-Marie Legendre was the first to publish evidence of the technique including an example in 1806. Gauss and Pierre-Simon Laplace later fully presented both the method and theory algebraically around 1809. The method was used principally in its early application for use in navigation and tracing the movement of planetary objects. However, it has since been implemented in a range of applications, with linear least squares (or ordinary least squares - OLS) regression used to generate parameter estimates of the linear regression model.

To present the theory of the OLS estimate the matrix form of the general linear regression model is introduced,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.3)$$

where \mathbf{y} is an $n \times 1$ vector containing the observed values, \mathbf{X} is an $n \times p$ (where $p = k + 1$) fixed matrix of predictors (i.e. $E(\mathbf{X}) = \mathbf{X}$), $\boldsymbol{\beta}$ is a $k \times 1$ fixed vector of unknown parameters and $\boldsymbol{\varepsilon}$ is an $n \times 1$ random vector of model errors. There are an infinite number of solutions for the unknown parameters $\boldsymbol{\beta}$ of the regression model. The OLS estimator looks to find \mathbf{b} to minimize the squared difference between the observed (\mathbf{y}) and fitted values ($\hat{\mathbf{y}}$) of the model (Myers 1990). The differences (or residuals) are contained in the residual vector $\boldsymbol{\varepsilon}$. Through the minimization of the sum of squared residuals (SSR) the OLS solution can be obtained.

$$SSR(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (2.4)$$

The partial derivative of SSR (eqn(2.4)) is derived with respect to $\boldsymbol{\beta}$ and set to zero. This minimizes the SSR and produces a system of equations known as the 'normal equations' (eqn(2.5)). Through the solution of the normal equations the least squares regression coefficients can be found (eqn(2.6)).

$$(\mathbf{X}^T \mathbf{X}) \mathbf{b} = \mathbf{X}^T \mathbf{y} \quad (2.5)$$

$$\mathbf{b}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (2.6)$$

Sample data is used to make inferences on the properties of the population model. This is based on certain assumptions about the population structure for the stochastic component $\boldsymbol{\varepsilon}$. These assumptions characterize the statistical properties of the estimated regression coefficients (\mathbf{b}). Classical least squares assumptions state that:

A1. A linear relationship exists between the response and the predictors;

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \cdots + \beta_k\mathbf{x}_k + \boldsymbol{\varepsilon} \quad (2.7)$$

A2. The predictors are strictly exogenous;

Zero conditional mean of the errors,

$$E(\boldsymbol{\varepsilon}|\mathbf{X}) = 0 \quad (2.8)$$

This ensures that the $\boldsymbol{\varepsilon}$ are uncorrelated with *each* predictor. This condition is satisfied when $\boldsymbol{\varepsilon}$ is independent of \mathbf{X} and $E[\boldsymbol{\varepsilon}] = 0$.

A3. The predictors are full rank;

There exists no perfect linear combination of the predictors.

$$\text{rank}(\mathbf{X}^T\mathbf{X}) = p = k + 1 \quad (2.9)$$

A4. Homoscedasticity of the errors;

The $\boldsymbol{\varepsilon}$ are independent of \mathbf{X} and $\boldsymbol{\varepsilon}$ has constant variance.

$$\text{Var}(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2 \quad (2.10)$$

A5. Non-autocorrelation of the errors;

$$\text{Cov}(\boldsymbol{\varepsilon}_q, \boldsymbol{\varepsilon}_w|\mathbf{X}) = 0 \text{ for all } q \neq w \quad (2.11)$$

A6. Normality of the errors.

$$\boldsymbol{\varepsilon}|\mathbf{X} \sim \text{NID}(0, \sigma_{\boldsymbol{\varepsilon}}^2\mathbf{I}) \quad (2.12)$$

Therefore, under assumptions 1-6 the errors will have the following distribution,

$$\boldsymbol{\varepsilon} \sim \text{NID}(0, \sigma_{\boldsymbol{\varepsilon}}^2\mathbf{I}) \quad (2.13)$$

The distribution of the errors will have the following influence on the distributional properties of the response. Under assumptions **A1-A3**,

$$E(\mathbf{y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = E(\mathbf{X}\boldsymbol{\beta}) + E(\boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} \quad (2.14)$$

With the addition of assumptions **A4-A5** the variance-covariance (VC) matrix is defined,

$$VC(\mathbf{y}) = E\left[(\mathbf{y} - E(\mathbf{y}))(\mathbf{y} - E(\mathbf{y}))^T\right] = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma_{\boldsymbol{\varepsilon}}^2\mathbf{I} \quad (2.15)$$

The distribution of the response \mathbf{y} under the complete least squares assumptions,

$$\mathbf{y} \sim \text{NID}(\mathbf{X}\boldsymbol{\beta}, \sigma_{\boldsymbol{\varepsilon}}^2\mathbf{I}) \quad (2.16)$$

The sampling distribution of the response and error is useful in determining the optimal properties of the OLS parameter estimates and under what conditions these properties are held.

2.2.3 Quality of an Estimator

The researcher will rarely acquire data that fits the least squares assumptions perfectly. Therefore, they have to be viewed as flexible to some extent. In practice, the process of random sampling should ensure that the predictors and the errors are independent and that the errors are random. Also the lack of a normal distribution of the errors (and subsequently the least squares estimators) can be avoided by obtaining a large sample size (when the study design allows). This is due to the asymptotic distributional assumptions of the generalized central limit theorem (Fisher 2011). However, the OLS estimate is sensitive to deviations away from assumptions and so any violations will naturally impact on the desirable properties of the estimator. A measure of the performance of an estimator ($\boldsymbol{\beta}$) can be attained through the mean dispersion error (MDE).

$$\text{MDE}(\mathbf{b}, \boldsymbol{\beta}) = E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})^T] = VC(\mathbf{b}) + (\text{Bias}(\mathbf{b}, \boldsymbol{\beta}))(\text{Bias}(\mathbf{b}, \boldsymbol{\beta}))^T \quad (2.17)$$

Eqn(2.17) illustrates that the MDE is calculated from measurements of the bias and VC matrix. When MDE is minimized, the performance of the estimator is maximised in this sense. The choice of estimator is based on an important 'trade-off' between accuracy and precision of the estimation.

Bias

Bias is defined as the difference between the expected value of the estimated parameter and the population parameter,

$$\text{Bias}(\mathbf{b}, \boldsymbol{\beta}) = E(\mathbf{b}) - \boldsymbol{\beta} \quad (2.18)$$

This definition can be used to observe the first important property of the OLS coefficient:

$$\begin{aligned} E(\mathbf{b}_{\text{OLS}}) &= E\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TE(\mathbf{y}) \\ (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} &= \boldsymbol{\beta} \quad \text{as } E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \end{aligned} \quad (2.19)$$

Under assumptions **A1-A3**, OLS is an unbiased estimator of the population parameters. The bias of the estimator (eqn(2.18)) is zero - i.e. $E(\mathbf{b}_{\text{OLS}}) = \boldsymbol{\beta}$. This implies that over a large sample, the average \mathbf{b} attained from OLS will equal the population $\boldsymbol{\beta}$. However, in application the user is likely to investigate a single sample. In some circumstances it may be useful to introduce a bias into the estimation to attain a lower MDE (for example, ridge regression will directly input a bias to attain a more stable $\mathbf{X}^T\mathbf{X}$ - see (Hoerl 1962)).

Efficiency

The VC (or dispersion matrix) of an estimator is defined as follows,

$$\text{VC}(\mathbf{b}) = E\left[(\mathbf{b} - E(\mathbf{b}))(\mathbf{b} - E(\mathbf{b}))^T\right] \quad (2.20)$$

The VC is a symmetric $k \times k$ matrix that provides the variance of the predictors on the diagonal and the covariance's above and below the diagonal. Under assumptions **A4-A5**, the OLS estimate has the following VC matrix:

$$\begin{aligned} \text{VC}(\mathbf{b}_{\text{OLS}}) &= E\left(\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right)\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right)^T\right) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TE(\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma_{\boldsymbol{\varepsilon}}^2(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1} = \sigma_{\boldsymbol{\varepsilon}}^2(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned} \quad (2.21)$$

When the assumptions are met, OLS has the smallest sampling variance of all linear unbiased estimators and is considered the "best" (i.e. minimum variance) linear unbiased estimator (or BLUE) by the Gauss-Markoff theorem. It is subsequently referred to as efficient in its class (see Rao (1999) for proof).

$$\mathbf{b}_{OLS} \sim N\left(\boldsymbol{\beta}, \sigma_{\varepsilon}^2 (\mathbf{X}^T \mathbf{X})^{-1}\right) \quad (2.22)$$

In practice these variance properties may not be satisfied. For instance, homoscedasticity of the errors (**A4**) is regularly violated in cross-sectional and time series data (Gujarati 2002). Whilst violation of this assumption will not affect the unbiased property of the OLS estimate, it can impact on the estimates of variance of the population parameters. Consequently, any inferences based on standard errors (such as hypothesis tests) could be problematic, often leading to an increased risk of type I errors (i.e. a false-positive result). Similarly, autocorrelation of the errors can lead to inflated t-values, which may result in the rejection of a true null hypothesis. Whilst the OLS estimate remains unbiased, violation of these assumptions will often lead to an inefficient estimate.

2.3 Inference in Epidemiology

The study of epidemiology concerns factors that affect the distribution and spread of disease. The primary aims are in understanding the causes of a disease and subsequently determining the means to control and prevent it at population level. The questions typically posed in epidemiological research relate to the association between an ‘exposure’ and a ‘disease’. The disease label refers to a response of interest, such as a medical, psychological or social condition and an exposure (or risk factor) is a variable of interest that the researcher believes to be causally related to the health outcome. Whilst statistical analysis can provide supporting evidence to a theory, causal inference remains a judgement based on study findings and/or biological knowledge.

The statistics outlined in section 2.2 are only useful in application if they are considered in context and are not intended to describe features beyond their means. The study of epidemiology is built around causal relationships and identifying clinical significance. The weight given to statistical significance in many studies is alarming, because (as observed in later chapters) estimates such as p-values and t-statistics are easily influenced by factors likely to be present in almost any epidemiological study – e.g. collinearity. Due to the complex nature of the environments present in epidemiology, a great deal of thought should be involved in constructing causal models and interpreting the analyses. In this section a focus is placed on the nature of the epidemiological study and the potential pitfalls in the complex relationship between statistical and clinical inference.

2.3.1 Study Design

To achieve the goals of research in epidemiology, a range of study designs have been proposed and implemented in the literature. 'Evidence hierarchies' are often used to demonstrate the relative influence of these designs on the progress of the field. As the questions and goals of the research differ, the order of the designs in the hierarchy is not an exact science. Some designs will not be suitable in all situations due to ethical and practical limitations. However, an illustration such as Figure 2.4 can provide a useful summary of the general consensus amongst researchers on the strengths of the evidence gained from the respective designs,

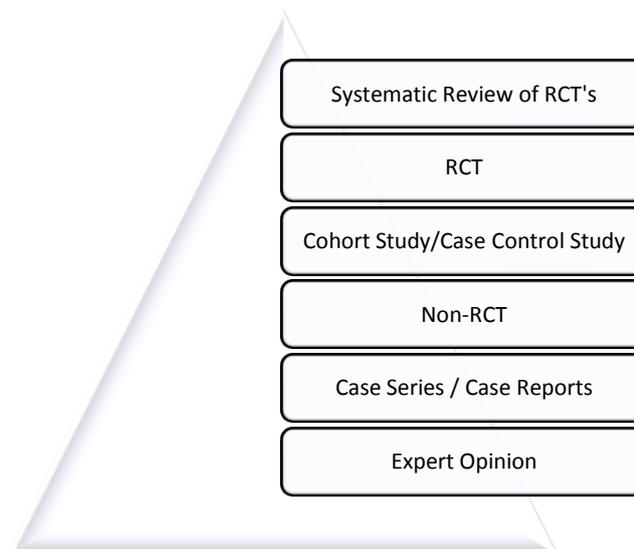


Figure 2.4: An example Hierarchy of Evidence.

In addition to illustrating the quality of the research designs, the pyramid representation has symbolised the quantity of published studies. High quality examples such as a systematic review or a meta-analysis have traditionally been rare in comparison to case reports, primarily due to monetary and time constraints in carrying out these studies. However, this is no longer the case with technological advancement improving the archiving of studies and the strengths of such designs being better recognised. The above hierarchy can be split into two categories – experimental and observational research. An experimental study involves the user assigning patients to treatment groups, whereas in an observational design subjects are recruited based on their exposure to some factor of interest – subject allocation to groups is not controlled.

Experimental studies include randomized controlled trials (RCT) in which the treatments are randomly assigned to subjects or non-RCT's in which no randomization occurs. The structure of an RCT means that subjects are followed up in the same manner beyond treatment allocation and receive the same health care. The benefit of this study design is that allocation bias can be minimized as any external factors to the treatment should be (approximately) balanced between groups. This gives the experimenter a greater control over baseline factors. At the highest point in the hierarchy is the systematic review. A systematic review is a summary of evidence surrounding a specific research question. A powerful systematic review will identify, assess, synthesise and interpret findings from all relevant publications (Hemingway and Brereton 2009).

As randomization does not occur, observational studies will give the researcher less control over the baseline factors of the subjects. The subjects can be matched on other factors, but this will not be as powerful as the randomization of an RCT. This can potentially bias the results. However, observational studies can provide a cheaper and less time consuming alternative to RCT's. In addition, there may be ethical considerations that make assigning patients randomly to treatment groups unfeasible. Also, for particular research questions, such as those aimed at rare diseases, large sample sizes and long study periods for RCT's would be required due to the infrequency of events. Although observational designs may be considered less robust than their experimental counterparts, the criteria of the study may dictate that an observational study is better suited to a particular research question.

Observational designs are split into descriptive and analytical forms. Descriptive studies - such as case reports, case series and cross-sectional surveys – are intended to identify and describe the incidence and trends in disease occurrence. For instance, a cross-sectional study can be used to take a 'snapshot' of the disease prevalence in a wide population at a particular time point. Analytical study designs are placed immediately above descriptive designs in the hierarchical structure. The aims of the analytical study are to examine why the disease is occurring, examine the nature of the causal mechanisms and quantify the relationship between the exposure and response. Popular analytical studies include case-control and cohort studies. The case-control involves the researcher selecting a group of individuals with the disease (i.e. cases) and a group without (i.e. controls), and the groups are subsequently compared. The cohort study involves the researcher grouping subjects who have been exposed to a risk factor of interest and another who have not. They are followed up to observe if and when the disease occurs in the future.

The main distinction between a descriptive and analytical study is the time point in which the response is measured. The cohort study looks prospectively at the disease occurrence (i.e. followed up over time), the cross-sectional study determines disease occurrence at the same time as the exposure or intervention and the case-control study retrospectively considers the exposure (often based on the recall of the individual). These studies are all susceptible to forms of bias, particularly as the researcher does not have the same level of control over the baseline variables as in an RCT. The studies must also account for hidden confounders and practical considerations to uncover the nature of any true causal relationships present amongst the variables of interest.

2.3.2 Causation and Causal Inference

The aim of an analytical or experimental study is to measure and explain the nature of any causal (or cause-effect) relationships. A 'cause-effect' relationship is defined when the incidence of a disease cannot increase without exposure to the factor of interest (Rothman et al. 1998). It is impossible to objectively prove a cause-effect relationship, instead epidemiologists will seek evidence to support a hypothesis (Scheutz and Poulsen 1999). Although no mathematical statement can be used to determine the existence of a causal relationship, a number of guidelines have been proposed. These attempts include the Henle-Koch postulates (1890), the Bradford-Hill criteria (1965) and Susser's criteria (1988). However, support of these criteria does not prove the existence of a 'cause-effect' relationship. Similarly, the absence of some factors can still occur when a causal relationship exists. External evidence and an element of judgement must play a role, which often leaves hypotheses open to debate.

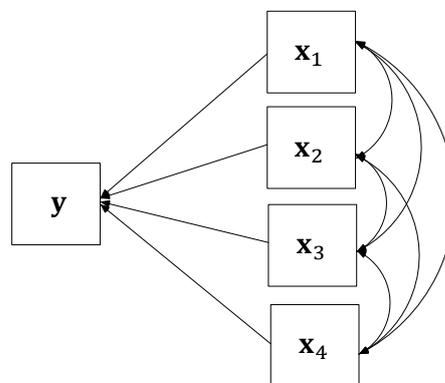


Figure 2.5: Path diagram representation of a multiple regression model.

For the linear regression models presented in section 2.2 causality is implied, however due to the statistical context of the discussion it has not been of primary concern. In epidemiology, even the simple linear regression model implies a causal relationship between the exposure x and the disease y . Figure 2.5 is an example of a path diagram (PD) representing a multiple regression model. This illustrates the causal relationships implied when a multiple regression is performed as part of a statistical analysis. The variables are represented by nodes (or vertices) and any relationships between two variables as arcs (or edges). The use of a directed arc (or arrow) demonstrates a causal relationship between two variables, with the direction of the arrow signalling the direction of causality. A double ended arrow (such as those that link the covariates) represents an implied covariance structure (i.e. no causal direction is specified). If the structure amongst the covariates is thought to differ from the standard linear regression model introduced in section 2.2, an advanced technique such as the directed acyclic graph (DAG) or a structural equation model (SEM) may be used for the analysis.

The use of a PD allows an enormous range of variations for the researcher to input external knowledge about the relationships of the variables. As causality cannot be proven using observational studies, the use of a directed arrow represents the users' *a priori* biological knowledge or a hypothetical relationship that they intend to assess. There is an implied sense of time in a PD, although the study design often dictates the nature of the temporal relationships. For instance, in a longitudinal study it is straightforward to identify the temporal flow from exposure to disease as the variables are measured over time. However, for a cross-sectional or questionnaire design, the subject will record both exposure/intervention and disease at the same time. Therefore, identifying which variable precedes the other can sometimes be difficult. A fundamental result of the causal structure is that the exposure precedes the disease. The path from exposure to disease (i.e. following the arrows) is labelled the "causal pathway".

A DAG is a PD such that no loops exist – i.e. a set of edges that can be followed to return to the same vertex (the PD in Figure 2.5 is not a DAG due to the vertices linking the covariates). The use of a DAG represents a great simplification of the environment in which the variables exist. It is a representation of the structure the researcher believes the sample to have been taken from – the study data may not adhere to this hypothetical construct. The nature of the implied causal relationships is not specified by the researcher. The effect may be harmful/protective, sufficient/necessary or effect modification may be present. However, the simplification provides a useful link between the statistical analysis

and the causal inferences made in epidemiological study. It allows the researcher to clearly communicate the hypothetical models and translate the findings to 'real life' situations. This opens the analysis to non-statisticians to debate the inclusion or exclusion of variables based on existing biological knowledge. DAG's are beneficial to consider which effects may be present under different causal relationships. From any study design, we can only obtain correlations between variables and not causality. Therefore, the relationships that are implied from *prior* knowledge can dictate the interpretation of any statistical phenomena that may be present in the analysis.

Confounding

Suppose a simple linear regression model (i.e. single predictor) is fitted to a set of data and we find the model to fit well. The exposure and disease are said to have a statistical association, but this does not prove a biological causal relationship exists between the variables. There are a number of reasons for the association that do not have to imply causation. Consider the example of a regression model for cardiovascular disease (CVD) with a single predictor of body mass index (BMI). After observing a significant regression coefficient, a researcher may present the conclusion that BMI is a risk factor for CVD. This statement implies that there is a causal relationship between BMI and CVD. However, justification of the link between observing an association (or lack of) and implying causality is not straightforward. The implied DAG presents the simple regression model as follows,



Figure 2.6: DAG demonstrating a causal relation between BMI and CVD.

A potential explanation for an inflated (or in some circumstances insignificant) association between exposure and disease can be the presence of one or more confounding variables. A confounder is a variable that is a cause, or a proxy of a cause, of the disease and also associated with the main exposure (i.e. BMI), but is not on the causal pathway (Rothman et al. 1998). A potential confounder may be the smoking status (SMK) of the subject.

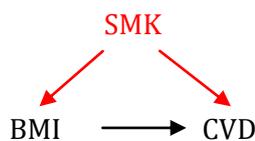


Figure 2.7: SMK as a confounder to the relationship between BMI and CVD.

Evidence suggests that SMK can lower BMI and is also a proxy for increased risk of CVD. The presence of the confounder enhances the coefficient in the simple regression model. Therefore, the researcher may decide to adjust for SMK by including it in the model. This example is a great simplification of the relationships present in this particular environment and there are likely to be multiple confounders, both known and unknown to the researcher. For instance, through adjustment for one confounder, additional confounding may be introduced into the model.

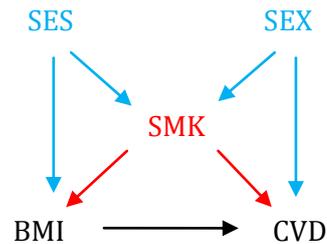


Figure 2.8: Potential confounders brought about by the inclusion of SMK.

If the confounder SMK is adjusted for in the model, social economic status (SES) and sex (SEX) may become confounders in their own right (i.e. both are causes of the disease and of the exposure through SMK). Therefore, these should be adjusted for in the model if the hypothesis is to be believed. This demonstrates the complexities when considering causality. The importance of identifying influential variables in the environment is apparent from this simple example – therefore the advantages of demonstrating causal links through DAG's are illustrated further. As there are likely to be numerous variables present in the study environment that can influence the relationship between the main exposure and the disease, the researcher must decide which variables to adjust for.

To control for confounding the researcher can choose to adjust for a 'sufficient' set of confounders **S**. Adjusting for **S** should ensure that there is no effect from confounders on the relationship of interest. A simple algorithm can be applied to check for sufficiency,

1. Remove all single headed arrows that leave the exposure variable;
2. Add a non-directed arc to connect each pair of variables that share a child (i.e. a node that is at the arrow end of an arc) in **S** or a descendent (i.e. a node that succeeds another but is separated by one or more nodes) in **S**;
3. Assess whether there is any unblocked 'backdoor path' (i.e. moving against the direction of the arrows) that goes from exposure to disease that does not pass through **S**.

When condition 3 is met, \mathbf{S} is a sufficient set to control for confounding. There is often more than one sufficient set to choose from. The 'optimal' set can be chosen to minimize the influence of the confounders. In addition, some sufficient sets may be unsuitable. For example, if the variables were not measured during the study or if particular variables had a greater measurement error this may also influence the decision.

Competing Exposures

Another potential influence on the relationship between the main exposure and the disease is the presence of 'competing exposures'. These are similarly defined as a cause, or proxy of a cause, of the disease, but are instead *not* a cause, or proxy of a cause, of the main exposure. The variable should not lie on the causal pathway. The example presented in the previous section is as follows if SMK were considered a competing exposure,

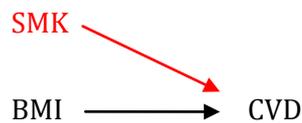


Figure 2.9: SMK considered as a competing exposure to BMI.

The competing exposure should have zero correlation with the main exposure in the population, however this may not be true in the study data due to chance sampling variation. Adjusting for the competing exposure in the model will not bias the estimate, however the association between a main exposure and an outcome will be incorrect in the single sample case. If this correlation structure is viewed over multiple samples, it may be hypothesized that there is some unobserved variable that precedes smoking status and BMI, and is causally related to both.

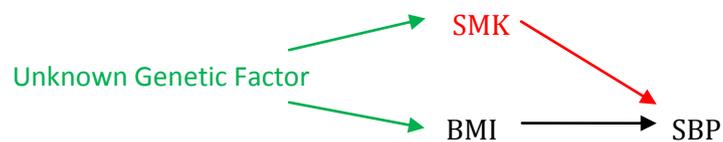


Figure 2.10: Hypothetical presence of a preceding unobserved variable.

If this were true, SMK would no longer be considered a competing exposure, but a proxy confounder. This would provide one explanation for the presence of an association in the sample. These variable types will be discussed in greater detail when collinearity is introduced in chapter 3.

2.3.3 Modelling Strategy

In epidemiological research it is important that *prior* knowledge be utilised to drive the modelling strategy. Whilst consistency across studies will add strength to a hypothesis, it cannot prove the model correct. A famous example is that of the hypothesized causal relationship between smoking status and lung cancer. The work of an American medical student Ernst Winder and a British scientist Richard Doll first brought the link to the attention of the public in the early 1950's. However, whilst numerous studies were conducted during the 1950's and 1960's demonstrating a strong association, the cigarette companies relied on the inability of statistical analysis to prove any causal relationship. The manufacturers proposed other explanations such as air pollution that could explain the presence of the correlation (i.e. a proxy confounder).

Ronald Fisher (a strong critic of the causal relationship), suggested a common genetic factor could play a similar role in this relationship. For instance, some genetic component that would influence the individual to smoke may also increase the risk of developing cancer. This hypothesis would never be proven. Jerome Cornfield later presented compelling evidence to demonstrate a direct causal relationship between exposure and outcome using an early meta-analysis of 14 observational studies (1966;1969). However, Fishers hypothesis of an unknown factor could still hold true despite the lack of evidence (Davey-Smith 2009). It is because of the inherent biases that accompany observational research that despite the mounting evidence, the cigarette manufacturers won every court case. Whilst a correlation structure can be repeatedly observed across studies, it is the clinical context that dictates the 'real world' interpretation of the statistics (however unlikely the counter-argument).

The DAG is a representation of what the researcher believes the population structure to be. These relationships are defined from *prior* knowledge or are hypothetical relationships to be assessed. They should not be driven by study data. The difference between the confounder and the competing exposure is that the confounder is a cause or proxy of a cause of the main exposure, whereas the competing exposure should not be. Consider the situation illustrated in Figure 2.11 - **X** is the main exposure, **Y** the response and **Z** is a potential confounder/competing exposure. If a correlation is observed in the sample data between **X** and **Z**, then external information should define the nature of the relationship (or if we believe one exists in the population at all).

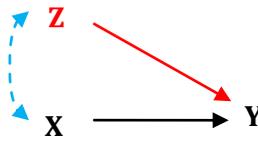


Figure 2.11: DAG demonstrating a sample correlation between two exposures.

If **Z** is labelled a confounder, identifying and adjusting for **Z** in the model will be beneficial to the precision of the estimate (as the correlation is expected). However, if they are defined as competing exposures, the correlation observed does not fit the hypothetical model, but may still be considered a product of sampling variation. Statistically, the covariance structure gained from the data could be considered acceptable in both cases. Whilst viewing sample correlation is considered an improvement to the estimation in the confounding case (i.e. improving precision), in the competing exposure situation it will be viewed as adding bias to the estimation. Therefore, identifying and labelling the variable as a confounder or a competing exposure prior to the analysis should define for the researcher whether or not it is sensible to adjust for **Z** in the analysis. This task becomes even more difficult when multiple external variables of the cause-effect relationship are identified and have to be defined.

There is always a tradeoff between the complexity of the population systems in epidemiology and achieving model parsimony. The nature of the epidemiologic environment is such that the full range of factors involved in most situations would be too difficult to identify and include in a single model – this would generally introduce unnecessary measurement error and bias. If a simple model can capture the complexity of the population structure adequately, model parsimony should be observed. This is a similar notion when identifying a sufficient set of confounders to eliminate external effects. The issue in many studies is that the model used is often too simple to assess the research question (i.e. ‘enough’ factors are not accounted for). However, by adopting a technique such as a DAG, the essential features of the model can be captured that can aid with linking the statistical results with the clinical interpretation. This can help with the identification of additional factors to consider and understand how to interpret the model. In the final section of this chapter another extension of the traditional algebraic regression model is considered by studying vector geometry. Vector geometry can play a similar role to DAG’s in simplifying the conceptual understanding of the estimates gained from the analytical regression model.

2.4 Vector Geometry of Linear Regression

Vector geometry offers a powerful illustrative tool that has the potential to provide a fresh perspective to that of the algebraic approach in multiple regression. Many of the complex relationships present in linear algebra can be translated into intuitive geometrical representations (Wickens 1995). For instance, Ronald Fisher (1915) used geometry to identify the exact distribution of the correlation coefficient of a sample of n pairs from a bivariate normal distribution - a problem that Pearson had struggled with prior to Fisher's breakthrough. However, it has since been mainly confined as an illustrative tool, often only utilized in scatter plot form. Herr (1980) proposes four theories for this;

1. There is a perception amongst statisticians that the analytic approach is 'traditional'. It would take a major shift in the balance of statistical work to encourage the use and acceptance of vector geometry in the same light.
2. Although papers such as Fisher's demonstrate the power of the geometrical approach, it is not immediately obvious to all. This suggests that whilst it is appropriate to "geometrical thinkers", it is not accessible to the majority.
3. Geometry is generally thought of as a mathematical tool and a skill. To teach it in statistics would require a capacity for abstract thought from the subjects.
4. It is believed (by some) that the geometric approach cannot achieve what the analytic approach can.

The first three theories would seem to have some standing and will remain reasons for why regression continues to be presented in analytical form. However, the final theory should not be readily accepted by statisticians. The role of geometry can play a substantial part in understanding how a method operates and identifying its limitations. The intention of the geometry used in this work is not to fully explain regression models – as we are only able to directly perceive a maximum of three dimensions – but rather to generate ideas that may not be obvious in analytical form. The geometrical representations are primarily intended to convert algebraic ideas to image. To understand the problems in the geometry and generate ideas about a solution, the methods may be carried out in more complex situations without the visual use of geometry.

2.4.1 Variable Space and Subject Space

The traditional graphical presentation in mathematics and statistics involves the user plotting observations as points on variable axes (e.g. a scatter plot). This format is referred to as variable space. It is an essential tool in statistics that gives an immediate summary of the data. Whilst this presentation is useful for gaining an insight into the observations, it limits our understanding of the relationship of the variables, which is one of the main objectives of this work. An alternative is to plot the variables in subject space. This means that observations become axes and the variables are plotted in the new space. The two plots are equivalent, but the emphasis switches from the observations to the variables (Wickens 1995).

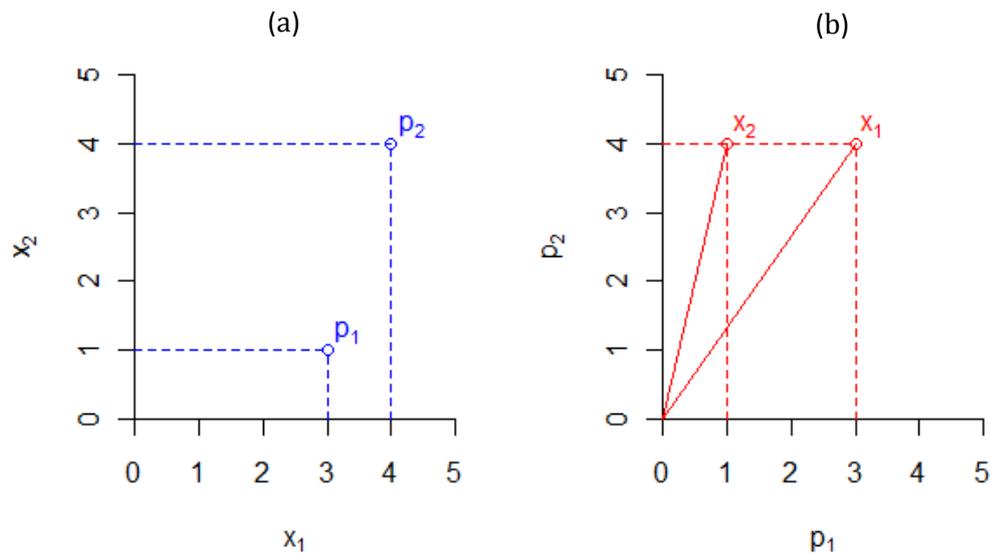


Figure 2.12: Observations (p_1, p_2) in (a) variable and (b) subject space.

An issue with the subject space representation is the dimensionality required to present the variables in this format. For example, for 50 observations on two covariates (x_1, x_2), we would theoretically require the variables to be plotted in 51-dimensional space (including the intercept). This problem is overcome by some abstract thought. If the variables were actually plotted in this space, only three points would actually need plotting (i.e. the intercept, x_1 and x_2). Therefore, to investigate the relationships between these points, only three dimensions are required. If the data is centered, the dimensionality further reduces by one. The points are effectively plotted on 'coordinate free' axes.

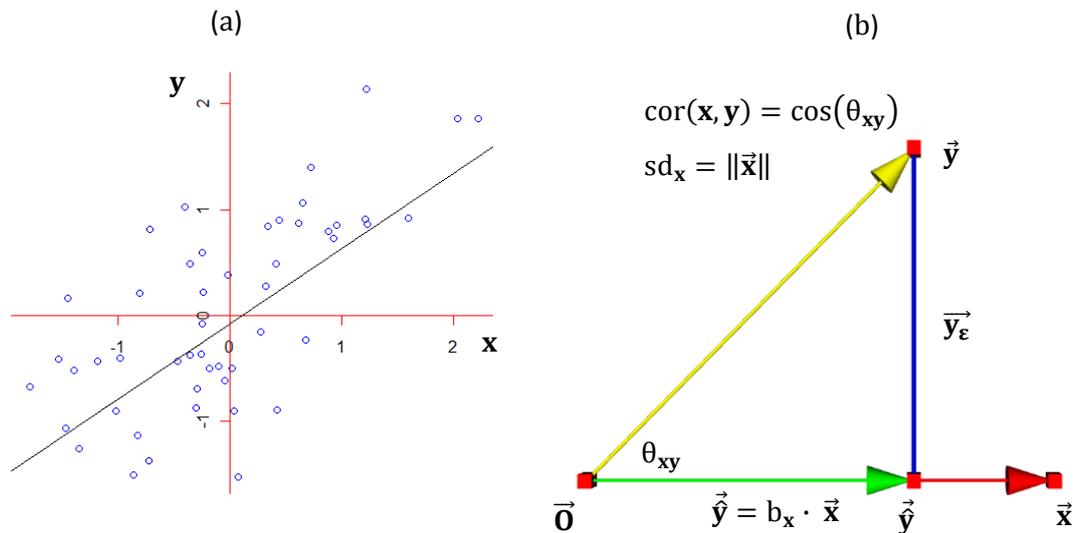


Figure 2.13: Simple linear regression in (a) variable space and (b) subject space.

The simple linear regression example in Figure 2.13b demonstrates fundamental results for geometry in subject space. The correlation of the response and predictor is represented by the cosine of the angle between the vectors θ_{xy} – i.e. $\cos(\theta_{xy}) = r_{xy} = R_y$. This result allows the user to calculate the position of any vector in relation to any other in the subject space based on the correlation structure. The standard deviation is demonstrated by the length (i.e. magnitude) of the vector. Therefore, if the variables are standardized (i.e. scaled to unit variance) they will be represented by a unit vector.

Figure 2.14 demonstrates a multiple regression with three covariates and the formation of confidence intervals by projection of the sampling error onto bivariate regression planes. Such models follow the same basic principles and will be considered in detail later in this thesis.

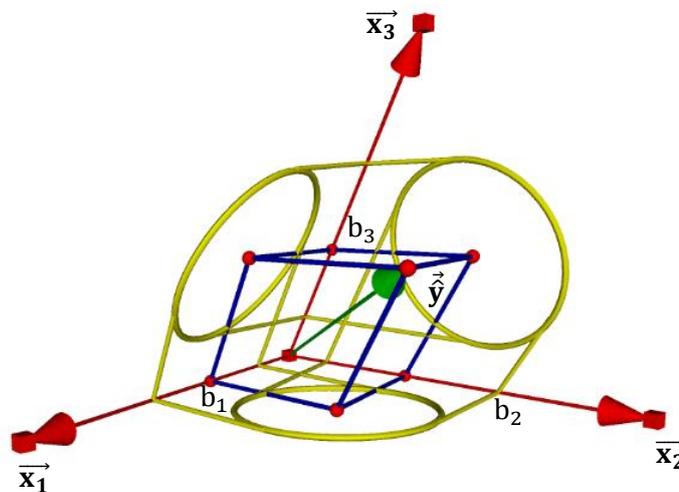


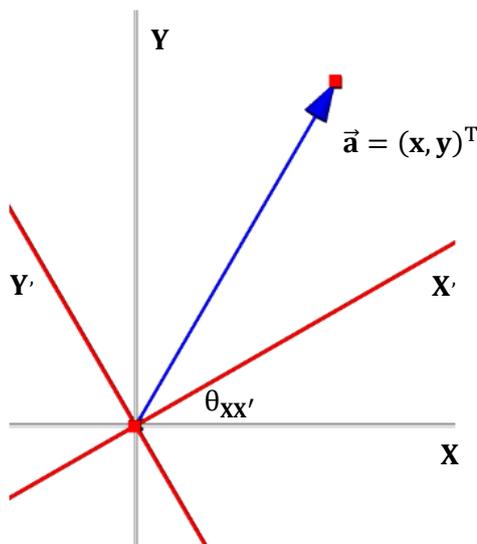
Figure 2.14: An illustration of the power of multivariate geometry.

2.4.2 Converting Analytic Ideas to Geometry

The vector geometry in this work is intended to provide a conceptual link to algebraic ideas. It appears appropriate at this stage to outline some of the basic results from linear algebra to help with the visualization in later sections – namely the form and role of rotation and projection matrices will be of particular importance to later work.

The Rotation Matrix

An axis rotation will transform a vector \vec{a} around a fixed point, whilst maintaining the distances and correlation structure with any other vectors. Consider the example in Figure 2.15. The axes are rotated by the angle $\theta_{XX'}$. The problem is to find the coordinates of the point (x, y) on the rotated $X'Y'$ axes,



$$x' = x \cdot \cos(\theta_{XX'}) + y \cdot \sin(\theta_{XX'})$$

$$y' = y \cdot \cos(\theta_{XX'}) - x \cdot \sin(\theta_{XX'})$$

In matrix form, the $k \times k$ rotation matrix \mathbf{R} transforms the coordinates as follows,

$$\begin{aligned} \mathbf{R}\mathbf{A} &= \begin{bmatrix} \cos(\theta_{XX'}) & \sin(\theta_{XX'}) \\ -\sin(\theta_{XX'}) & \cos(\theta_{XX'}) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\ &= \begin{bmatrix} x' \\ y' \end{bmatrix} \end{aligned}$$

Figure 2.15: An example of axis rotation.

The derivation of this result can be demonstrated by considering the correlations (and subsequently the angles) between the original and rotated axes. It follows that the matrices shown in eqn(2.23) are equivalent,

$$\begin{aligned} \mathbf{R} &= \begin{bmatrix} \cos(\theta_{XX'}) & \sin(\theta_{XX'}) \\ -\sin(\theta_{XX'}) & \cos(\theta_{XX'}) \end{bmatrix} = \begin{bmatrix} \cos(\theta_{XX'}) & \cos(\pi/2 - \theta_{XX'}) \\ -\cos(\pi/2 - \theta_{XX'}) & \cos(\theta_{XX'}) \end{bmatrix} \\ &= \begin{bmatrix} \cos(\theta_{XX'}) & \cos(\theta_{X'Y'}) \\ \cos(\theta_{Y'X}) & \cos(\theta_{YY'}) \end{bmatrix} \end{aligned} \quad (2.23)$$

The Projection Matrix

Consider U to be an n -dimensional vector space. W and V are defined as complementary subspaces of the original space U (i.e. such that any $\vec{u} \in U$ can be expressed as $\vec{u} = \vec{w} + \vec{v}$). A projection matrix P is an $n \times n$ matrix operator that projects \vec{u} along \vec{w} onto \vec{v} ; i.e. $P\vec{u} = \vec{v}$ (Meyer 2001). In this format, V and W are commonly referred to as the range and null spaces respectively. The matrix P must be idempotent – i.e. the vector \vec{u} is left unchanged in the range of V ,

$$\mathbf{P} = \mathbf{P}^2 \quad (2.24)$$

(Phatak and DeJong 1997)

The projection matrix is labelled ‘orthogonal’ if the following property holds,

$$\mathbf{P} = \mathbf{P}^T \quad (2.25)$$

The direction of projection is along the orthogonal complement to the projection plane.

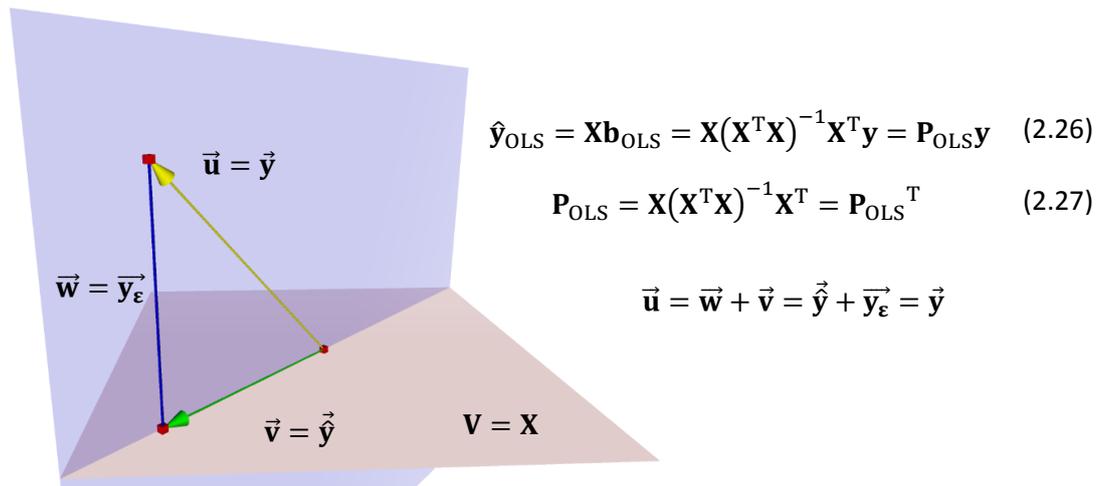


Figure 2.16: Orthogonal OLS projection of \vec{y} onto the regression space spanned by \mathbf{X} .

Geometrically, OLS is an orthogonal projection of \mathbf{y} onto the subspace spanned by \mathbf{X} (see Figure 2.16). Any projection matrix that does not adhere to the property in eqn(2.25) is labelled an ‘oblique’ projection. For an oblique projection, it is also necessary to specify the direction to which \vec{u} is projected onto V . The following is an example of an oblique projection of \vec{u} onto the space V (W and V are now not orthogonal),

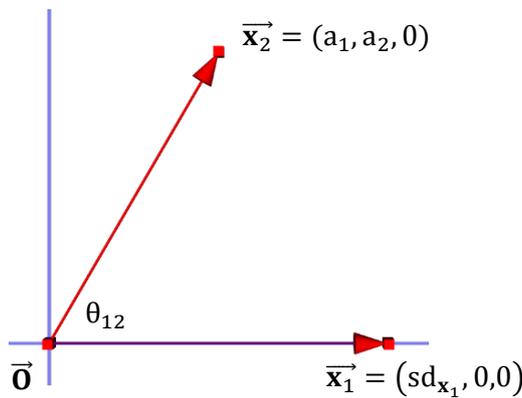
$$\mathbf{P} = \mathbf{V}(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T \quad (2.28)$$

(Phatak and DeJong 1997)

2.4.3 The Least Squares Estimate

In this section, a geometrical example of the OLS estimator is presented to illustrate vector geometry and demonstrate the process of converting algebraic concepts to image. The geometrical presentation allows us to investigate the interrelationships of the variables. These will be used in chapter 3 to explore the changes in estimates under varying conditions to improve our understanding of the behaviour of the OLS estimate.

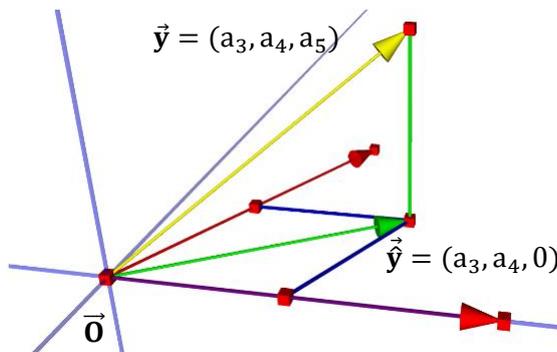
The mean centered two predictor case is first considered due to the simplicity of working in 3-dimensional geometry (i.e. two centered covariates and a single response). Correlations amongst \mathbf{x}_1 , \mathbf{x}_2 and \mathbf{y} are obtained from the data (which determine the angles between vectors) along with the standard deviation of each variable (determining the length). If $\vec{\mathbf{x}}_1$ is assigned to lie on the standard x-axis, the positional vectors of the variables can be constructed relative to this point. The positional vector of $\vec{\mathbf{x}}_2$ is calculated using the correlation between \mathbf{x}_1 and \mathbf{x}_2 , along with the standard deviation of \mathbf{x}_2 (i.e. $sd_{\mathbf{x}_2}$).



$$a_1 = sd_{\mathbf{x}_2} \cdot \cos(\theta_{12})$$

$$a_2 = sd_{\mathbf{x}_2} \cdot \sqrt{1 - \cos(\theta_{12})^2}$$

The position of the response (\mathbf{y}) and its orthogonal projection (see section 2.4.2) to the plane ($\hat{\mathbf{y}}$) spanned by \mathbf{X} can be similarly found using a partial regression coefficient,

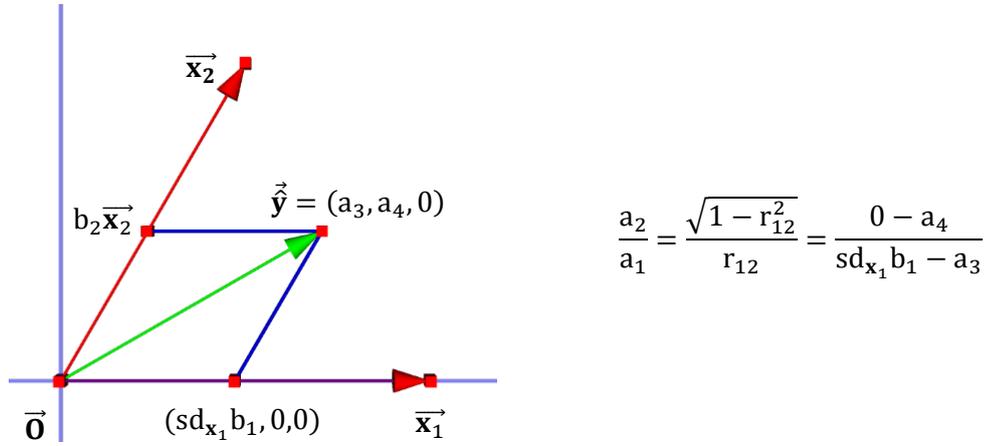


$$a_3 = sd_{\mathbf{y}} \cdot \cos(\theta_{y1}) = sd_{\mathbf{y}} \cdot r_{y1}$$

$$a_4 = sd_{\mathbf{y}} \cdot \frac{\cos(\theta_{y2}) - \cos(\theta_{y1}) \cdot \cos(\theta_{12})}{\sqrt{1 - \cos(\theta_{12})^2}}$$

$$a_5 = \sqrt{sd_{\mathbf{y}}^2 - a_3^2 - a_4^2}$$

The coefficient b_1 can be found by considering the intersection of the \vec{y} projection onto \vec{x}_1 ,



This simplifies as follows,

$$\begin{aligned}
 \left(\frac{\sqrt{1-r_{12}^2}}{r_{12}} \right) (sd_{x_1} b_{x_1} - sd_y r_{y1}) &= -sd_y \frac{r_{y2} - r_{y1} r_{12}}{\sqrt{1-r_{12}^2}} \\
 \frac{sd_{x_1}}{sd_y} b_1 &= r_{y1} - \left(\frac{r_{y2} - r_{12} r_{y1}}{\sqrt{1-r_{12}^2}} \right) \left(\frac{r_{12}}{\sqrt{1-r_{12}^2}} \right) \\
 &= \frac{(1-r_{12}^2)}{(1-r_{12}^2)} r_{y1} - \frac{r_{y2} r_{12} - r_{y1} r_{12}^2}{(1-r_{12}^2)} \\
 &= \frac{r_{y1} - r_{12} r_{y2}}{1-r_{12}^2} \tag{2.29}
 \end{aligned}$$

The result for b_2 follows a similar derivation. The OLS estimate is centered and unstandardized in this geometrical construction (although the predictor standard deviations are equal and \vec{y} is unit length). Centering has removed the intercept and allowed the geometry to be presented in one less dimension than required for subject space. The relationship expressed in eqn(2.30) will generate the standardized coefficient from its unstandardized form.

$$b_1 = \underline{b}_1 \cdot \frac{sd_y}{sd_{x_1}} \tag{2.30}$$

where \underline{b}_1 is the standardized regression coefficient for \mathbf{x}_1 . If \mathbf{y} remains centered and scaled to unit variance, the orthogonal projection of \vec{y} onto the vector \vec{x}_1 is the product of the coefficient estimate and the length of the vector in unstandardized form.

2.5 Conclusions

This chapter has provided an introduction to some of the basic concepts of regression analysis, both from a statistical and applied viewpoint. The discussion of regression in epidemiology (section 2.3) is intended to demonstrate some of the gaps between the statistical process of generating an estimate and the clinical interpretation in making use of it. However precise and detailed the association found in epidemiological studies, a cause-effect relationship cannot be proven empirically. Whilst viewed as the “gold standard” approach in epidemiology, an experimental study (even without error) can only strengthen a clinical inference, rather than provide a definitive proof.

“All of the fruits of scientific work, in epidemiology or other disciplines, are at best only tentative formulations of a description of nature”

(Rothman and Greenland 2005)

The use of the PD, DAG and SEM are an important feature of epidemiological work - placing statistics in an epidemiological environment. They allow the researcher to consider the nature of the relationships in the study environment and open the ideas to non-statistical thinkers. What the estimates represent can only be determined if external biological knowledge and potential influences are considered on the hypothetical cause-effect relationship. The confounder and competing exposure argument presented in section 2.3.2 is a prime example of this feature. Whilst the covariance structure obtained from the study data may be identical for both hypothetical models, the interpretation and modelling strategy can be entirely different based on the population model assumed.

Vector geometry plays a substantial role in this thesis in developing an understanding of regression and assessing the impact of collinearity. Rao (1999) utilises the geometry to demonstrate the theory behind the orthogonal projection minimizing the regression sums of squares. In section 2.4.3 the OLS estimate was constructed differently to the standard presentation in the literature. This will be used as an aid in demonstrating the impact of collinearity on the regression estimates. Although geometry is regularly used for illustrative purposes, the interest in this work is the change of the point estimates themselves, along with influences of factors in the model environment. A greater emphasis has been placed on calculating the estimates directly from the geometry. Rather than only discuss the ideas conceptually, this presentation allows for an understanding of the estimators and the role of alternative methods in handling collinear variables.

3. Measuring the Impact of Collinearity

Assumptions were outlined in chapter 2 that define the 'ideal' conditions for the OLS estimator. Under these assumptions, the estimator has uniformly minimum variance in the class of all unbiased estimators – MVUE (Myers 1990). It is also important to consider the impact on desirable accuracy and precision properties of the estimation when these conditions are not met. In chapter 3 the impact of entering linearly dependent covariates into the regression model is considered. The independence of covariates is a desired rather than implicit assumption of OLS (unless they exhibit a perfect dependency – i.e. $\text{rank}(\mathbf{X}^T\mathbf{X}) < k$ - see chapters 7 and 8). The estimator will remain the MVUE when the assumptions are satisfied; however the benefit of minimum variance amongst the unbiased class of estimators can be severely weakened. Subsequently, when the estimator is not precise, the unbiased property is of limited use in a single sample case.

The potential impact of collinearity on the accuracy and precision of an estimator is well recognised in the literature. However, methodology proposed to measure such a feature often overlooks the role of other factors in the study environment. One influence that is regularly ignored is the role of the response. Vector geometry can illustrate the crucial role that the covariance structure between the response and the predictors plays in dictating the impact of collinearity on the parameter estimates. Current diagnostic indices provide a measure of the degree of collinearity present. As such, any remedial action based on such a measure may be unnecessary or misplaced. In some circumstances collinearity can be 'favourable' to the modelling process. Indices should play a crucial role in assessing the impact of collinearity - both adverse and beneficial to the estimation. In this chapter the effects of collinearity in applied regression studies are investigated both on the statistical and interpretational properties of the estimate. An overview is presented of techniques currently in use in applied research along with an investigation of their merits in application using simulation studies. Limitations of existing techniques will dictate the properties desired for a new collinearity index developed in chapter 5.

3.1 What is Collinearity?

In regression analysis, collinearity (or more generally, ill-conditioning) describes the situation in which two predictors are highly correlated with one another. If there exists an approximate or exact linear relation amongst more than two predictors, it is referred to as multi-collinearity. The existence of highly collinear variables in a model does not directly violate the assumptions of least squares regression. The estimates remain unbiased and efficient in their class, but the effects can potentially devalue the analysis or any subsequent conclusions if the impact is disregarded (Freund and Wilson 1998).

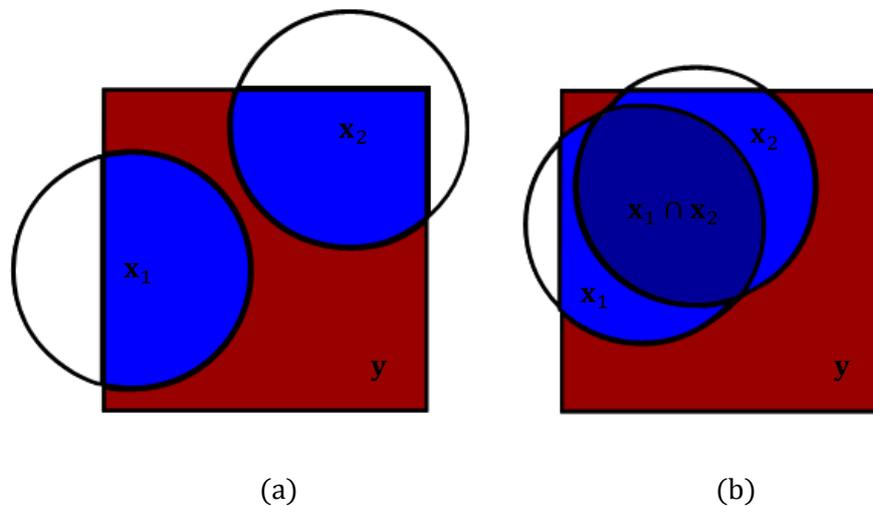


Figure 3.1: Illustrative effects of collinearity on a linear regression model.

The Venn diagrams in Figure 3.1 provide an interpretive illustration of the impact of collinearity. The linear dependence in the model exhibiting collinearity (Figure 3.1b) is illustrated by the relative ‘overlap’ of the covariates. As the levels of linear dependence increase amongst the predictors in a model, it becomes progressively difficult to partition the unique contribution of each predictor to the variation in the response. This results in unstable regression coefficients that are sensitive to changes in both sample observations and model specification. Elevated standard errors are brought about by the increased variability in the estimation of the model coefficient, resulting in a loss of precision.

Whilst the interpretation of the model coefficients can potentially suffer in the presence of multi-collinearity, the predictive power of the model is of less concern. The area of the response (y) overlapped with the predictors in Figure 3.1 represents the coefficient of determination (i.e. R_y^2). That is, the variation in the response accounted for by the predictors in the model. The approximate equality of this measure in both diagrams

illustrates that collinearity is not necessarily ‘harmful’ when the purpose of regression is prediction. In the presence of ‘substantial’ collinearity, the confidence intervals of the parameter estimates will become inflated with the instability of the OLS estimate. However, the user may still feel confident to interpolate, but not to extrapolate from the data. As a result, when the emphasis is placed on prediction in a study, researchers will often be more tolerant of high levels of collinearity amongst the covariates.

There are numerous reasons why collinearity arises in study data and it will almost always be present amongst variables in epidemiological and clinical examples. If collinearity exists amongst the predictors, it will influence the point estimates of the model along with commonly used statistical tests based on standard errors such as p-values, t-tests and F-statistics. To understand how to work with collinearity the impact must first be presented from a statistical perspective. The strict definition of collinearity describes a perfect (or exact) linear relationship between two covariates. For example, there exists a set of parameters c_j (for $j = 0, 1, \dots, k$) such that,

$$c_0 + c_1\mathbf{x}_1 + c_2\mathbf{x}_2 = 0 \quad (3.1)$$

Perfect multi-collinearity occurs when there exists one or more exact linear relationships amongst the covariates in the model (NB. In general, as there is no conceptual difference between “collinearity” and “multi-collinearity”, the former term is often used to describe both cases (Belsley 1991). This practice will be adopted throughout the thesis),

$$c_0 + c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \dots + a_k\mathbf{x}_k = 0 \quad (3.2)$$

When perfect collinearity is present, the design matrix \mathbf{X} is not full rank (i.e. $\text{rank}(\mathbf{X}) < k$). A solution to the OLS estimate cannot be computed using a regular matrix inverse. The rank deficient (or singular) matrix, means that $\mathbf{X}^T\mathbf{X}$ can only be inverted using a generalized inverse procedure. This situation appears extreme and unrealistic in practice, however it is certainly still possible in epidemiological study (e.g. see the lifecourse and APC models discussed in chapters 7 & 8 respectively).

Whilst perfect collinearity is occasionally encountered, a more common situation is ‘near-collinearity’. This is when there exists a high degree of collinearity amongst the covariates. By including the collinear covariates in a regression model, the covariance matrix produced will be near-singular. Consider the following $\mathbf{X}^T\mathbf{X}$ matrix for two covariates centered and scaled to unit variance,

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} \quad (3.3)$$

An estimate of the regression coefficients will still be calculated by OLS, however the determinant of $\mathbf{X}^T\mathbf{X}$ becomes increasingly small as the degree of multicollinearity grows (i.e. $r_{12} \rightarrow 1$).

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{1 - r_{12}^2} \begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix} \quad (3.4)$$

This results in an unstable $\mathbf{X}^T\mathbf{X}$ inverse, meaning that small changes in the measurements of the predictor variables can result in substantial deviations in the point estimates of the regression coefficients. O'Brien (2007) demonstrates the variance of a single predictor b_j ,

$$\text{Var}[b_j] = \frac{\sigma_\epsilon}{(1 - R_{\mathbf{x}_j}^2) \sum \mathbf{x}_j^2} \quad (3.5)$$

Where $R_{\mathbf{x}_j}^2$ is the squared multiple correlation of the j^{th} predictor regressed on the other predictors in the model. In the general case, when \mathbf{x}_j is regressed on covariates to which it shares a linear dependence, the variance of \mathbf{x}_j explained by the predictors will be high. Therefore, the variance of the regression coefficients will increase, indicating that precision is lost in the estimation. From this example it appears that collinearity is solely related to the VC of \mathbf{X} . Whilst it is true that $\mathbf{X}^T\mathbf{X}$ describes the degree of collinearity amongst the predictors, the role of other factors in the regression model can play a substantial part in dictating the 'impact' of this collinearity on the parameter estimates of the model.

3.2 Collinearity in Epidemiology

The impact of collinearity is often overlooked in epidemiological research. The pair-wise covariate associations should be negligible – a most unlikely scenario for biological and epidemiological data. In most epidemiological studies, it will be impossible to physically control for collinearity. It is a natural and unavoidable feature of the data. Small departures from independence can severely distort the interpretation of a model and the role of each covariate – resulting in increased inaccuracy as expressed through the regression coefficients and increased uncertainty as expressed through coefficient standard errors. The variability in the estimation of the regression coefficients can lead to a reduction in the

statistical significance of the coefficient. However, the point estimate may be substantive and therefore clinically significant. Researchers may mistakenly conclude that the statistical insignificance of a variable is due to clinical insignificance rather than collinearity and so ignore the result – yielding an elevated risk of Type II errors. The example commonly used in textbooks to illustrate collinearity is that, within a regression model, the overall variance of the dependent variable explained by the covariates is high, yet none of the covariates are statistically significant (Glantz and Slinker 2001; Kirkwood and Stern 2003). This can occur when the information given by each covariate is greatly ‘overlapped’ with other covariates due to collinearity (see Figure 3.1).

With increased parameter variance, small changes in the data can produce large swings in the parameter estimates. It can typically yield estimates with ‘incorrect’ sign and implausible magnitude (O’Brien 2007). These effects can severely hamper any clinical interpretation in a regression study. The emergence of the PD and the importance placed on causality has given rise to a number of variable types that influence the impact of collinearity. An example was presented in section 2.3.2 that considered the interpretation of collinearity in the confounding and competing exposures case. The distinction between these phenomena can only be made conceptually, as statistically they appear identical (Mackinnon et al. 2000). It should not only be assumed that collinearity can make correlated variables redundant. On the contrary; sometimes collinearity will generate spuriously strong associations between the predictors and the response. However, the interpretation of such coefficients will depend on external evidence and the hypothesized underlying causal structure of the population variables.

3.2.1 Causal Modelling

Correlation amongst variables does not imply causation. However, correlation is a necessary condition for causation, when all external variables are controlled for. After observing an association between an exposure and a disease, an epidemiologist will often wish to determine if any causal relationship exists between the variables. This is rarely a simple task. It is desirable to implement an experimental study design, to physically control for external influences on the main relationship. However, for practical and ethical reasons the randomization of treatments is not possible for a number of research questions. Studies such as RCT’s are generally only an option on a human population when considering potentially beneficial exposures. Whilst experimental studies can be

performed on animals and in vitro systems, there is still the question of how the hypothetical cause-effect relationship operates in humans. Therefore, the analyst is generally restricted to observational non-randomized studies for many research questions. The exposure will likely occur as a natural part of the subject's life – such as the area they live, where they work etc. – and the disease occurrence is recorded. The exposed and non-exposed groups can then be compared. The major issue with this approach (from an epidemiological standpoint) is that without the experimental control of external factors, the researcher may not be able to physically restrict the influence of intermediate variables and confounders. This is likely to be considered an analytical task.

Working within an Epidemiological Framework

The experimental process may begin with an ecological study - which considers analysis on the population rather than the individual. This is particularly useful for developing a causal hypothesis. However, in studying the population, characteristics are often assigned that may not be true of the individual. For instance, let us once again consider the classical cause-effect hypothesis between smoking and lung cancer. If data were gathered for the average cigarette consumption vs. lung cancer rates in ten different countries, a correlation would likely be observed. It would be tempting to conclude that this highlights an association between exposure and disease. However, this could be countered if the cigarette company suggests that it is those in the population that don't smoke that are at higher risk of developing lung cancer. As only averages from a population are available, this argument could not be disproved. An ecologic study will instead present a fast and cheap option to generate a hypothesis for future more detailed studies to assess.

A case-control or cohort study may be performed to attain data at an individual level. This should enable us to counter the initial argument of the cigarette companies in that it can be demonstrated that those individuals with the greatest cigarette consumption show a greater incidence rate of lung cancer – a statistical association between exposure and disease, but not necessarily a causal link. If we are confident that the data presents a real, rather than spurious association (e.g. generated by study design), our interest progresses to examining whether the relationship is causal.

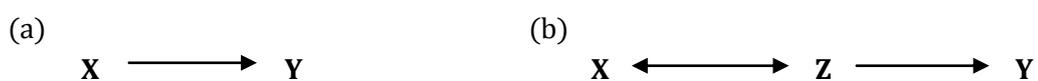


Figure 3.2: (a) A true causal link and (b) an association generated by confounding.

Figure 3.2 demonstrates two path diagrams that would produce a statistical association between **X** and **Y**. The example in Figure 3.2a illustrates a true causal relationship, whereas Figure 3.2b is an association generated by confounding. The distinction between these two examples is purely from a conceptual rather than statistical perspective. If a true causal association is observed between **X** and **Y**, then an intervention on **X** would affect the incidence rate of **Y**. If the association is a consequence of confounding, then an intervention on **X** may have no effect on the risk of disease **Y** (Gordis 2000). The PD defines the clinical hypothesis and the necessary intervention.

In Figure 3.2a, the variable **X** is both a necessary and sufficient cause of disease **Y**. This dictates that the disease could not develop without **X** and will always develop when **X** is present. This is rare in biology. In most circumstances, when subjects are exposed to a risk factor, some are more susceptible than others to develop the disease (e.g. a greater immunity). Similarly, there are often multiple risk factors attributable to the development of a disease. If other risk factors can cause the effect without the presence of **X**, then **X** becomes a sufficient, but not a necessary cause of **Y**. This is again rare, as the guaranteed development of a disease is seldom the product of a single factor. A more likely scenario is that **X** is necessary, but not sufficient to the disease. Alone, **X** may not cause **Y** to develop. However, the existence of **X** along with other factors will produce the disease. Finally, an exposure may be neither necessary nor sufficient. An intricate combination of risk factors may be required, but no specific combination is necessary to produce the disease. This is perhaps realistic of most chronic diseases (Gordis 2000).

A popular way to consider causal mechanisms is to employ the causal pie model (Rothman and Greenland 2005). Each pie represents a sufficient causal mechanism of a disease. The segments of the pie represent a component of the overall biological mechanism. For instance, the 'disease' may be metabolic syndrome (MetS), which is defined as a clustering of risk factors that are associated with a greater risk of developing cardiovascular disease and diabetes. The diagnosis of MetS is typically based on a combination of risk factors. For instance, the international diabetes federation (IDF) require (A) central obesity, along with two of (B) raised triglycerides, (C) reduced cholesterol, (D) raised blood pressure and (E) raised fasting plasma glucose. Risk factors are not all considered of equal importance. This is generally defined by the impact of each component on the incidence of the disease. The 'importance' of a component may change depending on the prevalence of others in the mechanism. To determine 'importance' is thus the role of statistics, rather than biology (Rothman and Greenland 2005).

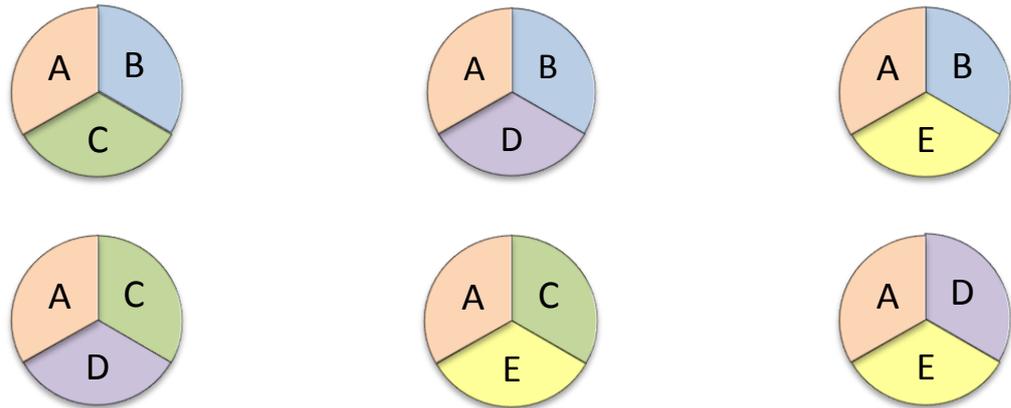


Figure 3.3: A causal pie illustration of MetS based on the IDF definition.

To characterize the IDF standards, there are six unnecessary but sufficient causal mechanisms to diagnose MetS (see Figure 3.3). Each consists of three ‘component causes’ that when occurring together are sufficient for the development of the disease. The ‘interactions’ of multiple component causes are common of most chronic diseases and rarely will a necessary cause be defined. Central obesity in the IDF definition is itself a necessary cause as it appears in every sufficient causal pie. Apart from central obesity, the other components are neither sufficient nor necessary to cause the disease. The blocking or eliminating of one will not prevent the disease, however each plays an important role in a causal mechanism. Mackie (1974) labels this the “INUS condition”.

The causal pie model represents a simplistic way to understand a potentially complex ‘causal web’. However, this methodology only partly enters the complexities of causal theory. For instance, more sophisticated models, such as DAG’s and SEM’s, have provided an added level of intricacy to the diagrams to better reflect the nature of ‘real world’ relationships (see section 2.3.2). The first consideration is that the causal pie implies that the ‘interactions’ amongst component causes are occurring at one point in time. In fact, a causal process may (and often will) occur as part of a temporal sequence of events. The use of directed arrows in DAG’s allows for a sequential arrangement of the causal relationships (i.e. generally left to right). Another benefit of DAG’s is in demonstrating the existence and role of intermediate variables and confounders.

The causal pie model represents a probabilistic approach to causality, whereas the DAG and SEM are labelled counterfactual approaches. These represent two different philosophical views. The probabilistic approach suggests that if A causes B, then the presence of A will increase the likelihood of B occurring. This is conceptually appealing for

epidemiology, as data is often incomplete or the nature of the mechanism in itself would seem probabilistic; e.g. smoking will in some, but not all cases, result in the patient developing lung cancer. Therefore, smoking can be labelled a probabilistic cause of lung cancer. In contrast, the counterfactual view of causality is that all things being equal - the presence of A will cause B. However, epidemiological study can rarely (if ever) ensure that “all things are truly equal”. The probabilistic argument is a ‘loose’ definition of the population model. In Rothmans pie model, the causal relationships amongst the variables are not specified. Only the likelihood of disease occurrence increases with the presence of component causes.

The DAG and SEM present a counterfactual version of causality as the models are not built on probability of disease occurrence, but specify direct causal relationships. The complexity of the model is raised; and with this, some of the uncertainty of probabilistic causality is removed. Each effect denoted by an arrow represents a hypothesized ‘true’ relationship, but in the study data this effect is moderated by biological and environmental factors. In approaching the complexity of a ‘real world’ system, the statistics and subsequent interpretation becomes progressively difficult. Whilst the pie model leaves many questions unanswered, the conceptual understanding of the causal mechanism still provides a valuable tool in application. For the discussion in this work, the DAG provides a means to demonstrate the statistical complexities of each application to study data. However, debate still remains as to the study of causality.

3.2.2 Understanding Third Variable Effects

MacKinnon et al. (2000) present three “third variable” examples in mediation, confounding and suppression where the addition of a third variable generates an effect on the main relationship. In each, it is the conceptual understanding of the phenomena that distinguishes the effects. The notion of a confounder was previously discussed in chapter 2 - this is once again considered along with the introduction of a ‘mediator’ variable.

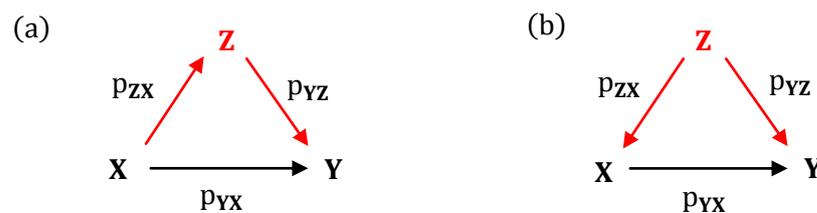


Figure 3.4: Third variable effects in which Z is (a) a mediator and (b) a confounder.

Mediation and Confounding

The first effect to be considered is that of a mediator (see Figure 3.4 a). Two causal paths can be defined from exposure to disease. The first is a path that links X directly to Y - where X is a 'direct cause' of Y as there exist no intermediate steps. The second is an 'indirect effect' that passes through the mediator Z . As demonstrated in the DAG, the indirect effect is defined by X being a cause of Z and Z a cause of Y .

A path coefficient indicates only the direct effect of a cause on an effect. In Figure 3.4, the path coefficients are denoted by P with a subscript ordering the effect to precede the cause.

$$X = e_x \quad (3.6)$$

$$Z = P_{ZX}X + e_z \quad (3.7)$$

$$Y = P_{YX}X + P_{YZ}Z + e_y \quad (3.8)$$

The exogenous variable X (i.e. has no arrows pointing toward it) is only determined by external factors and error - represented by e_x (this does not include measurement error due to the implicit regression assumption **A4** - see section 2.2.2). The equation for Z demonstrates that it is formed from X (as determined by the path coefficient weighting) and also specific unexplained factors and error contained in e_z . The endogenous variable Y (i.e. that has only arrows pointing toward it) is formed partly from X and from Z , in addition to the specific error - e_y . Substituting these relationships into the general correlation formula produces the following definitions,

$$\begin{aligned} r_{XY} &= \frac{\sum XY}{n} = \frac{\sum X}{n} (P_{YX}X + P_{YZ}Z) = P_{YX} \frac{\sum X^2}{n} + P_{YZ} \frac{\sum XZ}{n} \\ &= P_{YX} \text{Var}(X) + P_{YZ} r_{XZ} = P_{YX} + P_{YZ} r_{XZ} \end{aligned} \quad (3.9)$$

$$r_{ZY} = \frac{\sum ZY}{n} = \frac{\sum Z}{n} (P_{YX}X + P_{YZ}Z) = P_{YZ} \frac{\sum Z^2}{n} + P_{YX} \frac{\sum XZ}{n} = P_{YZ} + P_{YX} r_{XZ} \quad (3.10)$$

$$r_{XZ} = \frac{\sum XZ}{n} = \frac{\sum X}{n} (P_{ZX}X) = P_{ZX} \frac{\sum X^2}{n} = P_{ZX} \quad (3.11)$$

From eqn(3.11) it is clear that the path coefficient from X to Z (i.e. P_{ZX}) is equal to the correlation between the variables (r_{XZ}). This is a result that is always true when the child

(i.e. **Z** in this case) is only preceded by one parent (i.e. **X**) – i.e. a ‘simple’ regression. Subtracting $P_{YZ} r_{XZ}$ from both sides in eqn(3.9) gives,

$$P_{YX} = r_{XY} - P_{YZ} r_{XZ}$$

Substitute this into eqn(3.10) to gain the following,

$$\begin{aligned} r_{ZY} &= P_{YZ} + (r_{XY} - P_{YZ} r_{XZ})r_{XZ} \\ r_{ZY} - r_{XY}r_{XZ} &= P_{YZ}(1 - r_{XZ}^2) \end{aligned}$$

Which simplifies to the result in eqn(3.12).

$$P_{YZ} = \frac{r_{ZY} - r_{XY}r_{XZ}}{1 - r_{XZ}^2} \quad (3.12)$$

Therefore, the path coefficient of **Z** to **Y** is equivalent to the standardized beta weight from a regression model involving two predictors and a response (i.e. the regression coefficient b_z from the model $Y \sim X + Z$) – see eqn(2.29). Similarly, the path coefficient P_{YX} is the partial regression coefficient of **X** from the same model. The association between **X** and **Y** can be partitioned into the contribution of the main effect **X** and a mediator **Z**,

$$r_{XY} = \beta_X + \beta_Z r_{XZ} \quad (3.13)$$

Assuming all relations are linear and additive, the direct effect is simply calculated as the path coefficient P_{YX} . The indirect effect is the product of the coefficients P_{ZX} and P_{YZ} . The total effect is the summation of the direct and indirect effects (i.e. $P_{YX} + P_{ZX}P_{YZ}$). Rearranging eqn(3.10) for P_{YX} presents the following result,

$$P_{YX} = r_{XY} - P_{YZ}r_{XZ} \quad (3.14)$$

Insert eqn(3.11) and eqn(3.9),

$$P_{YX} + P_{ZX}P_{YZ} = (r_{XY} - P_{YZ}r_{XZ}) + r_{XZ}P_{YZ} = r_{XY} \quad (3.15)$$

The total effect is the unadjusted association of **X** on **Y** (i.e. the standardized coefficient from a simple regression model). The total effect is the coefficient of **X** with a contribution of the mediator. By adjusting for the mediator in the model, the regression coefficient of **X** is only the direct effect P_{YX} (or the total effect minus the indirect effect).

In the confounding case, the equations are adjusted as the direction of causality is reversed between \mathbf{X} and \mathbf{Z} - i.e. the confounder (\mathbf{Z}) becomes the exogenous variable,

$$\mathbf{X} = P_{ZX}\mathbf{Z} + e_x \quad (3.16)$$

$$\mathbf{Z} = e_z \quad (3.17)$$

$$\mathbf{Y} = P_{YX}\mathbf{X} + P_{YZ}\mathbf{Z} + e_y \quad (3.18)$$

This change in the structure of causality has no impact on the statistical computation of the path coefficients, only the conceptual understanding of the relationships. However, in the confounding case the difference between the total effect and the adjusted effect represents an estimate of confounder bias (Mackinnon et al. 2000).

The Geometry of Third Variable Effects

Statistical adjustment for a single ‘true confounder’ will reduce the bias of the point estimate toward the coefficient in the population model. Whereas, adjustment for a mediator will introduce bias to the estimation as it lies on the causal path. In both cases, the inclusion of the third variable in the model will generally reduce the magnitude of the coefficient from the simple regression. A mediator will explain part or even all of the association between exposure and effect as it shares the causal path. In confounding, a reduced coefficient will typically occur when “confounder bias” is removed from the model. Therefore, by controlling for the confounding (either physically or analytically), the user will generally reduce the potential of Type I error from a spurious relationship.

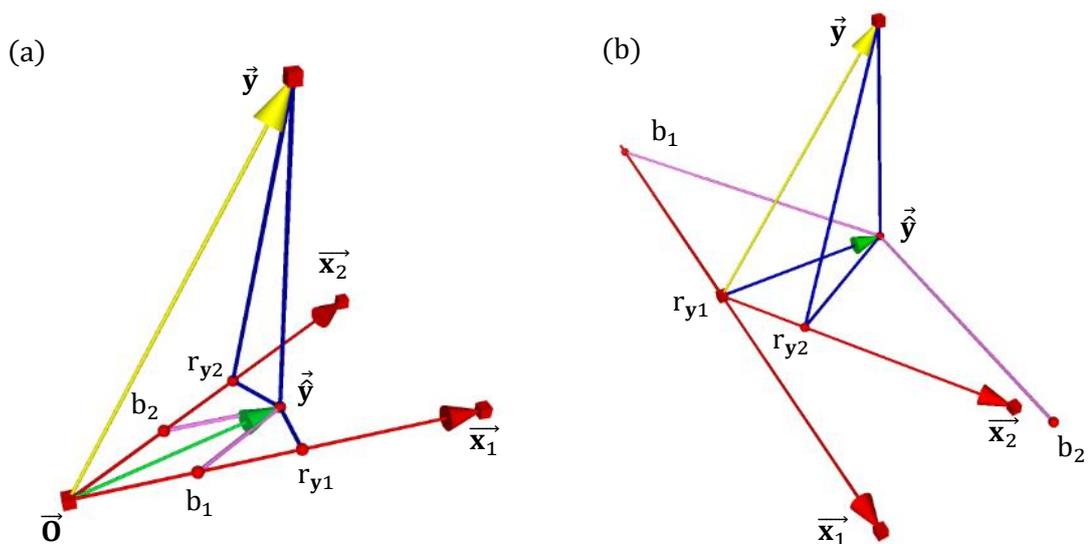


Figure 3.5: Statistical interpretation of (a) a confounding and (b) suppression effect.

Whilst the user may expect to see a reduced coefficient estimate when a high degree of collinearity is present amongst predictors, this should not be assumed. A suppression effect can inflate point estimates in both examples – see Figure 3.5 b. Consider the following example,

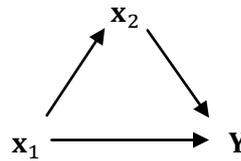


Figure 3.6: An example path diagram.

Standardized regression coefficients are as follows for the bivariable model,

$$\begin{aligned}
 b_1 &= \frac{r_{y1} - r_{y2} \cdot r_{12}}{1 - r_{12}^2} \\
 b_2 &= \frac{r_{y2} - r_{y1} \cdot r_{12}}{1 - r_{12}^2}
 \end{aligned}
 \tag{3.19}$$

Consider in a sample that a non-significant positive correlation is observed between x_1 and y (i.e. r_{y1}). In addition, the correlations r_{12} and r_{y2} are both found to be highly positive. If x_1 is considered to be the main effect and y to be the disease, then x_2 is defined a mediator. The value of the numerator in eqn.3.19a is likely to be negative (due to the negligible r_{y1}), whilst the denominator will be close to zero. This generates an inflated negative coefficient on b_1 . The coefficient b_2 is then enhanced and remains positive following the inclusion of the third variable.

The total effect in the mediation example is defined as the sum of the direct and indirect effects. If these were of opposite sign and equal magnitude, then adjusting for the third variable in the model would result in a “complete suppression” (Mackinnon et al. 2000). The direct and indirect effects cancelling each other, resulting in a statistically insignificant total effect. A suppression may also be termed “inconsistent mediation” or “negative confounding“, depending on whether the user wishes to be specific to the conceptual model employed. Depending on the values of r_{12} and r_{y2} , inclusion of x_2 in the model can enhance an already negative b_1 or can reverse the sign of a positive coefficient to negative. Even when the third variable has zero effect in the population, a single sample case is still likely to be interpreted as either a mediation/confounding or suppression due to sampling variation. Statistical insignificance and ‘incorrect’ hypothesized signs of the coefficients are potential ‘symptoms’ of collinearity, but neither is sufficient nor necessary for the presence of collinearity (Belsley 1991).

3.2.3 Development of a Causal Hypothesis

Consider again the two examples in Figure 3.4. If the path diagram shows **Z** to be a confounder then it should be controlled for. In theory, this decision can be made easily as **Z** is not on the causal pathway. Therefore, confounding bias is removed and the precision of the estimation for the direct effect improved. In contrast, the mediation example is more complex. The mediator **Z** will lie on the causal path between exposure and disease. Therefore, theoretically it is part of the total effect and shouldn't require adjustment. However, if in the sample no association is observed between **X** and **Y** (and this is repeatedly observed across studies), then should it be suggested that **X** is not a cause of **Y**? This is where the separation of causality and correlation becomes important. In almost any biological relationship there will be multiple unobserved intermediate steps (Gordis 2000).

A causal link between parental smoking and low birth weight is almost universally accepted. However, there will be numerous chemical reactions and genetic relationships that define the nature of this causal mechanism. This may not appear important in situations in which a desired association is observed which fits with biological theory; however in other circumstances the analyst is forced to take notice. Consider the classical birth-control example presented by Judea Pearl (2000),

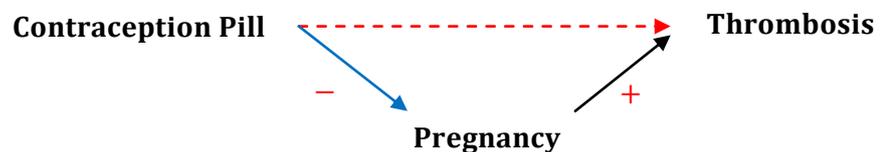


Figure 3.7: Classical birth Control example from Pearl (2000).

The study considers the effect of the contraception pill on the development of thrombosis. Evidence suggests that the pill reduces pregnancy and that pregnancy in turn increases the risk of thrombosis. Therefore, the indirect effect from exposure to disease is negative. To address the research aim we wish to know the effect of the pill on thrombosis whilst the mediation effect of pregnancy is held constant. If there were no direct effect, statistically adjusting for such a variable would create a spurious association between the pill and thrombosis (as pregnancy is on the causal pathway). A causal relationship could not be determined analytically and could potentially be misinterpreted as the presence of a direct relationship. The mediation effect of pregnancy must be physically controlled for as part of the study design (if practical limitations allow).

In this example, the groups under study should be women who used the pill post pregnancy and those who used other means than the pill to prevent pregnancy (Pearl 2000). This would block the causal relation between the pill and pregnancy, thus removing the mediated effect. Consider the addition of a fourth variable labelled **Z**,

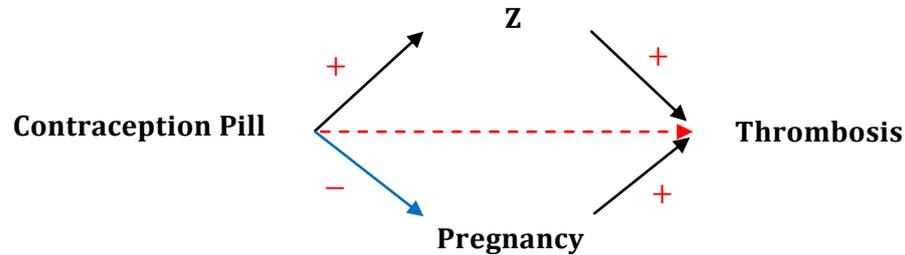


Figure 3.8: Presence of an additional mediator to the example.

In this example the hypothesized direct relation is mediated by an intermediate chemical mechanism, which in itself could contain a number of chemical reactions etc. Many of the direct effects stated in these hypothetical models are in fact a simplification of a complex set of indirect effects. The path coefficients leading to and from **Z** are both positive in this example and could cancel the negative indirect effect of pregnancy. Therefore, the total effect could be statistically insignificant. In isolation it may not be of great importance to replace the direct effect with an intermediate relation. However, the temptation to simplify the model by ignoring such factors can mislead the clinical interpretation. Causal theory would suggest not statistically adjusting for pregnancy, but if multiple pathways are present, the total effect can become difficult to interpret.

Demonstrating the effects of mediators, confounders and suppressors provides an insight into the complexities of the causal system. Whilst the causal pie model allows an overview of the sufficient mechanisms, the methodology is limited in understanding the complexity of the structure. The DAG allows the user to develop an understanding of the pathways and mechanisms at work in real world applications. The primary concern of this section is to demonstrate the epidemiological relevance of collinearity in the study data. The problems are complex on a statistical level, however they take on a new meaning when placed in a clinical context. This should be reflected in the understanding of collinearity and its effects. The discussion of causality discourages the notion that correlation in the sample must be a 'problem'. This is crucial as the discussion progresses towards measures of collinearity, often labelled collinearity 'diagnostics'. It is important to decide whether the index is 'diagnosing' a 'disease' or looking to understand the potential impact of collinearity, either detrimental or beneficial to the modelling process.

3.3 Collinearity Diagnosis

In epidemiological and clinical research, it is not surprising to find that many covariates are correlated, as they often share common physiological mechanisms, or measure different aspects of the same underlying mechanism. The question is not whether collinearity is an issue, but what the impact is on the modelling process. Indications that a collinearity effect is present (such as a change of sign or an inflated variance of the coefficients) must not be relied upon as a measure of collinearity. The suppression effect demonstrates that commonly experienced symptoms of collinearity are not necessary for such an effect to be present. Collinearity can also be beneficial to the estimation. Collinearity diagnostics have been developed to provide an indication of the 'severity' of collinearity in an applied regression study. They can provide a range of information, such as which variables are most highly involved in collinear relations and with which others they are linearly related to. This is intended to aid with model specification and to indicate an appropriate caution to place on the results of an analysis.

It is important to consider the balance between statistical and clinical knowledge. In section 2.3.2 two phenomena were considered that are statistically indistinguishable, but conceptually very different. Therefore, any statistical measure of collinearity must be balanced with a guided conceptual understanding of the model environment and application. For instance, a researcher can strive toward obtaining a set of predictors that minimize collinearity in a model, but if there is no theoretical basis for the variables chosen then the work is of limited use. Instead we must look towards an assessment of the impact of collinearity to guide the researcher with its likely effects on the analysis of the model. The index should assess the impact of collinearity on the model parameters and inform the user whether implementing OLS is the appropriate regression technique for the dataset. Myers (1990) suggests that the diagnostic tool should never be thought of as a guide toward whether an alternative technique will be successful, but rather as an indicator of the inefficiency of OLS. It will be important to consider features in the model environment, such as measurement error and sample size. In particular, the role of the response can play a key role in moderating the impact of collinearity on the variability of the model coefficient and the point estimate. These effects can both enhance and diminish the impact of collinearity – both potentially harmful to the interpretation and the clinical conclusions formed from the analysis. These features can have substantial damaging effects on inferences when the impact is misunderstood.

3.3.1 Developing a Collinearity Index

Variables considered to be clinically ‘important’ to the outcome, may produce insignificant t -values and p -values. Whilst this is a common feature of collinear data, it can also feasibly occur in non-collinear data. In contrast, a suppression effect can produce a coefficient with a much inflated statistical significance. The same applies for other common results of collinearity, such as changes of sign resulting from inclusion or exclusion of covariates. Another crude approach is to study the correlation matrix of the dataset directly. The thought behind this stems from the definition of collinearity - that there is a ‘significant’ correlation between two covariates. Although viewing high bivariate correlations can be seen as an indicator of collinearity, a lack of such correlations cannot equally be considered a sign that collinearity is not present. Multi-collinearity describes linear combinations amongst two or more covariates, which may not be seen directly from the correlation matrix. It is possible that covariates have low bivariate correlations, but there exists an underlying linear relationship amongst multiple covariates. This reasoning has led to the development of a number of collinearity diagnostics. However, the results of such indices are often extrapolated to explain features beyond the limitations of the diagnostic. The motivation of the method may not be fully understood and regression models are deemed to have ‘acceptable’ or ‘unacceptable’ levels of collinearity based on overly simplistic approaches and arbitrary ‘rules of thumb’.

3.3.2 Approaches to Diagnosing Collinearity

Belsley (1991) presents a useful summary of the current methods available to the researcher for general use as diagnostics for collinearity. These range from very simplistic approaches (such as the presence of naïve symptoms associated with collinear data discussed in section 3.1), to complex diagnostic tools that acquire a heavy computational cost (one such method will be developed in chapter 6). Two popular techniques in the variance inflation factor and condition index are first considered which are often applied in the regression literature. The methods developed in later sections of this thesis to study collinear data build upon these approaches and that of alternative estimators to OLS. The indices are both based on the dispersion matrix of the predictors (i.e. $\mathbf{X}^T\mathbf{X}$), but adopt different approaches in their assessment of collinearity amongst covariates.

The Variance Inflation Factor

The statistical discussion of collinearity in section 3.1 provides the basis for one of the simplest and most popular collinearity indices - the variance inflation factor (VIF) (Marquardt and Snee 1975). The theory of the VIF can be considered by presenting the variance of the estimated regression coefficient.

$$\text{Var}[b_j] = \frac{s_\varepsilon^2}{(1 - R_{x_j}^2) \sum (x_j - \bar{x}_j)^2} \quad (3.20)$$

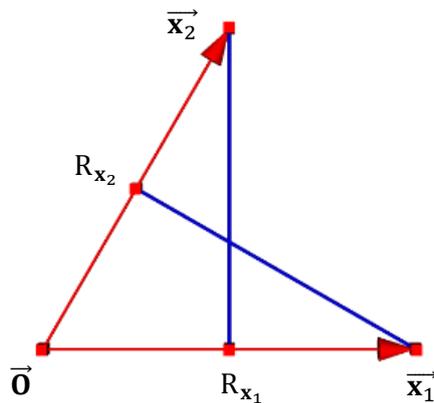
$$\text{VIF}_j = \frac{1}{1 - R_{x_j}^2} \quad (3.21)$$

VIF_j is the j^{th} diagonal element of the inverse correlation matrix of \mathbf{X} (Hocking 2003). As the degree of correlation between the j^{th} predictor and the remaining covariates increases, the value of $R_{x_j}^2$ will increase. If there were no correlations between the covariates this factor would disappear. Therefore, the quantity of $(1/1 - R_{x_j}^2)$ can be seen as an inflation term of the sample variance when collinearity is present in the model.

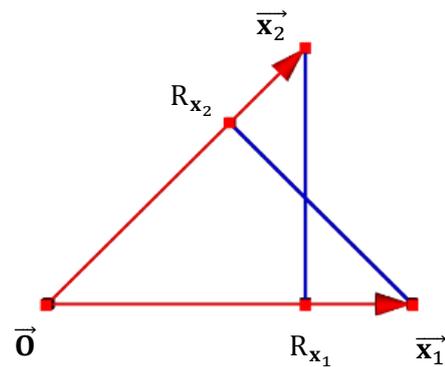
The method can be illustrated by considering the vector geometry of the VIF. Figure 3.9 demonstrates the geometrical construction of R_{x_j} for two predictors,

e.g.1 $\theta_{12} = \pi/3$, $\text{cor}(x_1, x_2) = 0.5$

e.g.2 $\theta_{12} = \pi/4$, $\text{cor}(x_1, x_2) = 0.71$



$$\text{VIF}_j = \frac{1}{1 - (1/2)^2} = \frac{4}{3}$$



$$\text{VIF}_j = \frac{1}{1 - (1/\sqrt{2})^2} = 2$$

Figure 3.9: An illustration of the construction of R_{x_j} using vector geometry.

R_{x_1} is found by an orthogonal projection of \vec{x}_2 onto \vec{x}_1 . In Figure 3.9, the vector \vec{x}_1 is assigned to lie on the standard x-axis in variable space, therefore R_{x_1} is identical to the x-coordinate of \vec{x}_2 (i.e. a_1 in the vector geometry of section 2.4.3).

$$R_{x_j}^2 = \cos^2(\theta_{12}) \quad (3.22)$$

$$VIF_j = \frac{1}{1 - \cos^2(\theta_{12})} \quad (3.23)$$

The trend of the VIF can be modelled by generating a simulation for two covariates x_1 and x_2 . Although in reality there are two VIF values, they are identical in the bivariable model as Figure 3.9 shows. Figure 3.10 displays the change in VIF as the correlation between the covariates increases (i.e. r_{12}).

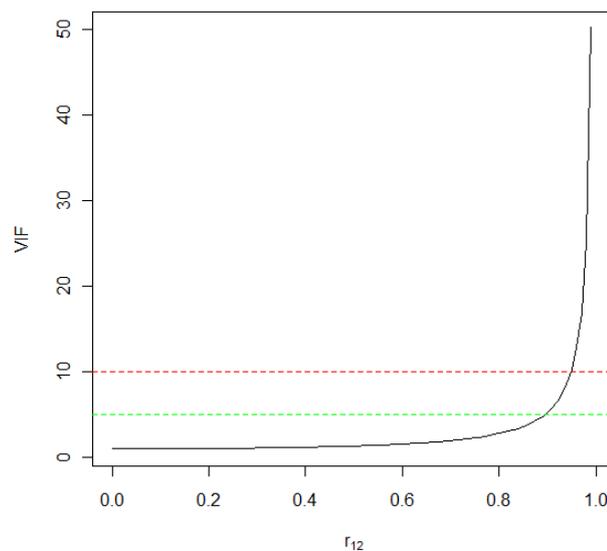


Figure 3.10: Trend of the VIF function in the bivariable regression case.

The VIF provides a measure of the impact of collinearity on the precision of individual regression coefficients (Fox and Monette 1992). Belsley discusses two common ‘rules of thumb’ to indicate ‘severe’ collinearity amongst predictors. Thresholds that are often suggested include a $VIF = 5$ and $VIF = 10$ (these have been indicated in Figure 3.10 by the green and red dashed lines respectively). Although the VIF trend becomes increasingly difficult to map in higher dimensional cases (i.e. with additional predictors included in the regression model), the bivariate example here demonstrates that at these arbitrary thresholds the variance begins to increase steeply.

The Singular Value Decomposition and the Condition Index

The condition index (CI) is a collinearity diagnostic regularly employed in regression studies. The theory can be demonstrated using an important result from matrix algebra - that any $n \times k$ matrix \mathbf{A} can be decomposed into the following set of matrices,

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T \quad (3.24)$$

where \mathbf{U} is an $n \times k$ matrix with orthonormal columns (i.e. the columns are the left singular vectors of \mathbf{A}), \mathbf{D} is a $k \times k$ diagonal matrix with non-negative elements and \mathbf{V} is a $k \times k$ orthogonal matrix (i.e. the columns are the right singular vectors of \mathbf{A}). This is a singular value decomposition (SVD). The matrix \mathbf{A} can be replaced with the mean centered $\mathbf{X}^T\mathbf{X}$ to gain the following SVD form,

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \quad (3.25)$$

Where \mathbf{V} contains the columns of eigenvectors \mathbf{v}_j of $\mathbf{X}^T\mathbf{X}$ and the diagonal elements of $\mathbf{\Lambda}$ (i.e. \mathbf{D}^2) are the corresponding eigenvalues λ_j . To attain the VC matrix of the mean centred \mathbf{X} , the $\mathbf{X}^T\mathbf{X}$ matrix should be divided by $n - 1$, which scales only the eigenvalues. Therefore, these are the eigenvectors of the VC matrix of \mathbf{X} . A full complement of non-zero eigenvalues can only be found if $\mathbf{X}^T\mathbf{X}$ is non-singular (i.e. full rank). They adhere to the characteristic equation,

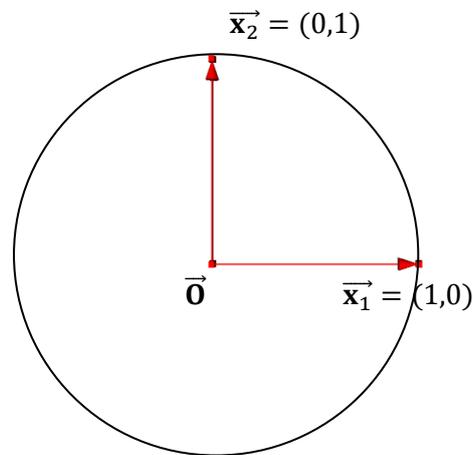
$$\det[\mathbf{X}^T\mathbf{X} - \lambda_j\mathbf{I}] = 0 \quad (3.26)$$

The λ_j can then be used to obtain the corresponding eigenvectors \mathbf{v}_j ,

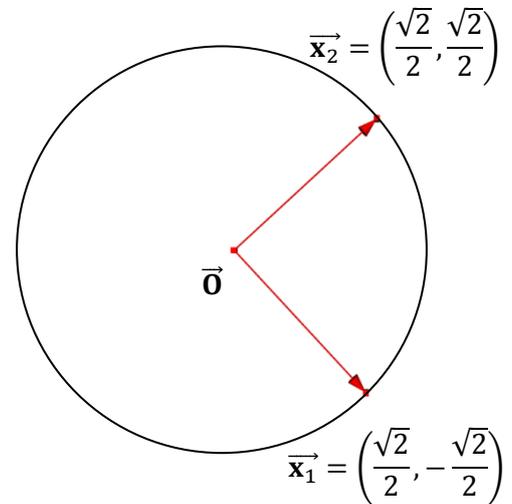
$$[\mathbf{X}^T\mathbf{X} - \lambda_j\mathbf{I}] \mathbf{v}_j = 0 \quad (3.27)$$

The set of \mathbf{v}_j and λ_j that give a non-trivial solution are the k eigenvectors and corresponding eigenvalues of $\mathbf{X}^T\mathbf{X}$. The benefit of obtaining the eigen-decomposition for collinearity diagnosis becomes apparent when we consider the vector geometry of the SVD. For this illustration eg. 2 is retained from the VIF geometry – i.e. a correlation of $\sqrt{2}/2$ between \mathbf{x}_1 and \mathbf{x}_2 . Geometrically, the effect of the eigenvalues and eigenvectors can be considered on an uncorrelated pair of covariates (i.e. a unit disc). The SVD performs two rotations by the eigenvectors either side of a scaling by the eigenvalues.

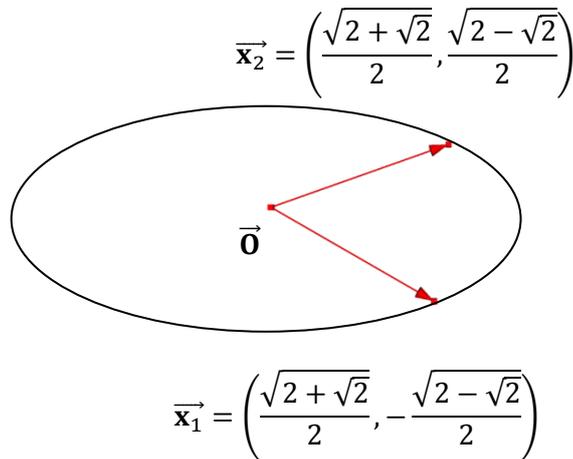
(a) Orthogonal Axes



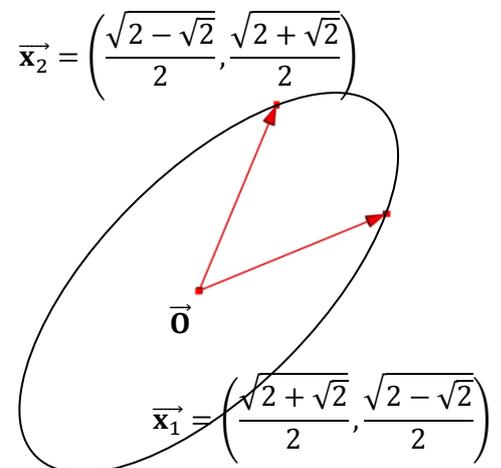
$$\mathbf{I}_{1 \times 1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

(b) First Rotation by \mathbf{V} 

$$\mathbf{V} = \begin{bmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}$$

(c) Scaling by Λ 

$$\Lambda = \begin{bmatrix} (2 + \sqrt{2})/2 & 0 \\ 0 & (2 - \sqrt{2})/2 \end{bmatrix}$$

(d) Second Rotation by \mathbf{V} 

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} 1 & \sqrt{2}/2 \\ \sqrt{2}/2 & 1 \end{bmatrix}$$

Figure 3.11: Illustration of the SVD decomposition of the VC matrix.

The process demonstrated in Figure 3.11 is labelled a principal axis transformation (Jackson 2003). The eigenvectors \mathbf{v}_j represent the direction cosines, demonstrating the angle between the old and new axes - i.e. $\cos^{-1}(\pi/4) = \sqrt{2}/2$. Therefore, $\vec{\mathbf{x}}_1$ is at 45° from the original $\vec{\mathbf{x}}_1$ axis and similarly at 45° from the original $\vec{\mathbf{x}}_2$ axis. This is demonstrated by elements v_{11} and v_{12} in \mathbf{V} respectively. The initial eigenvector rotation centres the main direction of variation on the original $\vec{\mathbf{x}}_1$ vector. The data circle is then scaled along the major and minor axes to represent the variation in the data. The final rotation centres the vectors on the \mathbf{x}_2 axis in (b) as the variation in this example is distributed evenly between the variables (i.e. $sd_{\mathbf{x}_1} = sd_{\mathbf{x}_2}$). The set of rotated axes are labelled "principal axes".

The eigenvalues scale the data circle (i.e. the identity matrix) to create a data ellipse when there exists correlation amongst covariates (Figure 3.11 c). The first eigenvalue gives a measure of the variance in the direction of greatest variance (i.e. the first principal axis). The second eigenvalue provides a measure of the variance of the axis orthogonal to the first and so on. Therefore, in an example with two collinear covariates, λ_1 will be relatively large in comparison to λ_2 . As the correlation between \mathbf{x}_1 and \mathbf{x}_2 grows, the difference between λ_1 and λ_2 will subsequently increase. As λ_2 decreases it is indicating the growing redundancy of the second dimension in the data. Thus, indicating a greater dependency amongst the predictors. The definition of what constitutes a 'small' eigenvalue becomes the issue from a diagnostic perspective (particularly when $k > 2$). Belsley (1991) suggests that the problem arises due to trying to identify a small eigenvalue relative to zero, when an improved suggestion is to define it against others in the study. The CI is a way of presenting the information gained from eigenvalues. Through considering eigenvalues as a ratio to the maximal variance, the thresholds become appropriate to the individual study.

$$CI_j = \frac{\lambda_{\max}}{\lambda_j} \quad (3.28)$$

Similar to the VIF, there are popular thresholds that researchers choose to use as an indication of severe multicollinearity. Belsley suggests 1-10 as an indicator of weak collinearity and 30-50 for moderate to high collinearity. However, the range of thresholds employed suggests that many lack valid statistical justification. They are arbitrary values suggested for the benefit of the user, but caution should be taken when placing too much emphasis on their results.

3.3.3 Application of Collinearity Indices

The VIF and CI are based exclusively on the dispersion matrix of the covariates (as illustrated in the geometrical examples). Therefore, they are labelled ‘correlation based’ diagnostics. These inform the user of the degree of collinearity amongst the covariates, but nothing of further factors involved in the model, thus neglecting to indicate its potential *impact* on the model parameters. If the user solely intends to assess the collinearity amongst the predictors, then measures based on correlation or covariance alone will be adequate. However, to make justifiable decisions on model specification, the user will require a greater detail of the model environment in which the collinearity exists. Prior knowledge and model features will dictate this impact and its subsequent interpretation.

The VIF and CI are popular diagnostic measures and both have strong interpretational benefits for the users understanding of collinearity. However, both have limitations with regard to our original aims for a diagnostic measure. For example, whilst the VIF can indicate the contribution of a variable to the presence of collinearity, it will not specify how many dependencies are present nor which variables are involved in particular dependencies. A further issue is with the promotion of a ‘rule of thumb’ to denote a ‘significant’ dependency. Introducing a ‘rule of thumb’ suggests a point at which the researcher should ‘act’ and try to remedy the ‘problem’. This misguided view of the model environment demonstrates a misunderstanding of the role of collinearity (in both a beneficial or adverse nature) and is dangerous for regression studies in epidemiology.

The Condition Index

The CI addresses some of the limitations of the VIF, whilst it still shares the common restrictions of a ‘correlation based’ index. An eigenvalue ‘close’ to zero will demonstrate the presence of a near-collinear relation and in turn a redundant dimension in the data. The issue of what constitutes a “small” eigenvalue is somewhat eased (but not negated) by adopting a ratio that places the eigenvalue in proportion to the largest in the study. The use of eigenvectors in partnership with the eigenvalues are sometimes used (primarily in methods such as principal components analysis – see chapter 4) to indicate which variables play a role in the strongest dependencies, but will not provide a complete solution. This index can be seen as an improvement to the VIF as a standalone diagnostic, but caution should still be placed in its application.

Recall that the eigenvector represents a rotation of the axes in subject space to centre them in the direction of maximal variance. Extending the example to three dimensions would produce an ellipsoid shaped data cloud (without perfect collinearity),

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1 & 0.95 & 0.05 \\ 0.95 & 1 & 0.05 \\ 0.05 & 0.05 & 1 \end{pmatrix}$$

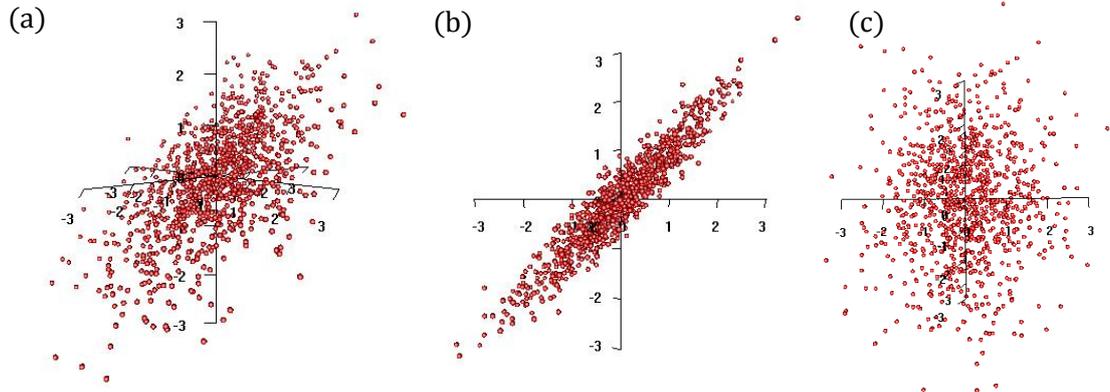


Figure 3.12: Results of a simulation from the theoretical correlation matrix.

The final eigenvalue, measuring the ‘flattest’ part of the ellipsoid, would indicate the linear dependency between \mathbf{x}_1 and \mathbf{x}_2 (i.e. an approximately redundant dimension). Figure 3.12c illustrates a random scatter of points (i.e. almost circular) demonstrating the independence of \mathbf{x}_3 with \mathbf{x}_1 and \mathbf{x}_2 . The correlation matrix produces the following SVD,

$$\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T = \begin{bmatrix} -0.71 & 0.01 & 0.71 \\ -0.71 & 0.01 & -0.71 \\ -0.01 & -1 & 1.97 \times 10^{-17} \end{bmatrix} \begin{bmatrix} 1.95 & 0 & 0 \\ 0 & 0.99 & 0 \\ 0 & 0 & 0.05 \end{bmatrix} \begin{bmatrix} -0.71 & -0.71 & -0.01 \\ 0.01 & 0.01 & -1 \\ 0.71 & -0.71 & 1.97 \times 10^{-17} \end{bmatrix}$$

Each eigenvector has been normalized to give a proportional weight to the original \mathbf{X} . Observe that \mathbf{x}_3 is approximately orthogonal to the third principal axis. This is to be expected as the variance in \mathbf{x}_3 is being explained almost entirely by the second principal axis (i.e. no rotation of \mathbf{x}_3 – i.e. $\cos(0) = 1$; \mathbf{x}_1 and \mathbf{x}_2 axes are orthogonal – i.e. $\cos(\pi/2) = 0$). Orthogonality between the original axis and the principal axis would indicate that it plays no part in the dependency. Therefore, it would seem natural to assume ‘large’ elements in the eigenvector with a ‘small’ associated eigenvalue, would indicate variables involved in a linear dependency. However, it should not be assumed that small elements in the eigenvector demonstrate that the variables are not involved in the relationships (Belsley 1991).

This misinterpretation arises from the use of the word ‘small’. A zero element in the eigenvector would naturally dictate that the variable is not involved in the dependency. However, a small element can still be involved in such a relation as demonstrated by a useful counter-example in Belsley (1991). The following relationship defines three covariates in a model,

$$\mathbf{X}_2 = -\frac{\mathbf{X}_1 + Q\mathbf{X}_3}{L} \text{ where } L = \|\mathbf{X}_1 + Q\mathbf{X}_3\| \quad (3.29)$$

The system is computationally singular (i.e. final eigenvalue will be zero). The eigenvector corresponding to the zero eigenvalue will be of the following form before scaling,

$$\mathbf{v}_3 = \begin{bmatrix} 1 \\ L \\ Q \end{bmatrix} \quad (3.30)$$

Therefore, a small Q will reduce the eigenvector element and seemingly its importance in the linear dependency. However, it is clear from eqn(3.29) that \mathbf{x}_3 plays an integral part in this relationship. Whilst Belsley’s example is extreme, the same problem can exist when the covariates are full rank. It is safe to assume that large eigenvector elements indicate a part in the dependency, however the reverse is not always true for small elements.

Variance-Decomposition Proportions

Belsley (1991) promotes the use of variance-decomposition proportions (VDP) in partnership with condition indices. The theory of the VDP can be illustrated in part by features already presented in both the VIF and the CI discussions. Consider once again the SVD of the VC matrix of regression coefficients shown in eqn(3.31),

$$VC(\mathbf{b}) = \sigma_\varepsilon^2 (\mathbf{X}^T \mathbf{X})^{-1} = \sigma_\varepsilon^2 \mathbf{V} \mathbf{\Lambda}^{-1} \mathbf{V}^T \quad (3.31)$$

The SVD is as follows,

$$(\mathbf{X}^T \mathbf{X})^{-1} = \mathbf{V} \mathbf{\Lambda}^{-1} \mathbf{V}^T$$

$$= \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1k} \\ v_{21} & v_{22} & \cdots & v_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ v_{j1} & v_{j2} & \cdots & v_{jk} \end{bmatrix} \begin{bmatrix} 1/\lambda_1 & 0 & \cdots & 0 \\ 0 & 1/\lambda_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\lambda_k \end{bmatrix} \begin{bmatrix} v_{11} & v_{21} & \cdots & v_{j1} \\ v_{12} & v_{22} & \cdots & v_{j2} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1k} & v_{2k} & \cdots & v_{jk} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^p \frac{v_{1j}^2}{\lambda_j^2} & \sum_{j=1}^p \frac{v_{1j}v_{2j}}{\lambda_j^2} & \dots & \sum_{j=1}^p \frac{v_{1j}v_{kj}}{\lambda_j^2} \\ \sum_{j=1}^p \frac{v_{2j}v_{1j}}{\lambda_j^2} & \sum_{j=1}^p \frac{v_{2j}^2}{\lambda_j^2} & \dots & \sum_{j=1}^p \frac{v_{2j}v_{kj}}{\lambda_j^2} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^p \frac{v_{kj}v_{1j}}{\lambda_j^2} & \sum_{j=1}^p \frac{v_{kj}v_{2j}}{\lambda_j^2} & \dots & \sum_{j=1}^p \frac{v_{kj}^2}{\lambda_j^2} \end{bmatrix}$$

Notice that each variance of b_j (highlighted in red) is split into k elements of eigenvectors, each associated with only one singular value (i.e. λ_j). Each ‘small’ eigenvalue (indicating one linear dependency) will proportionally inflate the eigenvector as the value lies in the denominator – thus inflating the overall variance of the coefficient, indicating covariate involvement in the dependency.

$$\eta_{kj} = \frac{v_{kj}^2}{\lambda_j^2} \quad \eta_k = \sum_{j=1}^p \eta_{kj} \quad \psi_{jk} = \frac{\eta_{kj}}{\eta_k} \quad (3.32)$$

(Belsley 1991)

Each VDP (ψ) will relate to precisely one near dependency. This concept is most easily presented and understood as part of a table (Belsley 1991). This can be demonstrated using the correlation matrix in Figure 3.12,

CI	Var(b_1)	Var(b_2)	Var(b_3)
1	0.023	0.023	0.004
1.97	0	0	0.955
39.1	0.977	0.977	0.001

Table 3.1: The VDP of the $X^T X$ matrix in fig.3.9.

Notice in this simple example that each variance composition sums to unity. The variance of the coefficient b_3 is almost entirely explained in the second CI (i.e. the second principal axis). The final CI is ‘large’, demonstrating the near redundancy of the final dimension. The VDP for this index contains two large elements on b_1 and b_2 . This demonstrates the high correlation between these two covariates and the approximate redundancy of x_3 in the linear dependency.

Next revisit the counter-example proposed by Belsley in eqn(3.29),

CI	Var(b_1)	Var(b_2)	Var(b_3)
1	0	0	0
4.59	0	0	0
NA	1	1	1

Table 3.2: The VDP of the counter-example used by Belsely (1991) in eqn(3.29).

The final condition index is undefined as the $\mathbf{X}^T\mathbf{X}$ matrix is singular. Notice that all covariates are equally weighted in the dependency. Although this trivial example is not particularly informative, the VDP have recognised the equal importance of the three variables in the perfect linear relation, even when the magnitude of each covariate in the relation differs.

All Subsets Regression

An interesting alternative to the VIF and CI is ‘all-subsets’ regression. The method is to regress each predictor on all possible combinations of the remaining covariates. The link to the VIF can be seen immediately in that $R_{x_j}^2$ performs a similar operation. However, $R_{x_j}^2$ only provides information against all remaining covariates and therefore fails to specify which particular predictors are involved in the relationships. All subsets regression would provide an answer to this limitation as all possible relationships are assessed and the corresponding analysis would flag those found to demonstrate a linear dependency. The cost is that the process is computationally heavy and would provide a substantial amount of information to digest. However, if this can be overcome then the method can provide a useful extension to the VIF measure when the results are analyzed with care.

Belsley (1991) highlights that although useful, this information would still be subject to the effects of collinearity if the subsets themselves exhibit linear dependencies (i.e. those cases we are particularly interested in). Any information gained from t -statistics and variances would be subject to the effects of collinearity. Therefore, for this method to be of use, it would require an understanding of which information could be used reliably – perhaps requiring the additional use of a collinearity index. In chapter 5 the matroid approach is developed, which borrows much from this idea, but instead of analyzing t -statistics and variances, uses collinearity indices such as the VIF to assess the subsets of covariates. Although this new method is promoted as a means of forming conceptual models for use in confirmatory analysis, they could equally be viewed as a powerful (although computationally laborious) method of assessing linear dependencies.

3.3.4 The Impact of Collinearity on Variance

Researchers will often use a ‘rule of thumb’ for the VIF to indicate ‘serious’ or ‘severe’ collinearity in a dataset. This will encourage the use of advanced procedures to relieve the problems of collinearity which can lead to additional complications in the analysis. If other factors had been accounted for in the initial assessment of the data, the need for such remedial action may be less than first thought. O’Brien (2007) questions the validity of any inference based solely on the VIF measure (although similar arguments can be extended to any ‘correlation based’ measure). The VIF provides a description of the collinearity present in the model. However, the variance equation of the regression coefficient demonstrates the influence of further factors in modifying the impact of collinearity,

$$\text{Var}(\mathbf{b}_j) = \frac{\sigma_{\epsilon}^2}{(1 - R_{x_j}^2) \sum \mathbf{x}_j^2} \quad (3.33)$$

These relate to (1) the coefficient of determination - R_y^2 , (2) the sample size and (3) the variance of the predictors. The vector geometry in Figure 3.9 illustrates that none of these additional factors have an impact on the value of the VIF, however when considered as part of a ‘model environment’ we begin to understand the potential effects of these factors. The geometrical illustration in Figure 3.13 is intended to demonstrate some of the effects of the model environment on the relative impact of collinearity.

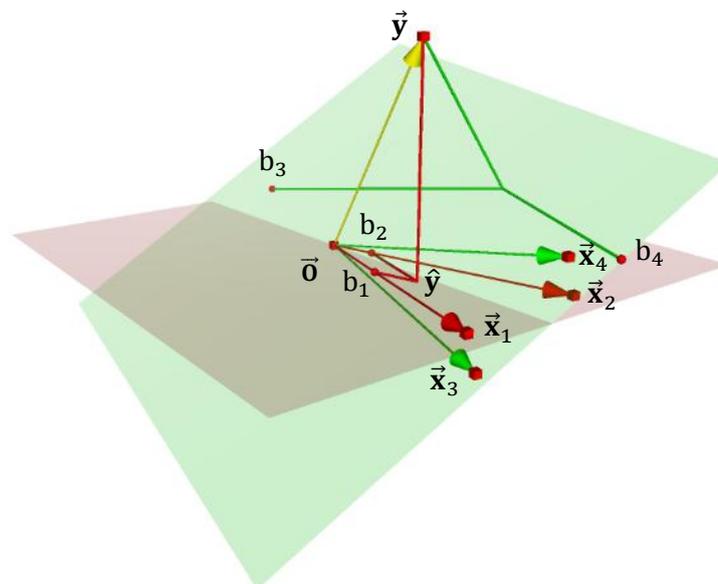


Figure 3.13: The effects of external factors on the impact of collinearity.

There is a rotation of the regression plane spanned by the green and red covariates in the models. These planes are to reflect changes in the position of the response (i.e. from measurement error, sampling variation etc.), but could equally (although against the basic assumptions of the linear model) reflect measurement error on the predictors. When the response is 'closer' to the regression plane in the green example (reflected by an increased R_y^2), a change in the slope of the plane will conceptually have less impact on the coefficient point estimates. Equally, an increase in sampling variation will counter this restriction of movement, thus increasing the variation – but again relative to the response. Further to that, an increase in correlation (reflected by a smaller angle between vectors \vec{x}_1 and \vec{x}_2) would demonstrate that small changes generated by sampling error could greatly impact on point estimates. Each of these effects can be moderated by a response vector lying close to the plane, indicated by a high R_y^2 .

The vector geometry emphasizes the effects of these factors on the impact of collinearity. From an applied viewpoint it would seem difficult to make any inferences regarding collinearity based solely on the $\mathbf{X}^T\mathbf{X}$, when any potential impact is governed by these factors. The influence of the response on the variance of the coefficients can be illustrated by substituting an unbiased estimate of the residual variance into eqn(3.20),

$$\text{Var}(b_j) = \frac{(1 - R_y^2) \times \sum(\mathbf{y}_i - \bar{\mathbf{y}})}{(1 - R_{x_j}^2) \times (n - k - 1) \times \sum \mathbf{x}_j^2} \quad (3.34)$$

$$= \frac{\sum(\mathbf{y}_i - \bar{\mathbf{y}})}{(n - k - 1) \times \sum \mathbf{x}_j^2} \cdot \frac{(1 - R_y^2)}{(1 - R_{x_j}^2)} \quad (3.35)$$

(O'Brien 2007)

The second fraction in eqn(3.35) highlights the extraction of a similar factor to the VIF which relates to R_y^2 instead of $R_{x_j}^2$. This is named the variance deflation factor (VDF) by O'Brien (2007). The name refers to the 'deflating' effect that a high R_y^2 can have on the variance of the regression estimates. The VDF is constructed in a similar way to the VIF,

$$\text{VDF} = 1 - R_y^2 \quad (3.36)$$

Therefore, a rule of thumb should account for the VDF as well as the VIF. Eqn(3.35) suggests that a $\text{VIF} \cdot \text{VDF}$ term would place the inflation of the sample variance in the context of the model. It is important to note that the vector geometry illustrates influences on the point estimates as well as variance of the coefficients. Both of these features are considered in our discussion and development of a new collinearity index.

3.3.5 Remedies of Collinearity

The results of an index such as the VIF or CI are frequently used to justify reducing the collinearity in the model. If a high degree of collinearity is identified amongst the variables, the user may remove collinear variables, enter linear combinations as a single predictor or employ alternative (often more complex) methodology. When the focus is on model specification, there is a great danger attached with allowing a statistical measure to dictate which variables are included in the model (e.g. forward, backward, stepwise selection (Derksen and Keselman 1992)). Misspecification of the model leaves the study susceptible to biased and inefficient estimates. This can occur when the user incorrectly removes a variable that is causally related to the response, without it being accounted for by others in the model (e.g. a sufficient set of confounders). Omitted-variable bias is present when parameters are over or underestimated, by leaving out important variables in the model (e.g. mediation, confounding, suppression examples).

Statistical models provide a great simplification of the intermediate steps that define 'component causes' in the real world. Third variable examples are a useful illustration of the theory of confounders, mediators etc. However, in application we are often faced with highly complex mechanisms. It is evident why spurious associations are often present in statistical analysis as the user rarely captures the complexities of the system under study. The 'remedies' focus on what we consider to be a 'problem' in collinearity from a statistical perspective. For the confounding example, a correlation was observed between **X** and **Z** in the sample data. Adjusting for the confounder in the model would increase the precision of the estimation. However, this is based on *a priori* knowledge to distinguish the effects from a competing exposure with a 'nuisance' correlation in the study data.

Removing a variable based solely on a statistical measure of collinearity is ill-advised. Also, the impact of collinearity is moderated by factors in the model environment, such as the predictors association with the response and the model error. Before any 'remedial' measures are employed, the researcher should be confident about the source and the extent of the 'problem'. It may be that in some circumstances the collinearity detected amongst the predictors can be used to the benefit of the estimation. Whilst collinearity diagnostics can play a crucial role in the specification of the model and understanding the estimates, the model environment and causal relationships should be accounted for before any remedial measures are explored.

3.4 Investigating the Role of the Response

The simulations in this section are intended to demonstrate the impact of the response on the coefficients from a linear regression model. The simulations are purely statistical, with no epidemiological context implied at this point. The index developed in this project must account for the impact of collinearity on the point estimate as well as the coefficient standard errors. The simulations in this section are designed to explore both of these changes through varying individual influences on the behaviour of the regression estimate under collinearity. In each simulation, only measures available to the researcher will be used, such as the correlation between the covariates and the correlations between each covariate and the response. In each case a regression situation is considered involving two covariates and a single response. Many of the potential ‘symptoms’ of collinearity can be illustrated using this simple model.

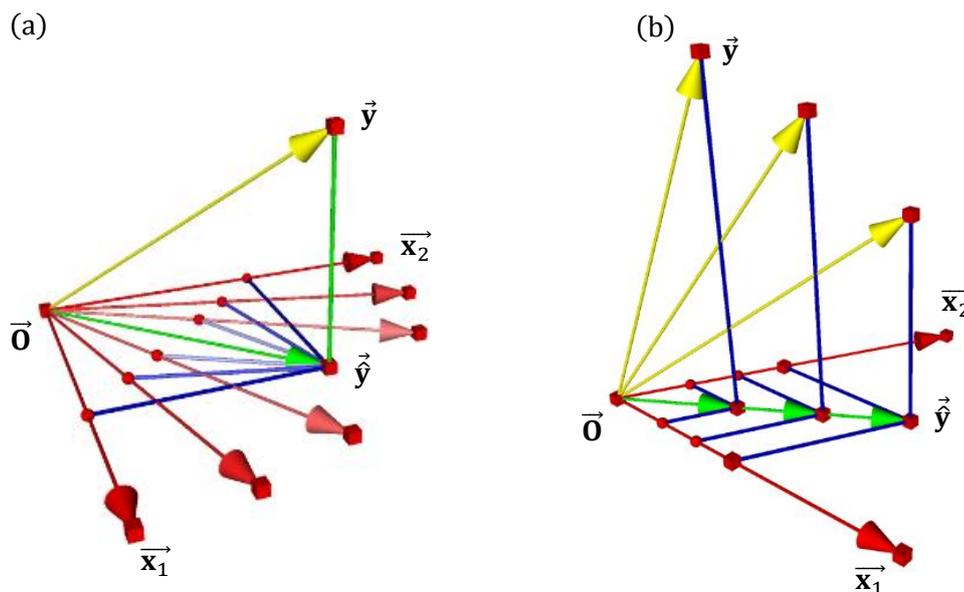


Figure 3.14: Models investigated in (a) simulation 3.1 and (b) simulation 3.2.

Simulation 3.1

The first simulation is designed to investigate the effect of increasing correlations between two predictors. Geometrically, the response is held at a constant angle from the regression plane (i.e. $\cos(\theta_{yX}) = R_y^2$), whilst the angle between the predictors is varied (i.e. r_{12}) (see Figure 3.14a).

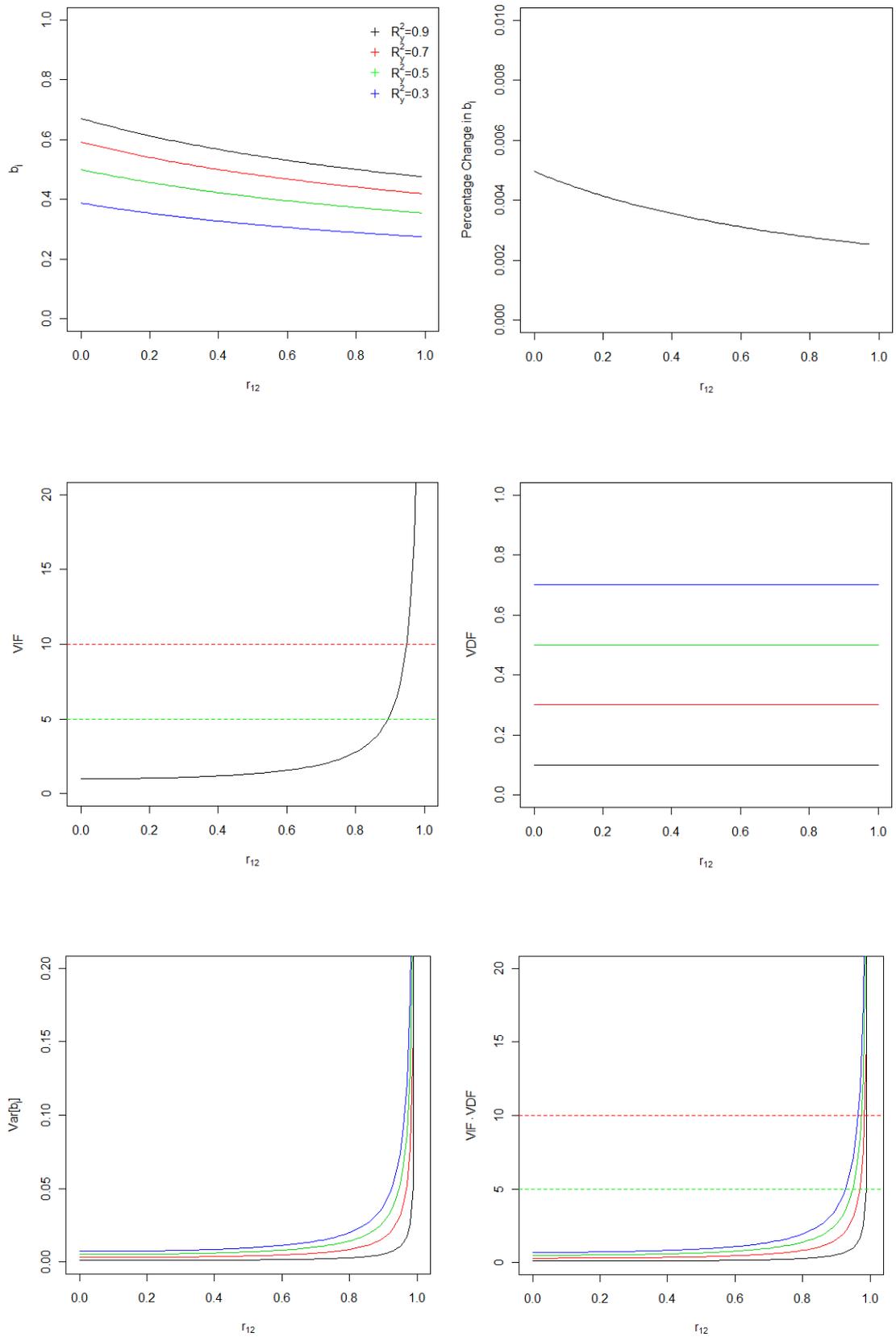


Figure 3.15: Parameter values from Simulation 3.1.

For the case in which the covariates are orthogonal (i.e. uncorrelated), the regression coefficients for the multivariable model are equal to those of the two individual univariable models for x_1 and x_2 . As the correlation between x_1 and x_2 is increased, the partial regression coefficients of the full model decrease by a common percentage change between simulations. There is no suppression effect in this example as the correlations (r_{12}, r_{y1}, r_{y2}) are positive. Hocking and Pendleton (1983) use the analogy of a picket fence, in which each picket represents a covariate. The greater the collinearity, the more overlap there is between the pickets – obscuring the view/role of each. If we imagine balancing a table on these pickets (i.e. a regression plane), the closer the pickets (i.e. the collinearity), the less stable the table will sit. This analogy also indicates that (when feasible due to time, monetary, study design constraints) it may be useful to collect additional data – i.e. more pickets (Freund and Wilson 1998). New data points could relieve a strong spurious collinearity and could reduce the variance of the coefficient estimates. However, there is no guarantee as to the benefit of such action and this must always be balanced against the often high monetary costs in acquiring such data.

The VDF remains constant throughout each simulation. It is only the VIF which changes to reflect the collinearity amongst the covariates. By considering different R_y^2 (i.e. the different colour markers), it is clear that the VDF indicates this change. As the angle between y and the regression plane is reduced (i.e. R_y^2 increases), the VDF decreases, subsequently deflating the variance of the estimates. The graph illustrates that this deflating effect has less impact when the collinearity is low. This is because the VIF remains relatively close to one and so the difference in VDF has little influence in deflating the variance. The common ‘rules of thumb’ of 5 and 10 for the VIF have been added to the graphs. This corresponds to a correlation of 0.90 and 0.95 respectively between covariates in the bivariable model. The logic can be seen behind this rule of thumb in that the VDF has relatively little effect on the variance until this point. However, it is clear that a large VDF can restrict the variance from substantially increasing beyond this point until ‘extreme’ collinearity is observed.

Simulation 2

In the second set of simulations the angle between the response and the regression plane is adjusted to investigate a change in the R_y^2 . The black, green and red lines relate to a VIF equal to 1, 5 and 10 respectively. These represent a baseline of orthogonality and two popular ‘rules of thumb’.

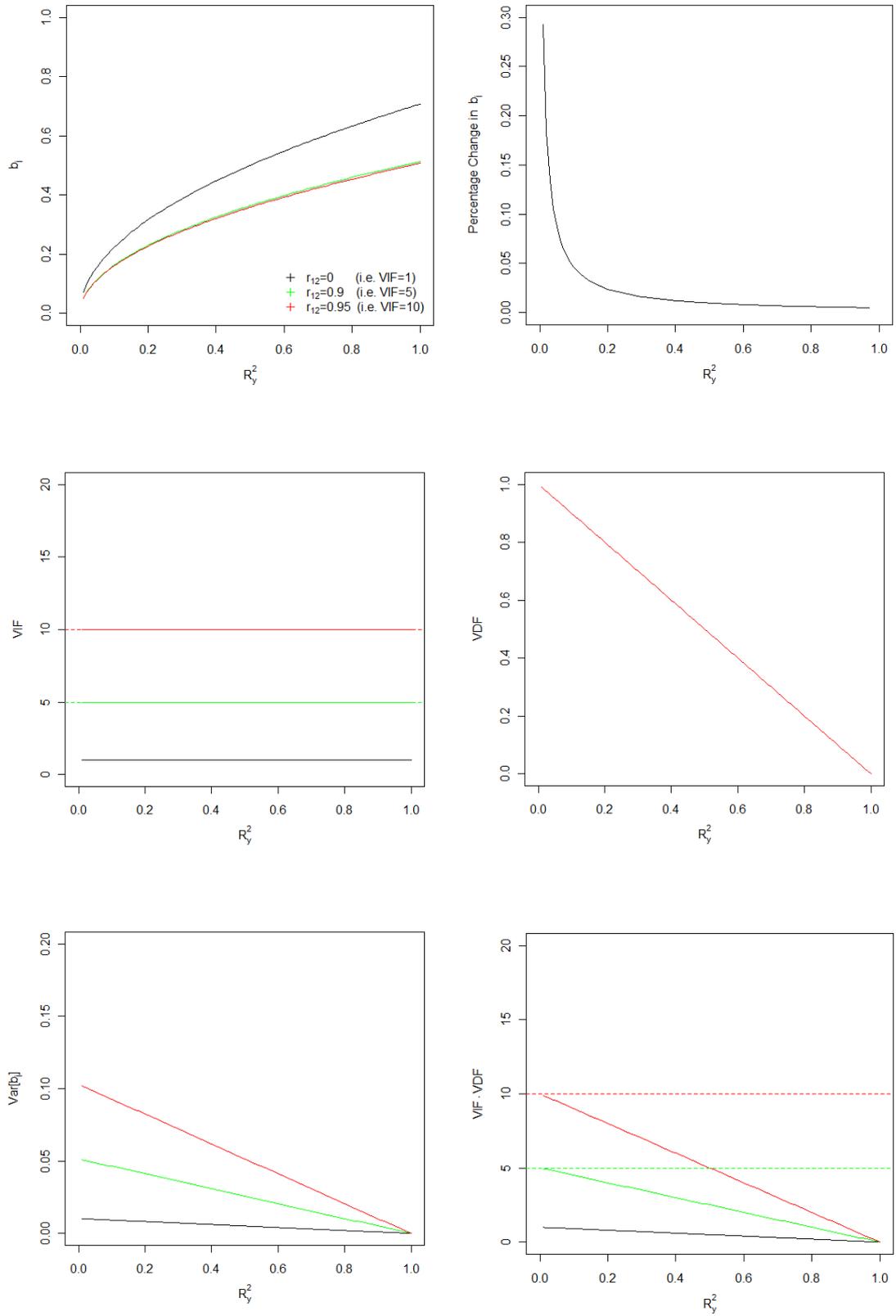


Figure 3.16: Parameter values from Simulation 3.2.

Figure 3.16 illustrates that whilst the VIF remains constant throughout each simulation, an increasing R_y^2 decreases the variance to a negligible level. When R_y^2 is close to unity, all of the variances are approximately equivalent. The VIF·VDF demonstrates that the green and red lines are equivalent when $R_y^2 = 0$ along with a VIF = 5; and $R_y^2 = 0.5$ along with a VIF = 10. Therefore, an arbitrary 'rule of thumb' based only on the correlation between the covariates alone is dangerous to base any decisions on. The response plays an important role in moderating the impact of collinearity on both the variance and point estimates of the coefficients.

Simulation 3.3

Simulations 3.1 and 3.2 demonstrated the common (and often expected) effects of collinearity on the parameter estimates of the regression model. From the vector geometry, the projected response is held constant to lie between \vec{x}_1 and \vec{x}_2 . This produced a percentage decrease on the regression coefficients upon increasing the degree of collinearity between the covariates. In simulation 3.3, the correlations are varied between the projected response and the covariates (see Figure 3.17).

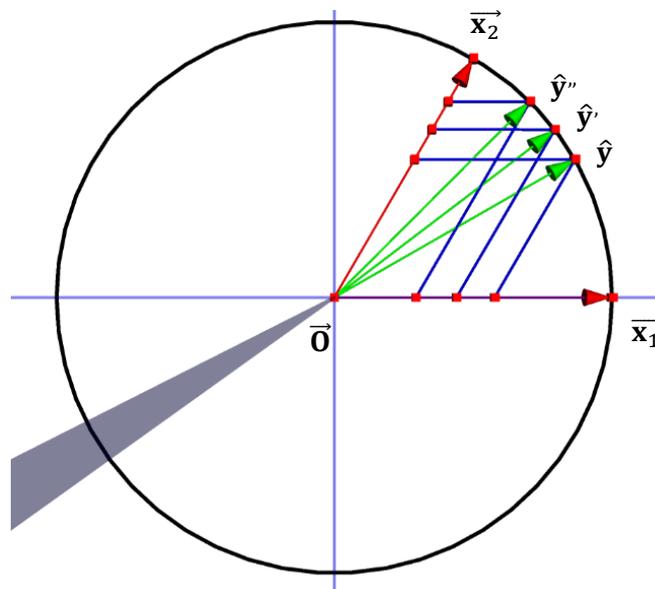


Figure 3.17: Effect of an unequal correlation between covariate and response

In this simulation, the response lies on the regression plane spanned by \vec{x}_1 and \vec{x}_2 (i.e. is explained entirely by the predictors). The position of the response on the plane is varied by its correlation (i.e. cosine angle) with the covariate \vec{x}_1 . Coloured markers are used to represent the correlations between predictors.

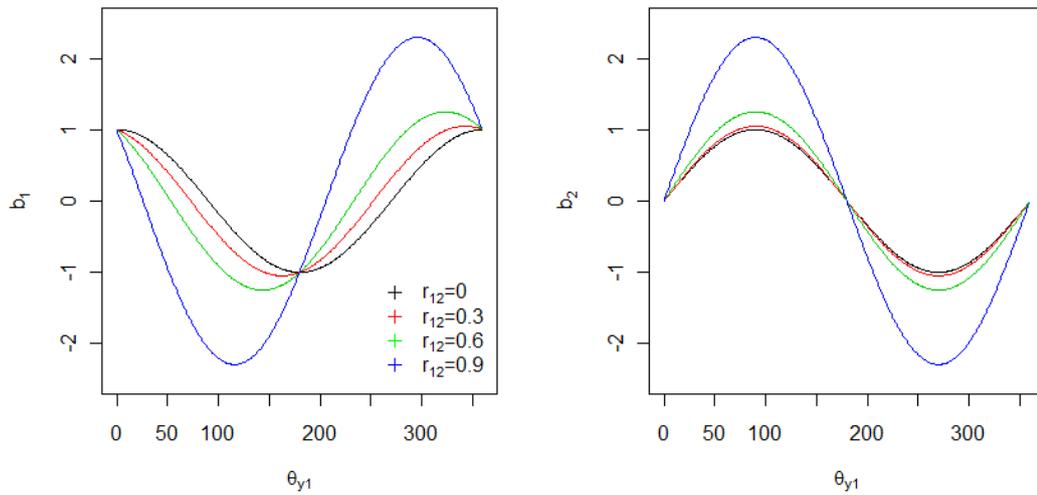


Figure 3.18: Change in b_1 and b_2 with a rotation of the response ($R_y^2 = 1$).

For simulation 3.3 a suppression effect can be observed in Figure 3.18. Regardless of the change in collinearity, the coefficient estimate on \mathbf{x}_2 is zero at 0° and 180° (as the angle is varied with respect to \mathbf{x}_1). The regression coefficient b_2 reaches unity when the projected response is parallel to the vector $\overrightarrow{\mathbf{x}_2}$. It is then maximized at 90° and 270° (i.e. when the response is orthogonal to $\overrightarrow{\mathbf{x}_1}$). The incremental increase in collinearity produces an inflation in the b_2 regression coefficient. In contrast, the graph of b_1 demonstrates a shift in the maxima toward the b_2 maxima with increasing collinearity. As was described for the suppression effect in section 3.2.2, the regression coefficient on $\overrightarrow{\mathbf{x}_1}$ becomes more negative, whilst the positive univariable coefficient on $\overrightarrow{\mathbf{x}_2}$ is enhanced. At the point in which perfect collinearity is present, the estimates could not be determined and we can imagine (if only theoretically) the curves equal with opposite sign. From this simulation we can begin to understand the potential impact of collinearity on the interpretation of the coefficients. Whilst a moderate collinearity may in turn have a moderate effect on the coefficients, a high collinearity can have a much inflated effect dependent on the position of the response.

Simulation 3.4

For the final simulation the change in angle between $\overrightarrow{\mathbf{x}_1}$ and $\overrightarrow{\mathbf{y}}$ is repeated. A low and high correlation between the covariates is considered separately ($r_{12} = 0.3, 0.9$). In each graph, the angle of the response to the regression plane is varied through $0 - 360^\circ$.

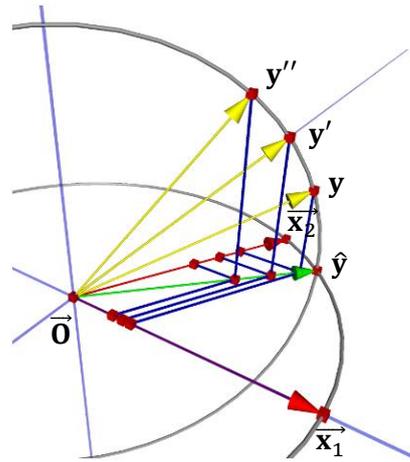


Figure 3.19: Simulation 3.4 – Varying R_y^2 on unequal correlations.

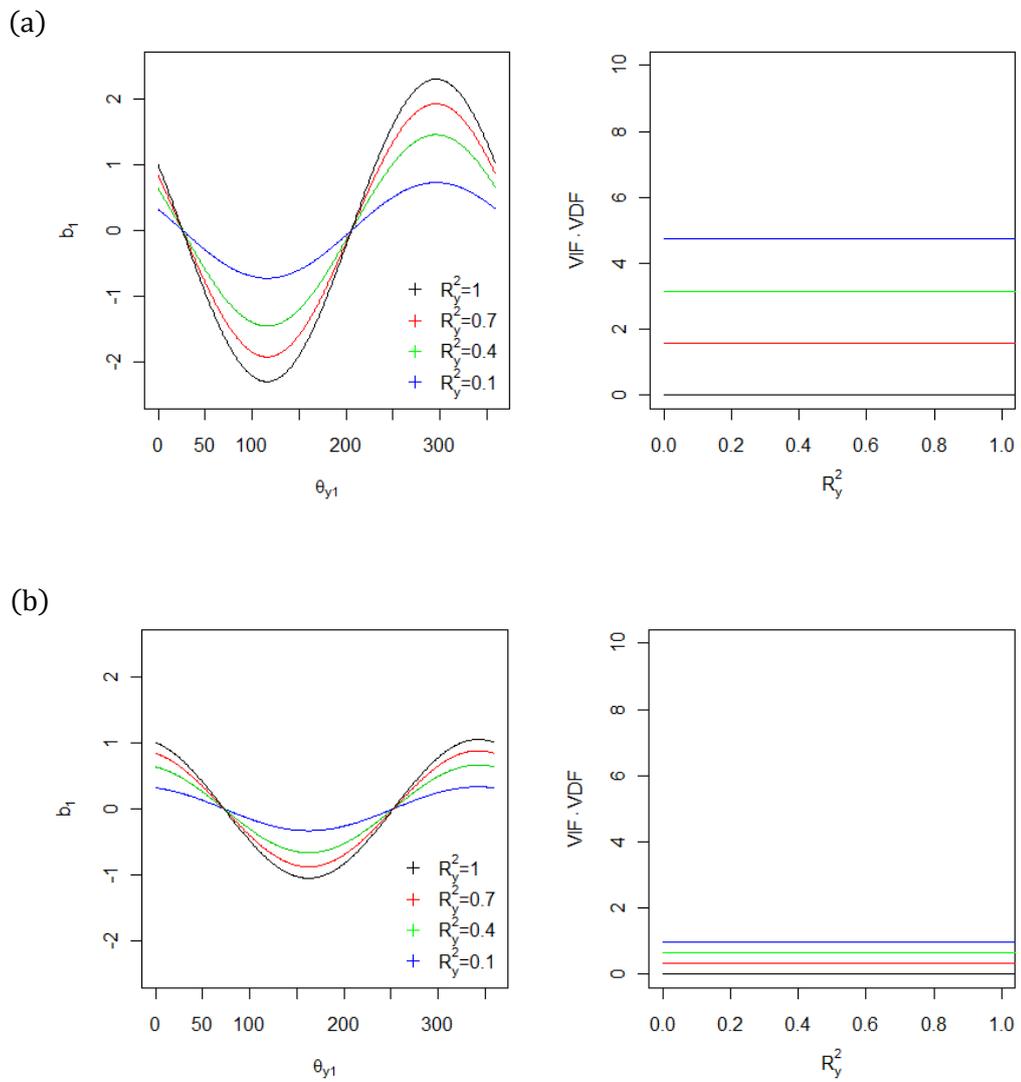


Figure 3.20: Results of simulation 3.4 for (a) $r_{12} = 0.9$ and (b) $r_{12} = 0.3$.

In Figure 3.20 the variance (and $VIF \cdot VDF$) remains constant throughout each simulation as the collinearity and R_y^2 are fixed. Therefore, the interest is in the point estimates of the coefficients. The inflation due to the suppression effect is much greater with a high R_y^2 . From this extension to simulations 3.1 and 3.2, it is clear that it is the impact on the point estimate, as well as variance that is essential for the new collinearity index to assess. The change from a population coefficient to an estimated coefficient (and subsequently a potential change in interpretation) may be much greater under different collinearity conditions, whilst the variance of the estimates remains constant. This is important for the development of a new index in chapter 5.

The simulations in this chapter have been designed to explore the influences of collinearity on the bias and standard errors of the coefficients. The VIF and CI (along with other 'correlation based' indices) are intended to measure only the degree of collinearity. Understanding the impact of collinearity is far more complex than 'correlation based' diagnostics can assess. The role of the response will have a major influence on the impact of collinearity on the point estimate and the standard errors of the coefficients. Therefore, the role of R_y^2 is essential in the variance and coefficient point estimates. The combination of a collinearity measure (e.g. $R_{x_j}^2$) and a measure of the response (e.g. R_y^2) appears potentially at the heart of a new index.

"... collinearity must not only be shown to be present but also shown to be adversely affecting the estimate of the given coefficient"

(Belsley 1991)

This quotation by Belsley should provide the motivation for the development of a fresh diagnostic measure. Keeping in mind the conceptual understanding of the variable types in epidemiological study, it cannot be said that any of the measures discussed in this chapter adequately address the latter part of this challenge. Examples provided discussing the role of biological knowledge in interpreting a confounding, mediating or competing exposure are testament to this. In addition, the influence of a suppression effect (contrary to many researchers understanding of the impact of collinearity) can play a major role in adversely affecting a researcher's interpretation of a model incorporating highly collinear variables. Statistical measures cannot distinguish such features, but can be built to work in partnership with this external knowledge. This is the primary motivation in the development of new techniques for applied regression studies.

3.5 Conclusions

If a change of sign is observed after including a third variable in a regression model, we may consider this a false result. We are often led to believe that this is a negative consequence of collinearity. Theoretically we could have all the information on the variables without measurement error or sampling variation and still experience this change of sign. Sampling variation may generate the discrepancy, but it may also be true of the population parameter. It is only when we bring the biological interpretation into the work that the coefficient begins to be questioned – it is not a problem from a statistical perspective or a failure of OLS. The estimate is still unbiased. The change of sign or magnitude of the coefficient may not be what we expect biologically and so it is labelled a ‘problem’ of collinearity. This is a difficulty of balancing disciplines. Statistics are often accepted when they fit with biological knowledge. However, the complexities of the real world systems are difficult to model with overly simplistic statistical models and this creates a bridge between statistical and clinical inference.

Epidemiology utilizes a number of advanced statistical methods. Consequently it is faced with the problem of applying statistics to the ‘real world’. It often appears to be limitations in the statistics that prevent us from ever truly answering these problems. However, as viewed in Pearl’s birth-control example in section 3.2.3, it is not the statistical method that is at fault, but rather the way in which it is used. Through including the mediator, the estimate does not reflect the interpretation of the causal model. Understanding the relevance of the statistics in the real world is one of the biggest challenges in applied statistics in epidemiology. The researcher should not automatically choose the blanket approach for all types of problems, but instead ask whether the assumptions and underlying model reflect the application. The collinearity diagnostics considered in this chapter stay very much within the realms of statistics. They do not address the epidemiological interpretation as they were not designed for such a purpose. It may be impossible for the statistics to achieve what we would like working within epidemiological restrictions such as study design and the seemingly guaranteed existence of collinear variables. The aim should remain to develop appropriate methods that address the epidemiological problem; to encourage the cooperation of the statistics with the clinical interpretation. This is the motivation for the development of a new index in chapter 5. It also begins our discussion regarding the use and application of ‘default’ statistical methods for the study of MetS in chapter 6.

4. Vector Geometry of Exploratory Analysis and Advanced Regression Methods

Basic concepts of vector geometry in correlation, rotation and projection were introduced in chapter 2 and used to derive the least squares estimator. Epidemiological study often dictates the need for more ‘advanced’ estimators than OLS. From an analytical perspective these techniques can appear complex. In chapter 4, principal components regression (PCR) and partial least squares (PLS) are considered which are in the class of ‘shrinkage’ estimators. The incorporation of such techniques into statistical software will often promote their use, as a thorough understanding of the methodology may not be required for implementation. Whilst the complexity must be respected, geometrical illustrations can be used to present and better understand such methods when possible. This is one of the reasons for discussing the OLS estimate in chapter 2. The intention was not to encourage the application of OLS, but instead to demonstrate a baseline approach that allows more complex methods to be better understood and in turn to encourage a considered application of the methodology.

A selection of advanced regression methods are introduced in chapter 4 that will be used in the remaining chapters of the thesis. Geometrical links between methods are presented where possible. The intention of this presentation is threefold; (1) to aid with the building of new methods from a common baseline. Links are considered between OLS and PLS, with the latter looking to maximize the covariance of the predictors and the response. This provides the motivation for a new collinearity index; (2) to understand the exploratory factor and clustering methodology and their uses in application. PCA is compared to factor analysis (EFA), focussing in particular on the underlying models assumed by each approach; (3) to understand what insights estimators such as PCR and PLS can provide in novel applications. The selection of methods in this chapter is very much dictated by the aims of later chapters. This is to provide a framework for the methodology that allows justification for their application in later work.

4.1 Geometry of Multivariable Regression

The geometrical extension from a bivariable regression model (i.e. $k = 2$) to a higher dimensional example (i.e. $k \geq 3$) can be made using the same basic principles that were presented in section 2.4. However, for the multivariable example, it is useful to understand the space in which the vectors exist. The n -dimensional subject space is first divided into two orthogonal subspaces – one containing systematic effects and one random effects (Wickens 1995). The ‘systematic space’ is then further divided into two orthogonal subspaces, consisting of a 1-dimensional intercept space and a k -dimensional effect space. The effect space contains the k centered regression vectors (e.g. in a bivariable model this is the regression plane spanned by the vectors \vec{x}_1 and \vec{x}_2). To this point a focus has been placed only on the projected response contained in the effect space (\vec{y}), however the response vector is further built of components in the intercept ($\vec{\alpha}$) and error space ($\vec{\epsilon}$).

$$\vec{y} = \vec{\alpha} + \vec{y} + \vec{\epsilon} \quad (4.1)$$

Collinearity is not changed upon mean centring and so this work is presented in the reduced $(n - 1)$ -dimensional space. However, collinearity will impact on the variation of the coefficient estimates, therefore the projection of \mathbf{y} onto the error space is important.

4.1.1 Incorporating Error into Vector Geometry

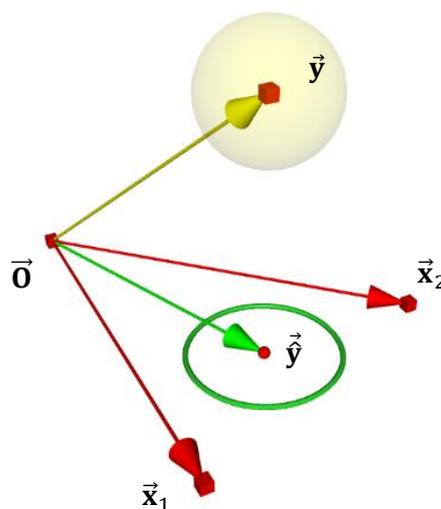


Figure 4.1: An illustration of error in vector geometry.

The error space is not simple to demonstrate in geometrical form due to the often high dimensionality (i.e. $n - (k + 1)$). The illustrations in this work have displayed only the systematic component, which has allowed us to work in fewer than the theoretically necessary dimensions. The random component is often illustrated as a uni-dimensional vector \mathbf{y}_ε , when in fact it is contained in the $(n - k - 1)$ -dimensional error space. Unfortunately, the 3-dimensional view available in human perception is rarely adequate to demonstrate this. The dimension of the space in which each vector lies controls its freedom to move. The impact of collinearity on the variance of the estimates can be considered in this manner. When k predictors are orthogonal, the effect space is k -dimensional. If there exists a perfect linear dependency amongst two predictors, they lie in a common dimension, meaning there is a redundant dimension in the k -dimensional effect space. This increases the dimensionality of the error space by one, thus allowing a greater movement of the error vector. This is reflected by the available degrees of freedom and prevents computation of OLS estimates under perfect collinearity (see chapters 7-8).

4.1.2 The Error Space

The linear model relies on the assumption that the ε_i (for $i = 1, \dots, n$ for n observations) are independent and identically distributed normal random variables with mean zero and constant variance σ_ε^2 (see section 2.2.2). This determines that the random error vector ($\vec{\varepsilon}$) represents n independent normal random variables. Independence of the errors (i.e. orthogonality) and constant variance (i.e. length) dictates that $\vec{\varepsilon}$ is spherically distributed in the geometry with radius $r = \sqrt{\varepsilon^T \varepsilon}$. The family of n -dimensional hyper-spheres provide the isodensity contours that define the distribution of $\vec{\varepsilon}$ (Belsley 1991). The contours are the set of points for which the values of ε_i give a constant probability C for the density function φ of ε (Gatignon 2003).

$$\begin{aligned} \varphi(\vec{\varepsilon}) &= \left[\frac{1}{\sqrt{2\pi}} e^{-\varepsilon_1^2/2} \right] \cdot \left[\frac{1}{\sqrt{2\pi}} e^{-\varepsilon_2^2/2} \right] \dots \left[\frac{1}{\sqrt{2\pi}} e^{-\varepsilon_n^2/2} \right] \\ &= \frac{1}{(2\pi)^{n/2}} \exp \left[\frac{-(\varepsilon_1^2 + \varepsilon_2^2 + \dots + \varepsilon_n^2)}{2} \right] = \frac{1}{(2\pi)^{n/2}} \exp \left[-|\vec{\varepsilon}|^2/2 \right] \end{aligned} \quad (4.2)$$

$$C = \varepsilon_1^2 + \varepsilon_2^2 + \dots + \varepsilon_n^2 = \sum_{i=1}^n \varepsilon_i^2 \quad (4.3)$$

(Wickens 1995)

Eqn(4.2) implies that the error vector is uniformly distributed in any direction (Wickens 1995). This means that the error vector can take any angle in the n -dimensional hyper-sphere with equal probability. This demonstrates the spherical properties of the error distribution. When considering only two predictors, the hyper-sphere is projected onto a 2-dimensional plane (regardless of the collinearity), therefore a spherical distribution is always observed in the regression space (see Figure 4.1). However, when the error distribution is projected onto a space consisting of more than two non-orthogonal predictors, the space is no longer constructed of standard orthogonal axes. Thus the spherical distribution will typically take the form of an ellipsoid (in the presence of collinear predictors) – an example of this is shown in Figure 5.22.

4.1.3 Least Squares Regression in Higher Dimensions

The OLS estimate in vector geometry can be viewed as a unique projection of \mathbf{y} onto the regression space spanned by \mathbf{X} . The projection that minimizes the residual error (i.e. the distance of the response to the regression plane) is the orthogonal projection to the regression space. This was demonstrated for the bivariable regression model in section 2.4. For the model involving two predictors and a response, the vector geometry can be fully presented using the conceptually appealing 3 dimensions that are required. When the geometry is extended to the 3 predictor model a focus must be placed instead on the projected response $\hat{\mathbf{y}}$ to remain in the 3-dimensional regression space.

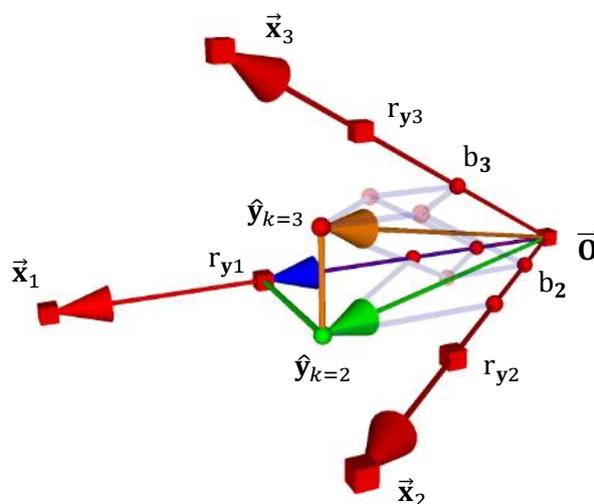


Figure 4.2: An example of a trivariable regression model.

The response projected onto the regression (or effect) space for the higher dimensional model can be divided into components. First, consider the simple regression model involving only the predictor \mathbf{x}_1 . The projected response would lie along the vector $\bar{\mathbf{x}}_1$ with length equal to the correlation r_{y1} (i.e. the blue vector). This is equivalent to the regression coefficient of the univariable model. The second component is orthogonal to \mathbf{x}_1 in the plane spanned by $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ (i.e. green component). This becomes the position of the projected response $\hat{\mathbf{y}}$ for the bivariable model including \mathbf{x}_1 and \mathbf{x}_2 (i.e. green vector).

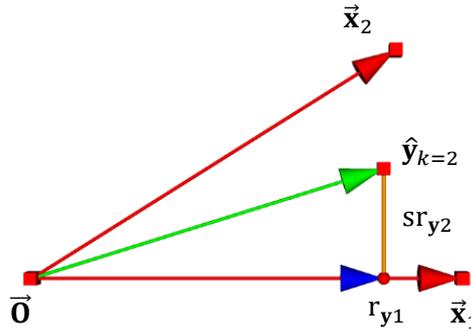


Figure 4.3: Construction of the projected response in 2D regression space.

This component is orthogonal to $\bar{\mathbf{x}}_1$, which demonstrates that \mathbf{x}_1 is held constant in the computation. It equates to a semi-partial correlation between \mathbf{y} and \mathbf{x}_2 (labelled sr_{y2_1}),

$$sr_{y2_1} = sd_y \cdot \frac{r_{y2} - r_{y1} \cdot r_{12}}{\sqrt{1 - r_{12}^2}} \quad (4.4)$$

This is equivalent to the correlation between \mathbf{y} and the residual vector of \mathbf{x}_2 regressed on \mathbf{x}_1 . Following the same principle, the third component is orthogonal to the plane spanned by \mathbf{x}_1 and \mathbf{x}_2 signifying that the addition to $R_{y_{k=2}}^2$ for the three predictor model is found by the correlation of \mathbf{x}_3 with \mathbf{y} , whilst \mathbf{x}_1 and \mathbf{x}_2 are held constant. This is the correlation of \mathbf{x}_3 and the residual vector from the bivariable model. This is labelled a partial correlation ($pr_{y3_{12}}$), with the covariates held constant listed in the subscript. The squared components sum to obtain $R_{y_{k=3}}^2$ (i.e. the length of $\hat{\mathbf{y}}$).

$$R_{y_{k=3}}^2 = r_{y1}^2 + sr_{y2_1}^2 + pr_{y3_{12}}^2 \quad (4.5)$$

b_j is found by the orthogonal projection of $\hat{\mathbf{y}}$ along the plane spanned by the remaining pair of covariates in the model onto $\bar{\mathbf{x}}_j$. Higher dimensional examples can be presented using the same principles of projection onto a lower dimensional space.

4.2 Exploratory Analysis

The role of exploratory analysis is to examine data without a pre-conceived statistical model or implied hypothesis. The purpose is to test assumptions for statistical inference, identify potential exposures and generate models or data structures. The role of confirmatory analysis is to verify models and hypotheses generated *a priori*. A small selection of exploratory methods will be outlined along with a discussion regarding the role of confirmatory analysis in applied work. This discussion is generally of a statistical nature, with a clinical context later assumed in chapter 6. This also acts as a precursor to advanced regression methods discussed in section 4.4 and applied in chapters 7-8.

4.2.1 Principal Component Analysis

PCA extracts weighted composites of manifest variables (known as principal components – or PC's), that are orthogonal (uncorrelated) and ordered by largest to smallest explained variance in \mathbf{X} . The motivation for a PCA is to retain as much of the explained variance of \mathbf{X} in the first few components, whilst reducing the dimensionality of the correlated covariates (Jolliffe 2002). The first PC (labelled \mathbf{z}_1) explains the greatest variability in the data, and each successive PC (\mathbf{z}_j for $j = 2 \dots k$) explains as much of the residual variance as possible. To demonstrate the construction of the PC's, recall the SVD of the $\mathbf{X}^T\mathbf{X}$ matrix presented in section 3.4.2,

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \quad (4.6)$$

From this composition, eigenvectors \mathbf{v}_j and eigenvalues λ_j are generated from the VC matrix of \mathbf{X} . The PC's are found in vector geometry by a rotation of \mathbf{X} using the eigenvector \mathbf{v}_j ,

$$\mathbf{z}_j = \mathbf{X}\mathbf{v}_j \quad (4.7)$$

where \mathbf{z}_j is a $k \times 1$ vector representing the j^{th} PC and \mathbf{X} is a mean centred $k \times 1$ matrix of observations. In PCA, the eigenvectors are referred to as vector 'loadings' or 'weight' vectors (i.e. the rotation matrix) and the elements of \mathbf{z}_j as the component 'scores' (i.e. a bilinear decomposition of \mathbf{X}). A further useful property is that the eigenvalues are proportional to the variance of their corresponding PC's (see section 3.3.2),

$$\mathbf{v}_j^T \mathbf{X}^T \mathbf{X} \mathbf{v}_j = \mathbf{z}_j^T \mathbf{z}_j = \lambda_j \quad (4.8)$$

As the matrix containing the eigenvectors is orthonormal (as defined by the SVD result), the PC's resulting from the transformation will be orthogonal to one another. The properties of the technique can be translated into their geometrical equivalent.

1. The sum of the squared eigenvector elements is equal to unity. This ensures that the new components will be on the same scale as the original \mathbf{X} .

$$\mathbf{v}_{1j}^2 + \mathbf{v}_{2j}^2 + \cdots + \mathbf{v}_{kj}^2 = 1 \quad (4.9)$$

Predictors and PC's pass through a common covariance ellipsoid in the geometry.

2. No variation is lost in the original data; it is reassigned to maximize the contribution of explained variance on the first component. Therefore, the sum of the variances in \mathbf{X} is equal to the sum of the variances in \mathbf{Z} .

$$\begin{aligned} \text{Var}(\mathbf{x}_1) + \text{Var}(\mathbf{x}_1) + \cdots + \text{Var}(\mathbf{x}_k) \\ = \text{Var}(\mathbf{z}_1) + \text{Var}(\mathbf{z}_2) + \cdots + \text{Var}(\mathbf{z}_k) \end{aligned} \quad (4.10)$$

In geometrical terms, the squared lengths of the vectors are equal,

$$|\vec{\mathbf{x}}_1|^2 + |\vec{\mathbf{x}}_2|^2 + \cdots + |\vec{\mathbf{x}}_k|^2 = |\vec{\mathbf{z}}_1|^2 + |\vec{\mathbf{z}}_2|^2 + \cdots + |\vec{\mathbf{z}}_k|^2 \quad (4.11)$$

3. The PC's are uncorrelated with one another.

$$\text{cor}(\mathbf{z}_s, \mathbf{z}_l) = 0 \quad \text{where } s \neq l \quad (4.12)$$

This dictates that the component vectors are orthogonal,

$$(\vec{\mathbf{z}}_s \cdot \vec{\mathbf{z}}_l) = 0 \quad (4.13)$$

4. The PC's capture the maximal contribution to global variance conditional on the orthogonality and sums of squares constraints.

The number of components will always equal the number of predictors (i.e. span the original k -dimensional space), however PCA creates a new set of axes for the data in which

the first few PC's will retain most of the global variation. This allows the user to retain fewer components, whilst still explaining a high percentage of the information provided in the original covariates (thus reducing the dimensionality of the data). This can be a very useful process, particularly for high dimensional data. The information (or signal) should be captured in the early components, whilst the latter PC's should contain mostly noise. However, as with any statistical technique, the quality of the resulting analysis is directly related to the 'quality' of the input data.

For a two predictor model, if the covariates are of equal variance the first PC will bisect the covariate vectors to ensure that the weighting is equal. This is equivalent to a 45° (i.e. $\cos(\sqrt{2}/2)$) rotation of the covariate axes onto the PC axes in variable space,

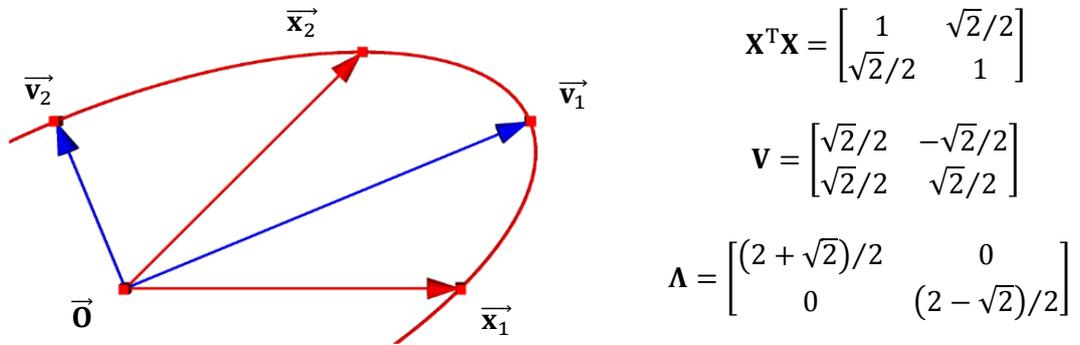


Figure 4.4: PCA analysis in subject space ($r_{12}=0.71$).

The bilinear decomposition produces the first weighting (i.e. \mathbf{v}_1) equal, due to the unit variance on each covariate and conditional on the unity constraint on the squared coefficients. The second PC is orthogonal to the first and adheres to the same unity constraint on the squared coefficients. This quantity is multiplied by the common standard deviation of the original covariates to obtain vector lengths when $sd_{x_j} \neq 1$.

$$r_{x_i z_j} = \sqrt{\frac{\sigma_{12}^2 |\bar{\mathbf{z}}_j|^2}{|\bar{\mathbf{x}}_1|^2 \left[\sigma_{12}^2 + \left(|\bar{\mathbf{z}}_j|^2 - |\bar{\mathbf{x}}_1|^2 \right)^2 \right]}} \quad (4.14)$$

Eqn(4.14) demonstrates the correlation between the old and new vectors calculated using the general correlation formula (see Wickens 1995). When the standard deviations are not equal, the first PC is weighted towards the covariate with the greater standard deviation in the bivariable example. This emphasizes the variance maximizing priority of a PCA.

4.2.2 Exploratory Factor Analysis

Factor analysis is similar to PCA in that constructs are formed from manifest variables (labelled factors rather than components). This allows the dimensionality of the data to be reduced by retaining a smaller number of factors. Although the procedures are mathematically similar, the philosophy behind the factor extraction differentiates the techniques. Factor analysis assumes the variation in a predictor to be generated by various hypothetical constructs, known as factors. These are attributes that cannot be measured directly, but we realise their effects through the observed variables.

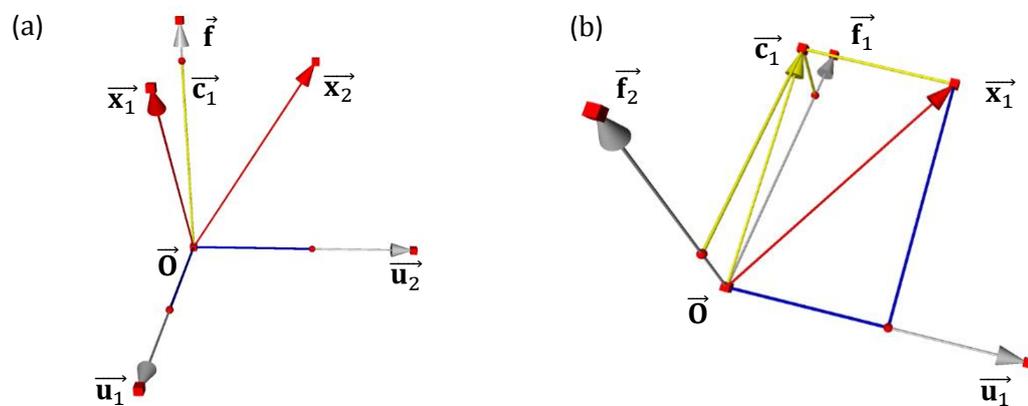


Figure 4.5: The decomposition of predictors into common and shared variation.

Figure 4.5a is a demonstration of the common factor approach. The constructs are formed by partitioning common variance (shown in yellow) from unique and error variance (shown in blue) (Costello and Osborne 2005). The common factor \mathbf{f} represents a single latent variable that is influencing both \mathbf{x}_1 and \mathbf{x}_2 . The additional effect of a unique factor \mathbf{u}_j for each covariate produces an imperfect correlation (i.e. $r_{12} \neq 1$). As such, geometrically the unique and common vectors are orthogonal and of unit length. The covariates lie in the space spanned by the common factor and a single unique factor. With a single common factor as illustrated in Figure 4.5a, the factor space is uni-dimensional, and the component \mathbf{c}_1 present in the factor space represents the common variance of \mathbf{x}_1 and \mathbf{x}_2 . If a second factor is introduced (as in Figure 4.5b), the factor space becomes two-dimensional. The unique vectors each remain orthogonal to the factor space (and each other), however the common vector $\overline{\mathbf{c}}_1$ does not have to lie on one of the orthogonal factors \mathbf{f}_1 or \mathbf{f}_2 , only in the space spanned by the factors. The \mathbf{c}_j are the factor scores and can be presented in as few dimensions as the number of factors retained (see Figure 4.6).

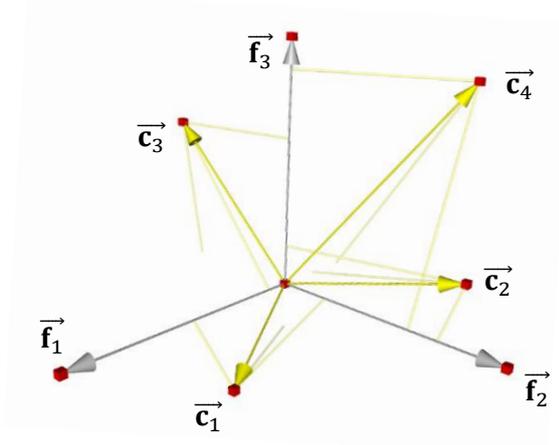


Figure 4.6: Geometrical Illustration of a 3-dimensional factor space.

The proportion of a variables variance that is due to shared variation is known as its communality – i.e. the squared length of \vec{c}_j (Wickens 1995). The unique variance is computed as the total variance minus the communality estimate. The loadings are found by the orthogonal projection of each \vec{c}_j onto each retained factor (e.g. Figure 4.6).

4.2.3 The Role of Confirmatory Factor Analysis

Methods of factor analysis are generally separated into exploratory (EFA) and confirmatory factor analysis (CFA). They are both based on the common factor model (see Figure 6.1), that assumes variables exhibiting high correlations are generated by common latent factors, whereas those that display low correlations by unique factors. An EFA looks to generate a theory about an underlying factor structure from a set of variables. It is typically used when there does not exist strong theory about the structure of the data. In contrast, CFA will take a hypothesized structure and analyze how well the data fits the structure based on certain constraints (often suggested by an exploratory approach).

The use of EFA in scientific research has been evident since the work of Pearson and Spearman in the early 20th century; however the development of CFA in the 1960's is a relatively recent addition. The techniques are commonly perceived as separate entities, however in applied research they should be viewed along a common continuum. When a researcher undertakes an EFA they will often have a 'hunch' about the form of the potential structure. It is ill-advised that the user simply 'throw' variables into the analysis without background reasoning (Floyd and Widaman 1995). Therefore, the driving difference is in the degree of *a priori* information specified by the user.

4.3 Cluster Analysis

Cluster analysis describes a range of algorithms and statistical methods for grouping objects or variables by their degree of association. It is often referred to as unsupervised learning (or classification), meaning that the cluster formations are undefined prior to the analysis. The term cluster usually refers to groups of subjects or objects from data, however there are examples of variable clustering methodology. Some of these methods are comparable to EFA and PCA, but produce 'distinct' (non-overlapping) clusters. Clustering techniques play an important role in clinical applications such as the classification of patients to diagnostic categories (Everitt et al. 1971) and groups reflecting high and low risk subjects (Avanzolini et al. 1991).

4.3.1 Hierarchical Clustering

Clustering methods are split into hierarchical and non-hierarchical techniques. Hierarchical clustering generates groups of highly associated objects that are nested within larger clusters with a weaker association (i.e. display a greater dissimilarity to other clusters). Non-hierarchical clustering procedures partition the dataset, producing a set of clusters that have no overhanging relationships amongst them. k -means clustering is a non-hierarchical method in which the number of clusters is chosen prior to the analysis and cases are iteratively reassigned by their distance from cluster means (see Figure 4.7b).

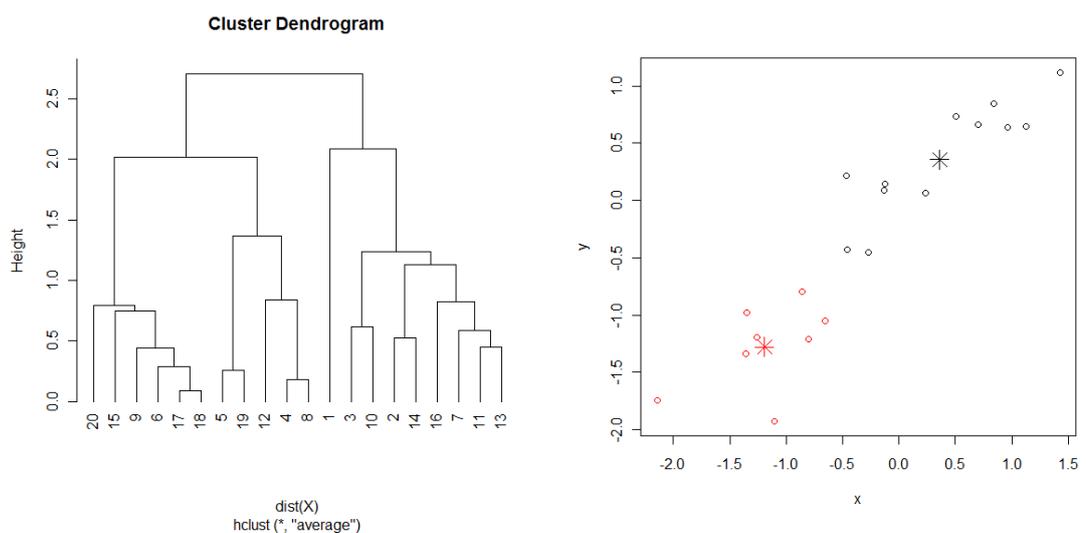


Figure 4.7: (a) Hierarchical and (b) k -means clustering on the $X^T X$ in section 3.3.3.

A focus is placed on hierarchical clustering due to the interpretational benefits of the dendrogram structure used in chapter 6 (see Figure 4.7a). Deciding to utilize a hierarchical cluster analysis, the user must decide whether to employ an agglomerative or divisive algorithm. Agglomerative (or ‘bottom up’) clustering involves all objects or variables being entered as single clusters in their own right. The algorithm merges the objects using a pre-specified measure of association until a single cluster comprising of all the objects is generated. Once a case is assigned to a cluster, it cannot be reassigned. The clusters can only grow in size. Divisive (or top down) clustering begins with a single cluster containing all the elements and splits the clusters until each element is separate to the rest.

Proximity and Linkage Criteria

The measure of proximity is a vital component of cluster analysis. It defines what it means for two observations or variables to be ‘similar’ (or equally important, to be ‘dissimilar’). There are a number of measures available to the user for continuous, categorical and binary variables (or mixtures). Green and Rao (1969) suggest that as data types and research objectives differ so greatly in applied research, there should be no “dominant” proximity measure. Instead, a range is often provided in statistical packages, usually split into distance and similarity measures. It is often cited as a difficult feature of clustering in application to relate these proximity measures to ‘real world’ concepts.

The linkage method defines the conditions to which the objects or clusters group to form a single cluster. Most clustering software will provide single linkage (i.e. elements are joined based on their closest distance to other elements) and complete linkage (i.e. based on furthest distance). However, these methods are sensitive to outliers and single linkage will tend to drag new elements into existing clusters rather than form separate clusters in their own right (known as ‘sequential joining’ or ‘chaining’). Criteria such as the median or centroid method may be preferable as they are based on a central measure of each cluster. Another popular criterion is Ward’s method (see Figure 4.7a). This method looks to minimize the group sums of squares within clusters, which is intended to retain as much information in the data whilst forming clusters of observations.

4.3.2 Visualizing the Results

The results of a hierarchical cluster analysis are traditionally visualized using a dendrogram (see Figure 4.7a). A dendrogram is a tree-like structure that displays the formation of

clusters as a series of horizontal and vertical lines. For the illustration in Figure 4.7a (from the R package *hclust*) the y-axis displays the measure of distance or similarity that each cluster is formed (i.e. the level of association amongst members of each cluster). The x-axis contains labels for the observations that are entered into the analysis. At the bottom of the dendrogram the objects are independent of one another and as the threshold for association is lowered (higher in the diagram), the clusters form.

When employing either an agglomerative or divisive clustering procedure a single cluster will be produced containing all the predictors and a level at which none of the variables are part of a cluster. Therefore, this can be viewed as a graduated scale in the strength of the cluster formations. Child (1990) criticises the subjectivity of not having a defined cut point to give a “sensible and representative number of clusters”. This can indeed be labelled subjective. However, the graduated clustering could be seen as an advantage in observing how the clusters develop and not needing to pre-define the number of clusters present (as in an EFA). This should aid with reproducibility as it would help to identify any consistency across study populations.

4.3.3 Variable Clustering

In Q-mode, observations are clustered in variable space, whereas in R-mode the data matrix is transposed and variables are clustered in object space (i.e. clustering rows or columns respectively). R-mode can be implemented using similarity measures such as Pearson’s r^2 or Hoeffding’s D statistic. Further analysis can be performed on a cluster analysis by applying a PCA to the predictors within the cluster and replacing with the first PC to analyze the structure of the ‘homogeneous’ groups (thus simplifying a PCA on the full set of variables). Clustering methodology is included in this chapter with a focus on variable clustering rather than the traditional object clustering. As a comparative method to EFA and PCA, the results of variable clustering would not produce abstract factors of loadings, but rather groups of observed variables. EFA is methodologically similar to a PCA. The difference lies in that the correlation matrix entered into the analysis is adjusted so that the diagonal elements equal the variances of the covariates. Therefore, the process must still invert the matrix, which is computationally heavy on examples consisting of a large number of variables. This problem is negated by clustering on similarity matrices and the interpretation eased by producing non-overlapping groups of covariates. R-mode clustering methods are often labelled discrete versions of factor analysis.

4.4 Shrinkage Regression

To this point the discussion has considered only OLS as a method of estimating regression coefficients of the linear model. However, the assumptions that underlie the method are regularly violated in application. The benefits of employing OLS are somewhat weakened and in some circumstances the method will simply not produce estimates (e.g. perfect collinearity). Alternative estimators have been developed for such situations. In this final section principal components regression (PCR) and partial least squares regression (PLS), both labelled shrinkage (or regularization) methods, are introduced and the benefits of application considered. These are considered particularly useful when data exhibits a high degree of collinearity. Shrinkage methods are labelled as such due to their variance shrinkage capability on coefficient estimates.

PLS is often described in a more complex form to PCR. However, the difference between the methods only exists in the weights used to construct the components. The methodological framework remains identical. Consider the bilinear model,

$$\mathbf{X} = \mathbf{CS}^T + \mathbf{F} \quad (4.15)$$

where \mathbf{S} contains the ‘weights’ of the decomposition (\mathbf{V} in PCR and \mathbf{P} in PLS) and \mathbf{C} contains the ‘loadings’ or latent variables (\mathbf{Z} in PCR and \mathbf{T} in PLS). Any unexplained \mathbf{X} -residuals are contained in \mathbf{F} . Geometrically, the loadings represent the new axes and the weights are a rotation from original axes. Loadings are entered into the linear regression model and coefficient estimates \mathbf{b}_* found by entering the latent variables as predictors,

$$\mathbf{y} = \mathbf{Cb}_* + \boldsymbol{\varepsilon} \quad (4.16)$$

The estimates can then be transformed back onto the original axes by a reverse rotation to provide an interpretable solution,

$$\mathbf{b} = \mathbf{Sb}_* \quad (4.17)$$

When the framework is presented in such a way it becomes simple to differentiate the motivations of these methods and understand key properties. This process is also the basis to the OLS procedure. Rotations have been added to the method, but the result is equivalent when all the variance in \mathbf{X} is retained (i.e. $\mathbf{F} = \mathbf{0}$). The length of the score vector represents the degree of variation explained in \mathbf{X} for each component.

4.4.1 Principal Component Regression

The bilinear decomposition of \mathbf{X} for PCR is as follows,

$$\mathbf{X} = \sum_{j=1}^k \mathbf{z}_j \mathbf{v}_j^T = \mathbf{ZV}^T \quad (4.18)$$

The PC's (\mathbf{Z}) are the latent variables (or loadings) and the eigenvectors (\mathbf{V}) are weights. There are no \mathbf{X} residuals when the full complement of components is retained, therefore no variance is lost in the bilinear decomposition of \mathbf{X} . The eigenvectors rotate the original axes to ensure the contribution to global variance is maximized on the first few components. The user will often wish to project only onto the first few PC's (i.e. discarding the $k - m$ components with the least explained variance). The linear regression model for PCR is as follows,

$$\hat{\mathbf{y}} = \mathbf{Z}_m \mathbf{b}_{\text{PCR}}^m + \boldsymbol{\varepsilon} \quad (4.19)$$

where \mathbf{Z}_m is an $n \times m$ matrix consisting of the first m PC's of $\mathbf{X}^T \mathbf{X}$. The $\mathbf{b}_{\text{PCR}}^m$ is an $m \times 1$ vector containing the m regression coefficients for the corresponding PC's. The parameter estimates $\mathbf{b}_{\text{PCR}}^m$ that minimize SSR on the rotated axes are found in the same way as the OLS estimates, but replacing the original covariates with the selected PC's.

$$\mathbf{b}_{\text{PCR}}^m = (\mathbf{Z}_m^T \mathbf{Z}_m)^{-1} \mathbf{Z}_m^T \mathbf{y} \quad (4.20)$$

Using eqn(4.18) and eqn(4.6) the following result is obtained,

$$\mathbf{Z}_m^T \mathbf{Z}_m = \mathbf{V}_m^T \mathbf{X}^T \mathbf{X} \mathbf{V}_m = \mathbf{V}_m^T \mathbf{V}_m \boldsymbol{\Lambda}_m \mathbf{V}_m^T \mathbf{V}_m = \boldsymbol{\Lambda}_m \quad (4.21)$$

The variance (or length) of the PC's are proportional to the eigenvalues and the PC's are uncorrelated (i.e. $\boldsymbol{\Lambda}_m$ is a diagonal $m \times m$ matrix). Using this property, eqn(4.20) can be rewritten in the following way,

$$\mathbf{b}_{\text{PCR}}^m = \boldsymbol{\Lambda}_m^{-1} \mathbf{V}_m^T \mathbf{X}^T \mathbf{y} \quad (4.22)$$

If the PC's are interpretable, the parameter estimates of the component scores may be of some use. However, a simple transformation can attain estimates on the original axes.

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{Z}_m \mathbf{b}_{\text{PCR}}^m = \mathbf{X}\mathbf{V}_m \mathbf{b}_{\text{PCR}}^m \quad \mathbf{b} = \mathbf{V}_m \mathbf{b}_{\text{PCR}}^m \quad (4.23)$$

$$(k \times 1) = (k \times m) \times (m \times 1)$$

When $m < k$, k regression coefficients are still obtained on the original \mathbf{X} column space (as demonstrated in eqn(4.23)), however they will no longer be equal to the OLS estimates. Variance has been discarded and is contained in the \mathbf{X} -residual matrix. Another way to consider the computation of the PCR components is to find the first eigenvector of the SVD matrix of \mathbf{F} (left and right singular vectors are identical). The first residual matrix would contain the centered data matrix \mathbf{X} . Once component scores have been found, these can be subtracted from the residual matrix and the process is repeated until the m components have been found. Whilst this is less elegant than the 'one step' SVD process, a sequential algorithm based on the residual matrix is the basis of many iterative processes used in statistical software (e.g. NIPALS - see Figure 4.9). The fitted values and direct computation of the regression coefficients in this space can be calculated as follows,

$$\hat{\mathbf{y}}_{\text{PCR}}^m = \mathbf{Z}_m \mathbf{b}_{\text{PCR}}^m = \mathbf{Z}_m (\mathbf{Z}_m^T \mathbf{Z}_m)^{-1} \mathbf{Z}_m^T \mathbf{y} = \mathbf{X} \mathbf{V}_m \mathbf{\Lambda}_m^{-1} \mathbf{V}_m^T \mathbf{X}^T \mathbf{y} \quad (4.24)$$

$$\mathbf{b}^m = \mathbf{V}_m \mathbf{\Lambda}_m^{-1} \mathbf{V}_m^T \mathbf{X}^T \mathbf{y} \quad (4.25)$$

This is of the form of the standard result for OLS presented in eqn(2.6). Vector geometry can be used to describe the relation of the reduced component estimates of PCR to the original OLS estimator. Phatak (1997) demonstrates that inserting $(\mathbf{X}^T \mathbf{X})(\mathbf{X}^T \mathbf{X})^{-1}$ into the equation clarifies the link between the estimators,

$$\begin{aligned} \mathbf{b}_{\text{PCR}}^m &= \mathbf{V}_m \mathbf{\Lambda}_m^{-1} \mathbf{V}_m^T (\mathbf{X}^T \mathbf{X})(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ &= \mathbf{V}_m \mathbf{\Lambda}_m^{-1} \mathbf{V}_m^T \mathbf{V}_m \mathbf{\Lambda}_m \mathbf{V}_m^T \mathbf{b}_{\text{OLS}} = \mathbf{V}_m \mathbf{V}_m^T \mathbf{b}_{\text{OLS}} \end{aligned} \quad (4.26)$$

As the matrix $\mathbf{V}_m \mathbf{V}_m^T$ is idempotent and symmetric, \mathbf{b}_{PCR} is an orthogonal projection of the \mathbf{b}_{OLS} solution. When $m = k$, the matrix \mathbf{V}_m will be equal to \mathbf{V} . As \mathbf{V} is orthonormal (i. e. $\mathbf{V}\mathbf{V}^T = \mathbf{I}$), when the complete set of k PC's are entered into the model the solution of eqn(4.26) is equivalent to OLS. Eqn(4.26) demonstrates that $\hat{\mathbf{y}}_{\text{PCR}}^m$ is simply an orthogonal projection of $\hat{\mathbf{y}}_{\text{OLS}}$ onto the space spanned by the first m PC's (see Figure 4.8). PCR is intended to increase the precision of the estimates by reducing noise and retaining artefact of interest from the original predictors.

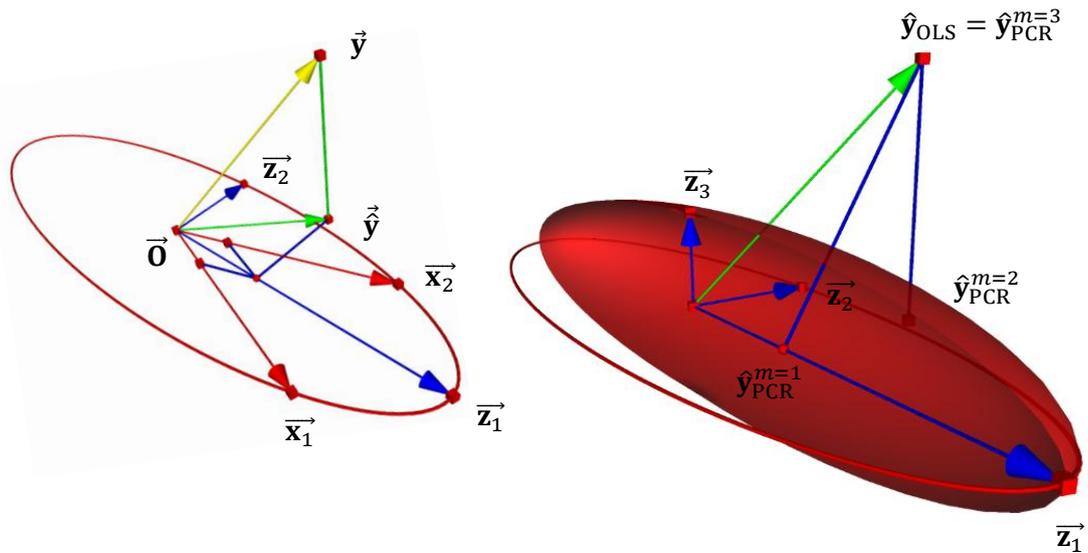


Figure 4.8: (a) PCR on two covariates with the first PC retained and (b) on three covariates demonstrating the $\hat{\vec{y}}_{PCR}^m$ projections onto regression surfaces.

The estimates are biased as information is omitted from the bilinear decomposition of \mathbf{X} (i.e. the bias-variance tradeoff). It is important to question whether those PC's with eigenvalues close to zero should be removed from the regression. As has been shown throughout the computation, the construction of the PC's is based only on the predictors in the regression (i.e. $\mathbf{X}^T\mathbf{X}$) and so it may not be the optimal approach to choose to generate the regression estimates. PLS provides a useful alternative for the researcher.

4.4.2 Partial Least Squares Regression

In practice, it may be that those PC's with small eigenvalues could be good predictors of the response. Therefore, removing these predictors from the regression model may have a damaging effect on the efficiency of the prediction and the stability of the estimation (Rao 1999). The components that are retained in the model (i.e. those with the greatest eigenvalues) may have a limited association with the response. An alternative shrinkage method based on a similar procedure is PLS. PLS was initially developed in the 1970's by Herman Wold as a tool for the social sciences but is now heavily used in chemometrics and bioinformatics. As opposed to the global variance maximizing process that defines PCR, PLS attempts to maximize \mathbf{X} and \mathbf{y} covariance (i.e. $\mathbf{X}^T\mathbf{y}$). The components are ordered by their maximal covariance with the response. As such, removing PLS components should retain the explanatory power of the model.

Similar to PCR, PLS seeks to obtain linear combinations of the manifest covariates that are orthogonal to one another (i.e. PLS components). A similar bilinear decomposition of \mathbf{X} can be formed such that,

$$\mathbf{X} = \mathbf{T}_m \mathbf{P}_m^T + \mathbf{F}_m \quad (4.27)$$

$\mathbf{T}_M = \mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_M$ are the m PLS latent variables (or scores) and $\mathbf{P}_M = \mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M$ are the weights (i.e. loadings) on the components. The OLS estimator is once again used to obtain PLS coefficients on the latent axes which are then transformed back to attain coefficients on the original axes.

$$\mathbf{y} = \mathbf{T}_M \mathbf{b}_{\text{PLS}}^M + \boldsymbol{\varepsilon}_M \quad (4.28)$$

$$\mathbf{b}^M = \mathbf{P} \mathbf{b}_{\text{PLS}}^M \quad (4.29)$$

Although this is presented for a single response (i.e. \mathbf{y}), the process can be extended to a multiple response example (see De Jong (1993)). Whilst the links between PLS and PCR should be clear from this bilinear presentation, the computation of the weight matrix \mathbf{P} in PLS is computationally heavier.

A linear combination has to be obtained for both the predictors and the response (Phatak and Dejong 1997). This is achieved by an iterative process to compute the optimal ‘weight vectors’ to maximize covariance between \mathbf{Y} and \mathbf{X} . Geometrically, \mathbf{y} is uni-dimensional which means that the process can only optimize for one component at a time. A sequential series of uni-dimensional projections maximizing covariance are required. Once a component is formed, the residual matrix for both \mathbf{X} and \mathbf{Y} (i.e. the variance not explained by the PLS component) is entered again as the data matrix and the procedure is repeated. Similar to PCR, a maximum of k components can be formed, whilst fewer can be retained (i.e. $m < k$) to reduce the dimensionality of the data. This has been considered particularly valuable in fields such as chemometrics and bioinformatics in which data often exists with a great number of variables and few observations. Due to the nature of the extraction these components are not necessarily ordered by explained variance in \mathbf{X} . By the iterative process of maximizing covariance, the method of obtaining the weights \mathbf{P} can be explained by considering the algorithms that perform PLS. Two of the most popular are NIPALS developed by Herman Wold (1973) and SIMPLS proposed by Sijmen De Jong (1993). For our single response, these algorithms are equivalent (Phatak and Dejong 1997). The general process is illustrated in Figure 4.9. Both \mathbf{X} and \mathbf{y} are initially mean centered to provide the starting point for either PLS algorithm.

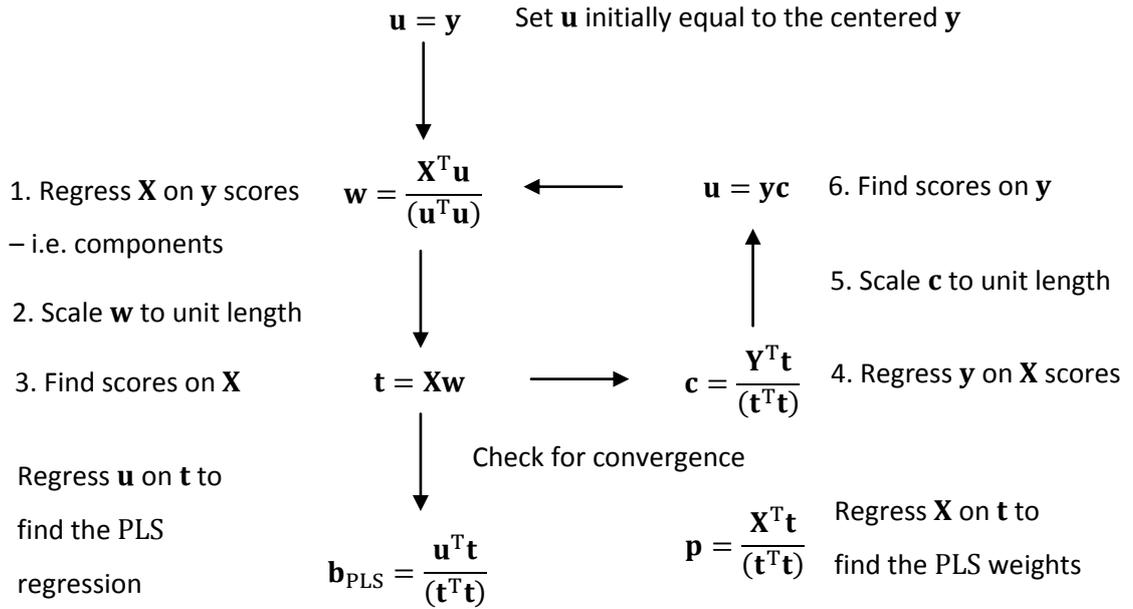


Figure 4.9: Graphical presentation of the NIPALS PLS algorithm.

The left side of Figure 4.9 finds the component scores on \mathbf{X} (i.e. \mathbf{T}_M) and the right side the component scores on \mathbf{y} (i.e. \mathbf{U}_M). The first factors \mathbf{t}_1 and \mathbf{u}_1 are simply weighted combinations of the mean centered \mathbf{X} and \mathbf{y} respectively (i.e. $\mathbf{t}_1 = \mathbf{X} \mathbf{w}_1$; $\mathbf{u}_1 = \mathbf{y} \mathbf{c}_1$) (Manne 1987). These weights can also be found as the initial pair of left and right eigenvectors from the SVD of $\mathbf{X}^T \mathbf{y}$. For the work in this thesis, \mathbf{y} is uni-dimensional and so the right singular vector is equal to one (this is later discussed in section 5.3).

$$\mathbf{w}_1 \mathbf{X}^T \mathbf{y} \mathbf{c}_1 = \mathbf{w}_1 \mathbf{X}^T \mathbf{y} = \mathbf{t}_1^T \mathbf{u}_1 = (n - 1) \text{cov}(\mathbf{t}_1, \mathbf{u}_1) \quad (4.30)$$

(Dejong 1993)

There are similarities to PCR in that \mathbf{w}_1 provides the rotation weight on \mathbf{X} and \mathbf{c}_1 on \mathbf{y} , however only the covariate axes are rotated as the response is univariate. Convergence is found when the change in \mathbf{t} between iterations is smaller than a set percentage. When the PLS component that maximizes $\mathbf{t}_1^T \mathbf{u}_1$ is found, the matrices \mathbf{X} and \mathbf{y} are deflated by retaining only the residuals from the regression of \mathbf{X} and \mathbf{y} on \mathbf{t}_1 and the process is repeated. This loop continues until the full M orthogonal components \mathbf{T}_M are extracted. The vector \mathbf{y} is projected onto the first m PLS components extracted from the data,

$$\mathbf{b}_{\text{PLS}}^m = (\mathbf{T}_m^T \mathbf{T}_m)^{-1} \mathbf{T}_m^T \mathbf{y} \quad (4.31)$$

Each weight \mathbf{w} is calculated from the residual matrix \mathbf{X} produced after each projection,

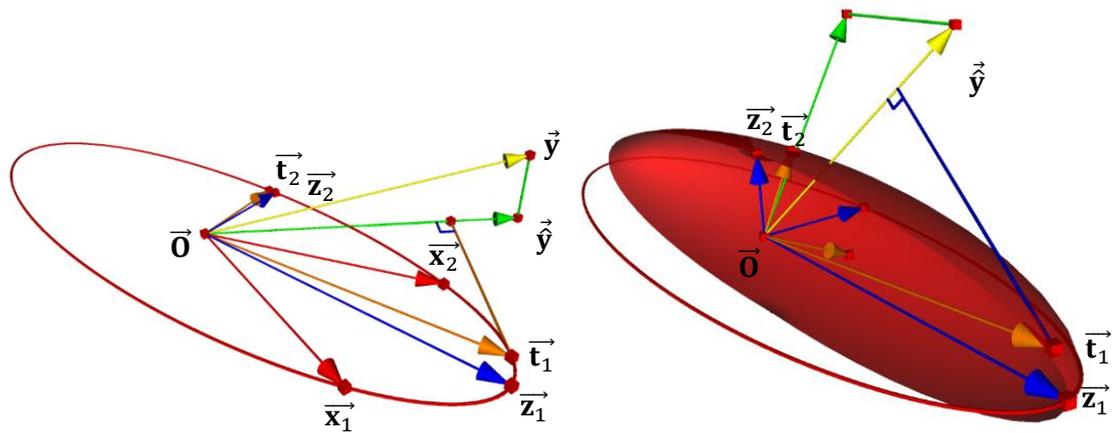


Figure 4.10: (a) The covariance maximizing procedure of PLS on the first component of a two predictor model and (b) the PLS axes of a three predictor model.

The first PLS component is found at the point in which the plane normal to the projected response \hat{y} is at a tangent to the covariance ellipsoid (see Figure 4.10b). The second and third components must lie on the plane that passes through the origin and is orthogonal to the first PLS component. \hat{y} is projected orthogonally onto this plane and the tangent to the ellipse is used to form the second PLS component. The third is orthogonal to the first pair of components lying on the ellipse. Notice that in this illustration, the third PLS component explains more \mathbf{X} -variance than the second. Unlike PCR, PLS components represent the covariance structure and so they are not necessarily ordered by variance explained in \mathbf{X} .

4.5 Conclusions

This chapter introduces a small selection of regression methods which are intended as a preliminary discussion for applications in later chapters. The methods of EFA, PCA and clustering represent approaches to understand the underlying structure amongst a set of collinear covariates. They each experience benefits and limitations for such a purpose and provide ideas for which new approaches can be developed for similar applications. PCR and PLS are useful estimators for highly collinear data. The vector geometry of PLS in particular provides the motivation for a new collinearity index that is developed in chapter 5. The covariance maximizing feature links in well with our discussion regarding the role of the response. PCA and PCR can be viewed as suffering similar limitations to correlation based measures such as the VIF and CI. The methods of PCR and PLS also provide a useful tool for analyzing perfectly collinear data which will be discussed in chapters 7-8.

5. Developing an Index to Measure the Impact of Collinearity

The merits of a ‘correlation based’ collinearity index were considered in chapter. The common limitation of these methods is that they describe the degree of collinearity amongst predictors, rather than indicate its potential *impact* on the modelling process. An alternative was considered in the VIF · VDF measure which places this impact in the context of the model. When the deflation effect of the response on the variance is accounted for, the initial assessment of ‘problematic’ collinearity may be somewhat reduced – rendering the use of a correlation based ‘rule of thumb’ to indicate ‘severe’ collinearity a rather crude technique. Chapter 5 focuses on the challenge of comparing two sets of regression estimates, for use in model assessment and model selection. The impact of collinearity on the point estimate will be considered in addition to the variance. Vector geometry is utilized to generate ideas for how such an index could be formed.

An example application in body fat analysis is used to illustrate the development of our index and demonstrate a potential use. The intention is to place the theory in an applied epidemiologic setting and to be able to provide an interpretation for the results. The latter point will be stressed in particular as a statistical measure can only achieve so much without the use of clinical *a priori* information to define the setting. Comparisons will be made to the VIF and VDF indices that were discussed in chapter 3. The focus in particular is on the relationship between the response variable and the correlated covariates. The VIF · VDF considered this effect on the variance of the coefficient estimates, but this is only one aspect of common collinearity ‘symptoms’. Parameter estimates considered to have ‘incorrect’ sign or implausible magnitude are often features associated with collinear variables. Although a sign change is considered a clinical rather than statistical ‘problem’, the magnitude of the deviation of the coefficients could potentially be used as a tool for assessing the impact of collinearity. It could also provide a ‘global’ comparison between nested models for a potential use in model building.

5.1 Example Application

Body fat has been widely shown to be an important risk factor for diabetes and cardiovascular disease (Gregg et al. 2005). A recent study by Romero-Corral et al. (2010) suggested that patients with a normal body mass index (BMI), but a high body fat content were still at high risk of cardiovascular mortality. These patients were labelled “normal weight obese” - highlighting that it is not sufficient to diagnose a subject based on their weight or physical appearance alone. In this chapter data are analyzed from a study by Penrose et al. (1985). The study records the percentage body fat, weight and height, along with 10 body circumference measures of 252 men.

	Body fat	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm
Body fat												
Weight	0.61											
Height	-0.03	0.49										
Neck	0.49	0.83	0.32									
Chest	0.70	0.89	0.23	0.78								
Abdomen	0.81	0.89	0.19	0.75	0.92							
Hip	0.63	0.94	0.37	0.73	0.83	0.87						
Thigh	0.56	0.87	0.34	0.70	0.73	0.77	0.90					
Knee	0.51	0.85	0.50	0.67	0.72	0.74	0.82	0.80				
Ankle	0.27	0.61	0.39	0.48	0.48	0.45	0.56	0.54	0.61			
Biceps	0.49	0.80	0.32	0.73	0.73	0.68	0.74	0.76	0.68	0.48		
Forearm	0.36	0.63	0.32	0.62	0.58	0.50	0.55	0.57	0.56	0.42	0.68	
Wrist	0.35	0.73	0.40	0.74	0.66	0.62	0.63	0.56	0.66	0.57	0.63	0.59

Table 5.1: Pearson correlations for the Penrose data.

Body fat was estimated using hydrodensitometry (i.e. underwater weighing). The process involves each individual being lowered into a tank of water, expelling the air from their lungs and remaining still whilst measures are taken using adapted scales. This process is repeated three times and averaged to attain a measure of percentage body fat.

Whilst hydrodensitometry is considered highly accurate (Hortobagyi et al. 1992), it is inconvenient and uncomfortable in practice. The covariates in this study are used to explore body composition using these external measurements of body circumference. The aim is to see how these highly correlated variables can be used to create an ‘optimal’ model to explain percentage body fat. This would potentially allow for a reduction in the number of measurements required, reducing study length, cost and participant burden whilst maintaining measurements that most accurately represent total body fat. Such studies have become particularly important with a recent focus on body fat distribution (Canoy 2010;Chen et al. 2011;Mundi et al. 2010).

5.2 Potential Applications of the Index

As with the use of any statistical measure, a researcher should not rely exclusively on study data to assess the validity of a model. *Prior* information should be incorporated into the analysis to tailor the assessment to a particular discipline or setting. If the impact of collinearity is considered to be a 'problem', then to assess that problem we need an idea of what the population structure is. For instance, suppose \mathbf{x}_1 and \mathbf{x}_2 are two uncorrelated predictors in a population, but the sample values are correlated. If we believe the population values to be uncorrelated (i.e. an *a priori* assumption), the expectation is that the multivariable coefficient estimates of the response regressed on both \mathbf{x}_1 and \mathbf{x}_2 are unchanged compared to their univariable regression coefficients. The fact that the sample values are correlated causes the univariable and multivariable estimates to differ.

We might believe there to be a correlation in the population, and wish to 'use' this as adjustment for confounding. A weaker or stronger correlation may be observed in the sample, again due to sampling variation. Now we wish to know the extent of the discrepancy between the 'true' adjustment (in the population) and the observed adjustment (in the sample). We may only 'guesstimate' (from external data) what the population correlation is supposed to be, but nevertheless be interested to know the impact of the sampling variation on our estimated multivariable coefficients in the sample compared to that in the population, and perhaps to have this assessed in relation to the discrepancy between the population zero correlation scenario and the true population correlation, i.e. relative to the ideal adjustment.

Another potential use for such an index is in model selection. In clinical research we may be given a number of clinically plausible predictors to include in a regression model and be required to produce a 'minimum subset' of covariates. The index is performing a similar task to the previous example. It is still comparing the disparity between two sets of estimates, but now the intention is to compare models consisting of different predictors. The index would need to provide an assessment of the impact of collinearity that is being introduced into the model by including these predictors. Further to that, the clinician may insist on biologically relevant predictors to be included in the final model and so a range of 'minimum subsets' may be produced based on the initial requirements and external knowledge. This index may take the form of comparing each model to a baseline set of estimates (formed from univariable models), or it will compare two sets of multivariable models (i.e. sample estimates in both cases).

The form of the index does not change – only how it will be employed. Whichever example the theory is built around, we need to assess the extent of disparity between two sets of point estimates with reference to collinearity. It would be an unrealistic target for this chapter to fully address each of these challenges individually, but the intention is to build a general apparatus that could be developed to perform such operations in the future. There is also the additional consideration when incorporating the response variable of the likely deviation under sampling variation. The coefficients gained from least squares regression are unbiased (under the assumptions outlined in section 2.2.2), however in a single sample case this is of little use if the variance is high. This chapter describes the development of a ‘global’ measure of estimate deviation, approaches to incorporate uncertainty in the point estimates and considers the individual roles of covariates in contributing to a ‘global’ impact on the modelling process.

5.3 Incorporating the Response into Existing Diagnostics

It would seem natural to seek parallels between existing indices, such as the CI and VIF, to developing new ‘covariance based’ measures. The VIF · VDF has achieved this to provide a simple extension to the VIF. However, like the VIF, this index will not identify which variables are involved in particular dependencies. The CI and VDP provide such a measure on the covariance of the predictors. Further extension is required for the VIF · VDF to maintain the goal of incorporating information of the response variable. First consider the scaled covariance between predictor and response $\mathbf{X}^T \mathbf{y}$. This represents a vector of covariance’s (see eqn(5.1)) and a geometrical projection of $\vec{\mathbf{X}}$ onto $\vec{\mathbf{y}}$ (see Figure 5.1a),

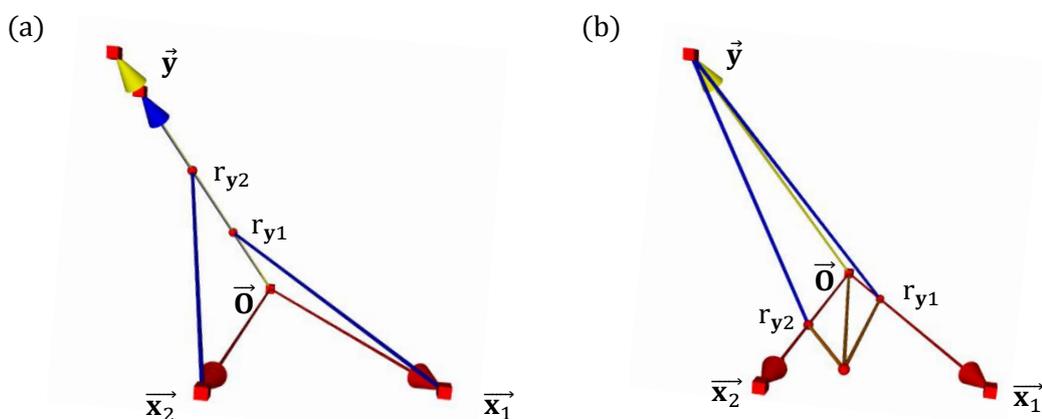


Figure 5.1: Geometrical projection of (a) \mathbf{X} onto \mathbf{y} and (b) \mathbf{y} onto \mathbf{X} .

$$\mathbf{X}^T \mathbf{y} = [\text{Cov}(\mathbf{X}_1, \mathbf{y}) \quad \text{Cov}(\mathbf{X}_2, \mathbf{y}) \quad \dots \quad \text{Cov}(\mathbf{X}_k, \mathbf{y})]^T \quad (5.1)$$

This vector (with a single response) corresponds to the following SVD,

$$\mathbf{X}^T \mathbf{y} = \mathbf{L} \mathbf{M} \mathbf{N}^T \quad (5.2)$$

where \mathbf{L} is a $k \times w$ matrix consisting of the left singular vectors of $\mathbf{X}^T \mathbf{y}$, \mathbf{M} is a $w \times w$ diagonal matrix of eigenvalues and \mathbf{N} is a $w \times w$ orthogonal matrix containing the right singular vector of $\mathbf{X}^T \mathbf{y}$. Whilst the $\mathbf{X}^T \mathbf{X}$ matrix is represented geometrically by a hyper-ellipsoid, the $\mathbf{X}^T \mathbf{y}$ only contains a single direction vector. It is a projection of \mathbf{X} onto the surface of the response vector \mathbf{y} (see Figure 5.1a). As \mathbf{y} is uni-dimensional, each projected vector of \mathbf{X} onto \mathbf{y} will lie on the same single dimension (i.e. $w = 1$). Therefore, \mathbf{M} contains only one eigenvalue. This is the length of the co-variance vector scaled by $(n - 1)$,

$$\mathbf{M}_{11} = (n - 1) \sqrt{\text{Cov}(\mathbf{x}_1, \mathbf{y})^2 + \text{Cov}(\mathbf{x}_2, \mathbf{y})^2 + \dots + \text{Cov}(\mathbf{x}_k, \mathbf{y})^2} \quad (5.3)$$

The right singular vector \mathbf{N} is the eigenvector of the matrix $\mathbf{y}^T \mathbf{X} \mathbf{X}^T \mathbf{y}$. This is a scalar value, therefore the eigenvector is equal to 1. The left singular vector \mathbf{L} is the eigenvector of $\mathbf{X}^T \mathbf{y} \mathbf{y}^T \mathbf{X}$. This will take the form of a normalized vector of co-variances (i.e. scaled to unit length). It is a vector weighted by the original covariance elements. This represents the angle between the first principal axis of the $\mathbf{X}^T \mathbf{X}$ matrix and the \mathbf{y} vector (i.e. the rotation to the projection vector). The complete SVD for the $\mathbf{X}^T \mathbf{y}$ is as follows,

$$\mathbf{X}^T \mathbf{y} = \mathbf{L} \mathbf{M} \mathbf{N}^T = \begin{bmatrix} \text{Cov}(\mathbf{x}_1, \mathbf{y}) / \|\mathbf{L}\| \\ \text{Cov}(\mathbf{x}_2, \mathbf{y}) / \|\mathbf{L}\| \\ \vdots \\ \text{Cov}(\mathbf{x}_k, \mathbf{y}) / \|\mathbf{L}\| \end{bmatrix} \left[(n - 1) \sum_{j=1}^k \sqrt{\text{Cov}(\mathbf{x}_j, \mathbf{y})^2} \right] [\mathbf{1}] \quad (5.4)$$

The information for this deconstruction is contained in the $\mathbf{X}^T \mathbf{y} \mathbf{y}^T \mathbf{X}$ matrix (as the right singular vector gives the same eigenvalue and an arbitrary eigenvector). However, the matrix is still uni-dimensional and so only the first eigenvalue will be non-zero (i.e. \mathbf{M}_{11}^2). To work with this matrix would provide only the eigenvectors orthogonal to the first and so provides little more useful information than the $\mathbf{X}^T \mathbf{y}$ can provide for a diagnostic criteria. A direct extension to $\mathbf{X}^T \mathbf{X}$ of this form would seem of little use.

To place $\mathbf{X}^T \mathbf{y}$ in the same space as $\mathbf{X}^T \mathbf{X}$ we may consider the reverse projection - $\mathbf{X}^T \mathbf{y}$ will give the same SVD as $\mathbf{y}^T \mathbf{X}$ and therefore could be viewed as an orthogonal

projection of \mathbf{y} onto \mathbf{X} (see Figure 5.1b). The distance from the origin to the point of projection on \mathbf{X} is equal to each of (1) the covariance between \mathbf{X} and \mathbf{y} , (2) the least squares coefficient from the univariable model and (3) the square root of the coefficient of determination from the same univariable model. Notice the similarities with VIF geometry presented in Figure 3.9, with this process now obtaining R_y^2 for the VDF. Projecting \mathbf{X} onto \mathbf{y} or \mathbf{y} onto \mathbf{X} is analogous to obtaining an equal VIF in the bivariable model.

In Figure 5.2, the two points of the projected response are projected once again parallel to the covariates to form a composite direction vector (shown in orange). This vector represents a weighted version of a principal axis corresponding to the covariance elements of the $\mathbf{X}^T\mathbf{y}$ matrix (instead of $\mathbf{X}^T\mathbf{X}$ in the principal axis transformation).

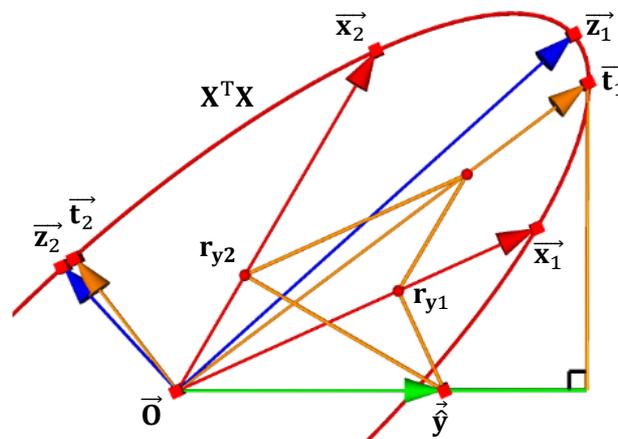


Figure 5.2: Covariance maximizing procedure for two covariates.

The construction of principal axes maximizing variance (shown in blue) and covariance (shown in orange) are demonstrated in Figure 5.2. The length of the blue vectors represent the eigenvalues used for the CI. In comparison, the first principal axis maximizing covariance is of shorter length, demonstrating the compromise in explained variance on the predictors with explaining variance of the response. Note that this covariance vector can also be found by the orthogonal projection of $\hat{\mathbf{y}}$ (or \mathbf{y}) onto the $\mathbf{X}^T\mathbf{X}$ ellipse. The r_{yj} indicated in this model are the regression coefficients from the univariable models (i.e. an orthogonal projection of $\hat{\mathbf{y}}$ onto the covariate vectors). The covariance maximizing axes in the bivariable model are identical to the two components found in a PLS analysis. The composite vector formed in the direction of the two univariable r_{yj} projections is the first PLS component, labelled \mathbf{t}_1 . This hints at making use of the covariance maximizing axes formed from a PLS analysis in developing a new collinearity index.

5.4 Assessment of Deviation

To measure the disparity between two sets of estimates, they need to be placed in a comparable setting. Vector geometry can do so by considering both in a common space. The first step is to compare the estimates from two univariable models to the multivariable model formed from collinear predictors. The ‘traditional’ geometrical representation would be to project $\vec{\hat{y}}$ orthogonally to the vectors to find the univariable point estimates b'_j and parallel to the opposite covariate to find the multivariable estimates b_j (this is illustrated in Figure 5.3).

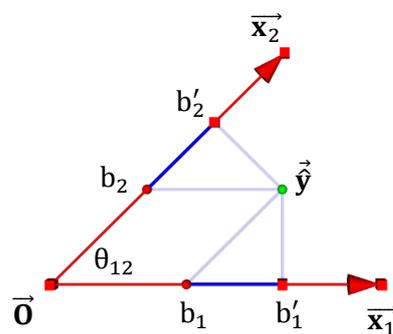


Figure 5.3: The projection of \hat{y} to generate univariable and multivariable estimates.

The distance between the projections would indicate the change in regression coefficients on each predictor ($b'_j - b_j$) (i.e. shown in blue). To gain a ‘global’ measure of the impact on the model the individual univariable estimates are presented as a single multivariable model involving uncorrelated predictors. The position of $\vec{\hat{y}'}$ representing the projected response for the univariable model is the product of the change in regression coefficients, relative to the collinearity in the model (i.e. shown in green in Figure 5.4).

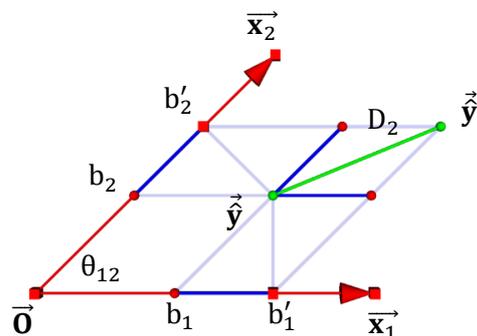
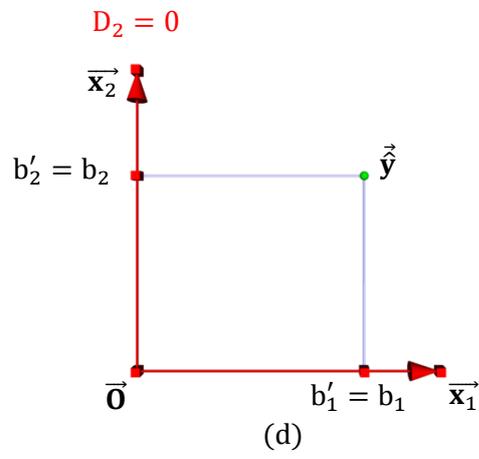
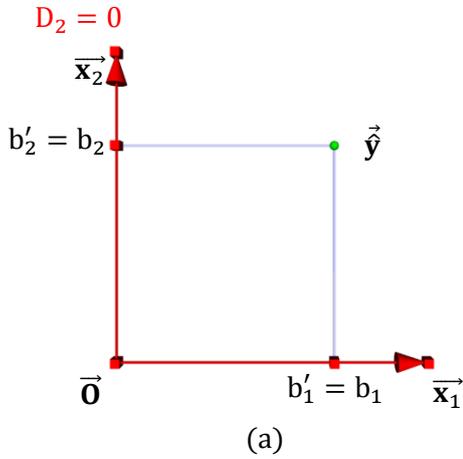


Figure 5.4: Placing uncorrelated univariable estimates onto collinear axes.

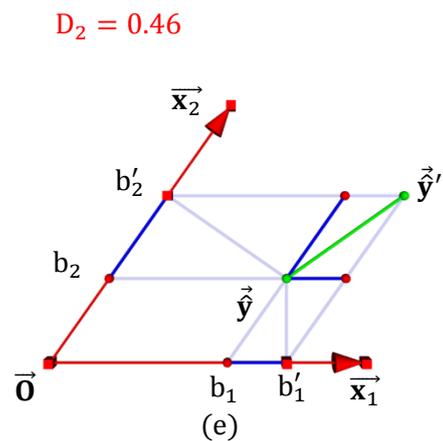
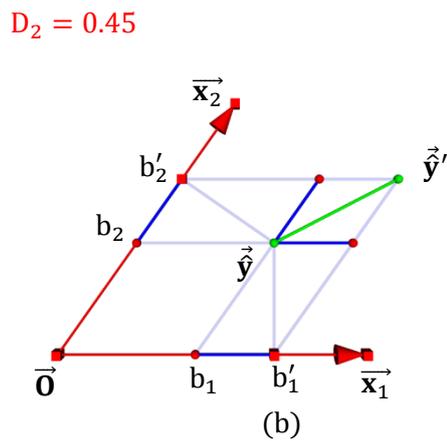
This removes the effect of the non-orthogonal projection and places the estimates on common collinear axes. This will generate a ‘global’ measure of deviation (shown in green).

$r_{y1} = 0.7, r_{y2} = 0.7$	$r_{y1} = 0.75, r_{y2} = 0.65$
------------------------------	--------------------------------

$\theta_{12} = 90^\circ, \text{cor}(x_1, x_2) = 0$



$\theta_{12} = 55^\circ, \text{cor}(x_1, x_2) = 0.57$



$\theta_{12} = 20^\circ, \text{cor}(x_1, x_2) = 0.94$

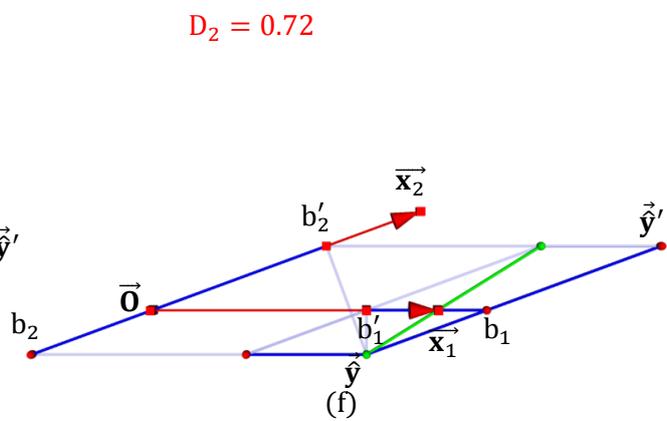
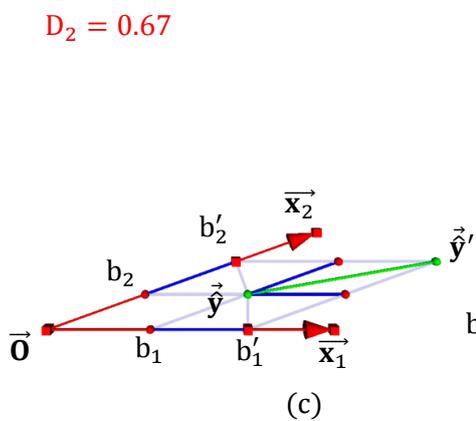


Figure 5.5: Deviation of point estimates in different collinearity conditions.

Figure 5.6 illustrates an adjustment in the discrepancy measure (labelled D_2 for the bivariable example and more generally D_k for k predictors) under changing collinearity conditions.

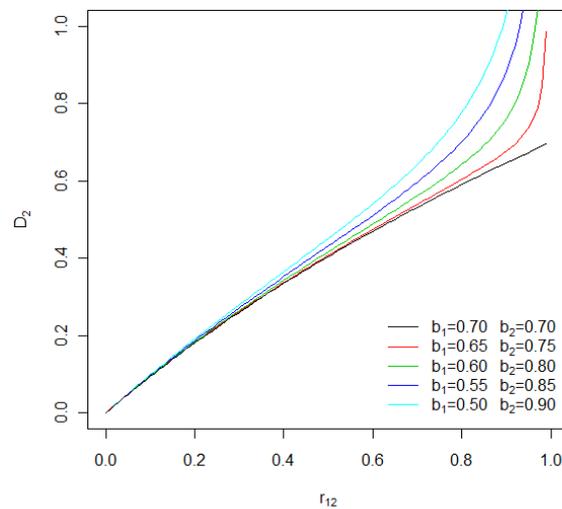


Figure 5.6: Change in D_2 under varying collinearity conditions.

When the covariates are independent of one another (i.e. $r_{12} = 0$), the regression coefficients for the univariable and multivariable models are identical. Therefore, the measure of deviation at baseline is always zero (i.e. $D_2 = 0$). As the correlation is increased in the equal univariable coefficient case (e.g. Figure 5.5 – (a), (b) and (c)), the D_2 index tends toward the value of the common univariable coefficient. At this point, the predictors are entirely overlapped (i.e. perfect collinearity) and the variance of the response explained by the multivariable model is equivalent to that being explained by either univariable model (e.g. the black line in Figure 5.6).

In the model with unequal univariable coefficients (e.g. Figure 5.5 – (d), (e) and (f)), D_2 will tend toward infinity. At high levels of collinearity, the disparity introduced into the model can far exceed that of the ‘equal coefficient’ case. Example (f) of Figure 5.5 demonstrates a change of sign on x_2 , whilst the point estimate of x_1 is enhanced. Comparatively, example (c) has the same VIF, however does not exhibit a change of sign and demonstrates less movement in the coefficients. This highlights an advantage of the D-index over the VIF. The change of sign is not of particular interest statistically, but it represents a potential change in the clinical interpretation. Under sampling variation, these deviations can become enhanced or diminished with a potential impact on the conclusions of the study.

5.4.1 Calculating the Index

The calculation of D_2 can be demonstrated by splitting the 'global' deviation into two components (labelled α_2 and β_2 in Figure 5.7).

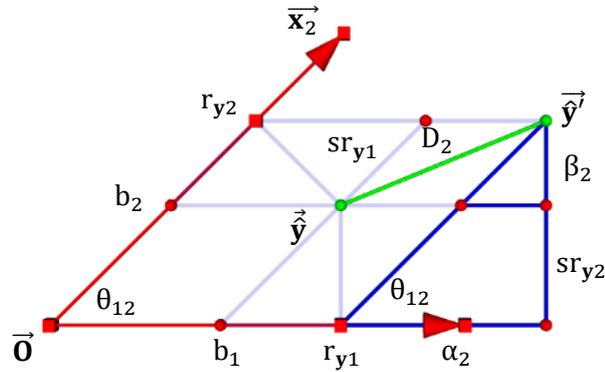


Figure 5.7: Computation of the D-index.

The first is parallel to \bar{x}_1 (α_2) and the second orthogonal to \bar{x}_1 (β_2). The derivation of the D_2 index (using the triangle highlighted in blue in Figure 5.7) can be shown as follows,

$$\alpha_2 = r_{y2} \cos(\theta_{12}) = r_{12}r_{y2} \quad (5.5)$$

$$\begin{aligned} \beta_2 &= r_{y2} \sin(\theta_{12}) - sr_{y2} \\ &= r_{y2} \sqrt{1 - r_{12}^2} - \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{1 - r_{12}^2}} \\ &= \frac{r_{y2}(1 - r_{12}^2) - r_{y2} + r_{y1}r_{12}}{\sqrt{1 - r_{12}^2}} \\ &= r_{12} \cdot \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{1 - r_{12}^2}} = r_{12}sr_{y1} \end{aligned} \quad (5.6)$$

The index is then calculated as the squared sum of orthogonal components,

$$D_2^2 = \alpha_2^2 + \beta_2^2 = (r_{12} r_{y2})^2 + (r_{12} sr_{y1})^2 = r_{12}^2 R_y^2 \quad (5.7)$$

$$D_2 = r_{12} R_y \quad (5.8)$$

For the bivariable comparison to univariable model, D_2 is equal to $R_y \cdot r_{12}$ (or $D_2^2 = (R_y \cdot r_{12})^2 = R_y^2 \cdot r_{12}^2$ - squared to be proportional in magnitude to the VIF). When the predictors are orthogonal, D_2^2 will equal zero, as regardless of the variance explained by the model there is no impact of collinearity in the change between the estimates – i.e. \hat{y} and \hat{y}' are identical. Similarly, when $R_y = 0$, including both of the predictors in the model has explained zero variance in the response - therefore collinearity can have no impact on the model coefficients (contrary to any indication from the VIF). As the collinearity increases, its relative impact on the predictors is modified by the variance explained in the bivariable model. If the R_y is high, the index will indicate that the collinearity impact is greater than for a low bivariable R_y on the same collinear predictors.

5.4.2 Interpretation of the Index

The use of vector geometry to generate this index provides insights into its interpretation and flexibility for future development. The composite direction vector formed from the two univariable regression coefficients is in the covariance maximizing direction on a single dimension (or a single component PLS model – see Figure 5.2). This is the vector labelled \hat{y}' . The \hat{y} vector represents the OLS estimation (i.e. an orthogonal projection of the response onto the regression surface). By definition, this is the covariance maximizing direction in the bivariable model and is equivalent to a PLS with maximum components retained. Therefore, D_2 is measuring the distance between an uncorrelated composite vector on a single dimension and the collinear predictors of the bivariable model – i.e. a comparison of PLS models. This can represent the impact of collinearity in moving from an uncorrelated *prior* to a correlated sample.

The length of the vector $\overrightarrow{\hat{y}'}$ is equal to the summation of the two univariable r_y^2 estimates, relative to the degree of collinearity (i.e. by the cosine rule $2r_{y1}r_{y2}r_{12}$ is added). The length of $\overrightarrow{\hat{y}}$ is equal to the R_y found in the bivariable model. Therefore, the magnitudes of the vectors follow this interpretation of a univariable to bivariable comparison of models. The work in this chapter initially focuses on this comparison (i.e. a baseline of uncorrelated predictors), however the interpretation will follow to comparing to a correlated baseline *prior* (i.e. non-orthogonal projections to represent *a priori* knowledge or a model with correlated predictors). Potential extensions and uses for the D-index will be described throughout this and later chapters.

5.4.3 Comparison to a Variance Based Index

A comparison of D_2^2 and VIF values for each bivariable model in the body fat data is displayed in Table 5.2, with the former in the lower triangle and the latter in the upper.

	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
Height	1	1.11	1.05	1.04	1.16	1.13	1.33	1.18	1.11	1.12	1.19
Neck	0.03	1	2.60	2.32	2.17	1.94	1.83	1.30	2.15	1.64	2.25
Chest	0.03	0.31	1	6.20	3.20	2.14	2.07	1.30	2.13	1.51	1.77
Abdomen	0.03	0.40	0.56	1	4.24	2.43	2.19	1.26	1.88	1.34	1.62
Hip	0.06	0.21	0.34	0.53	1	5.09	3.11	1.45	2.21	1.42	1.66
Thigh	0.04	0.16	0.27	0.39	0.31	1	2.77	1.41	2.38	1.47	1.45
Knee	0.09	0.14	0.26	0.37	0.27	0.21	1	1.6	1.85	1.45	1.79
Ankle	0.01	0.06	0.12	0.14	0.13	0.09	0.10	1	1.31	1.21	1.47
Biceps	0.03	0.15	0.26	0.31	0.21	0.19	0.14	0.06	1	1.85	1.67
Forearm	0.02	0.10	0.17	0.17	0.12	0.10	0.08	0.03	0.11	1	1.52
Wrist	0.02	0.13	0.23	0.27	0.16	0.10	0.11	0.04	0.10	0.05	1

Table 5.2: D_2^2 in the lower triangle with VIF's in the upper triangle.

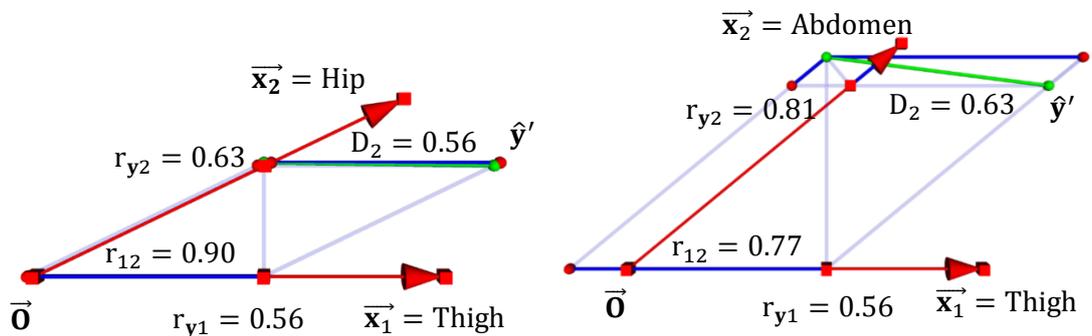


Figure 5.8: D_2^2 and VIF for two pairs of predictors in the body fat data.

There are seemingly a number of discrepancies regarding the measurement of collinearity between these indices (as expected between correlation and covariance based indices). Figure 5.8 illustrates one example to demonstrate the geometrical interpretation of the impact. The VIF is larger for the Thigh/Hip model than the Thigh/Abdomen model to reflect the increased correlation between the predictors. In comparison, D_2 indicates that the impact of collinearity is greater in the Thigh/Abdomen model due to the increased R_y^2 . If the VIF is adjusted by VDF, the Thigh/Hip model would remain the higher value as R_y^2 would have a greater deflation effect on the variance.

For the VIF · VDF, the response is considered a deflating element of the variance on each predictor. However, for the D-index a greater R_y , coupled with high collinearity amongst the predictors, has the opposite effect.

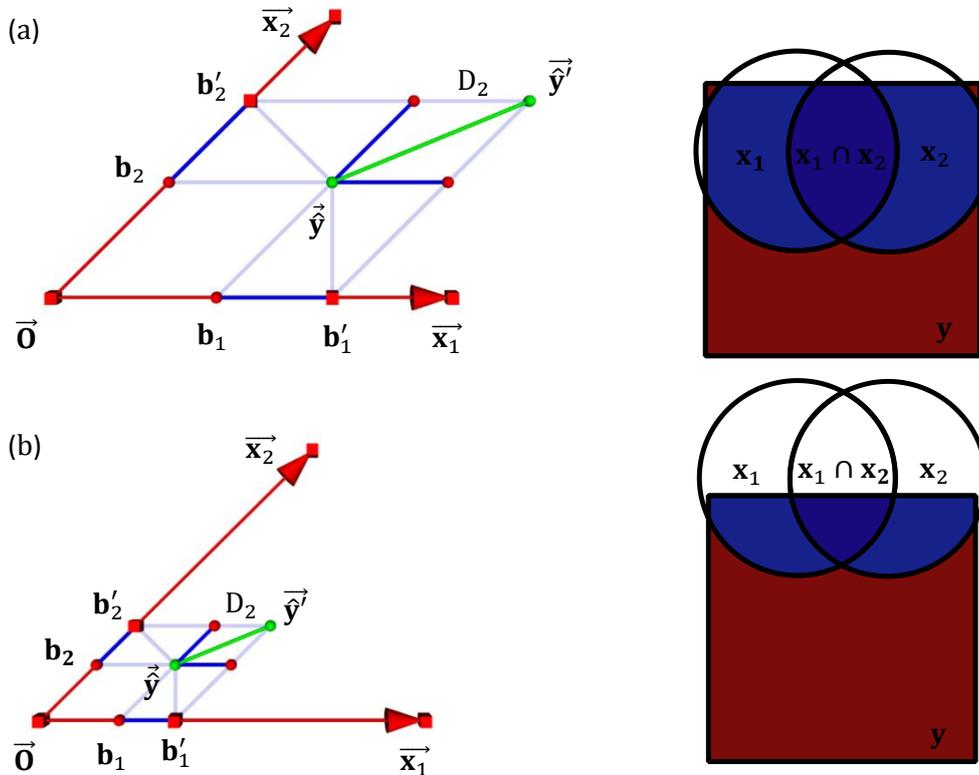


Figure 5.9: The role of the response in the impact of collinearity.

The VIF is a comparison to baseline of unity for orthogonal predictors and r_{12} is a comparison to baseline of zero. The VIF is adjusted by VDF (for VIF·VDF), whilst r_{12} is adjusted by R_y (for D_2) to incorporate the response. At baseline (i.e. orthogonality), the VIF · VDF will remain at unity (if it were considered a ‘global’ measure – i.e. $R_y^2 = 0$ for the baseline) whilst D_2 will remain zero. The measures are used to represent an *impact* on the univariable predictors when they are entered together in a bivariable model. If the covariates are highly collinear, but also explain a high proportion of the response, then the variance explained attributed to each predictor reduces by a greater amount moving from the univariable to bivariable regression models (reflected in the point estimates). However, the greater the variance explained in y , the greater the precision of the estimates in the bivariable model. D_2 is highlighting a greater deviation in the point estimates (an effect of sampling variation if we believe the predictors to be orthogonal in the population), whilst the VIF · VDF is an increase in precision of the estimates. The difference rests on our conceptual understanding and is not necessarily measuring the extent of a ‘problem’.

A key assumption in this discussion is that $VIF \cdot VDF$ is represented as a 'global' measure. This is not actually reflective of the true index. In the bivariable case the VIF is always equal for both predictors (as demonstrated in Figure 3.9), however this is not necessarily true of models when $k > 2$. Therefore, the baseline should be presented as the VIF adjusted by VDF for each predictor, which would not be unity if r_{y1} or $r_{y2} > 0$. It is important to stress this difference in the measures to facilitate the progression of the index discussion and development. It demonstrates the novelty of generating a 'global' index in D_k . Comparisons must be made to the $VIF \cdot VDF$ in the current discussion as there does not appear a direct alternative in diagnostic literature that provides such a measure. In examples beyond the bivariable model (i.e. $k > 2$), representing the $VIF \cdot VDF$ as a global measure would require some form of generalized R_x . In comparison, the D_2 can be extended by utilizing vector geometry (this is demonstrated in section 5.7).

The use of vector geometry to generate this measure provides a general framework for its development, but it very much remains a statistical index. The application and clinical assumptions are crucial to define its utility in practice. For instance, if the predictors are believed to be orthogonal in the population, the D_2 would provide an indication of the disparity in estimates under sampling variation. If instead some correlation was assumed in the population model, then the baseline would need to be adjusted to reflect this correlation. From a model building perspective, D_2 represents an overall change in the estimates (either beneficial or detrimental to the estimation) after including a new predictor. For this purpose, the statistical measure would require more information. Which predictors would generate the greater impact and which would it appear more beneficial (from a purely statistical perspective) to retain in the model?

The links between the D-index and PLS components aid with the interpretation of this measure, but it may also suggest an additional use for the index to those previously discussed. For example, with a greater overlap of the collinear variables to the response (i.e. a greater D_2), a single component PLS may become preferable over the two component model. Further to this, a single component model is in the direction of the univariable predictors (assumed to be uncorrelated), whilst the full component model is equivalent of an OLS regression. Therefore, the D-index could be adjusted to measure deviation between additional PLS component estimates by considering different projections in the vector geometry. This would reflect the impact of additional collinearity with each component added and could indicate an optimal number based on an interpretive rather than predictive criteria (e.g. such as the PRESS – see Allen (1974)).

5.5 Deconstructing the Impact of Collinearity

This section aims to summarize the information provided by the VIF, VIF·VDF and D_2 index and compare the utility of each to deconstruct the impact of collinearity. The regression model consisting of the predictors ‘weight’ and ‘abdomen’ in the body fat data provides a model to demonstrate the features of the indices.

e.g. 1

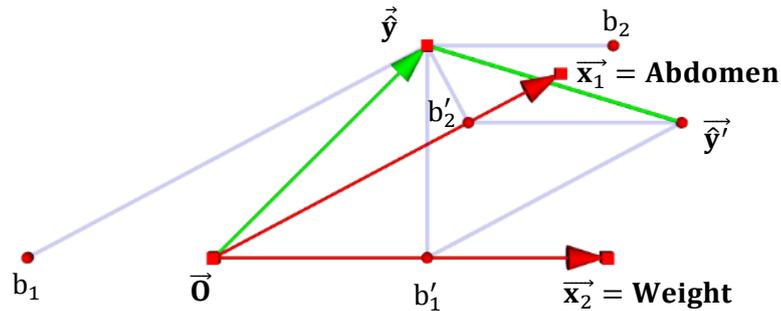


Figure 5.10: Vector Geometry of the Weight + Abdomen model.

The intention of ‘diagnosing’ collinearity is to deconstruct an example such as this into manageable information that aids with our interpretation of the impact. To test the ability of each index to perform the task, a further two further examples are generated on the same collinear axes of predictors (i.e. an equal r_{12} is used in all models).

e.g. 2

e.g. 3

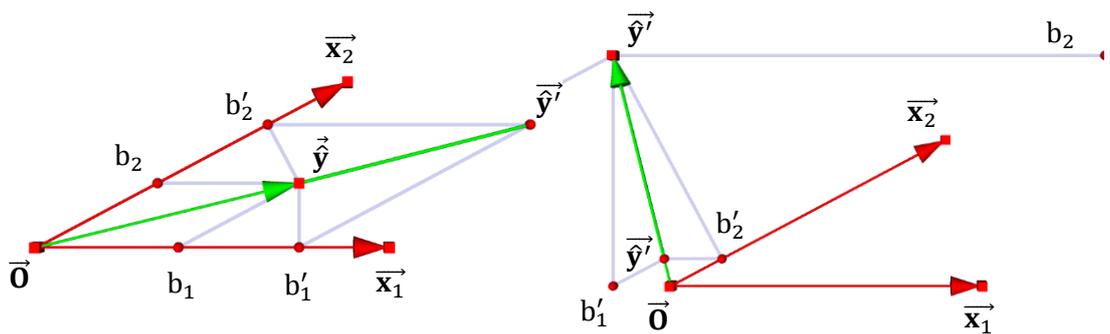


Figure 5.11: A high vs. low association between response and predictors.

The bivariable \hat{y} length is held constant (i.e. an equal R_y^2), whilst the correlation of the response with the predictors is allowed to vary. Importantly, the r_{12} and R_y^2 used in each of these models ensures that our ‘global’ measures will not be able to distinguish between these examples.

5.5.1 Simulation Study

The simulations in this section are intended to highlight limitations of the D-index for use in model building and provide ideas for an extension that could differentiate models as demonstrated in Figure 5.10 and Figure 5.11. The model including predictors ‘weight’ and ‘abdomen’ provides the example (see Figure 5.10). To avoid confusion with the real data, these predictors are replaced by \vec{x}_1 and \vec{x}_2 in the simulations. For both simulations the correlation between \hat{y} and x_1 is varied between $(-1 \dots 1)$. In simulation 5.1 only the correlation between \hat{y} and x_1 is allowed to change. The R_y^2 and r_{12} are held constant at 0.72 and 0.89 respectively. This ensures that the VIF, VIF·VDF and D_2 indices each remain constant. The focus is instead on the change in the point estimates and variance of the predictors to provide further detail of the collinearity impact.

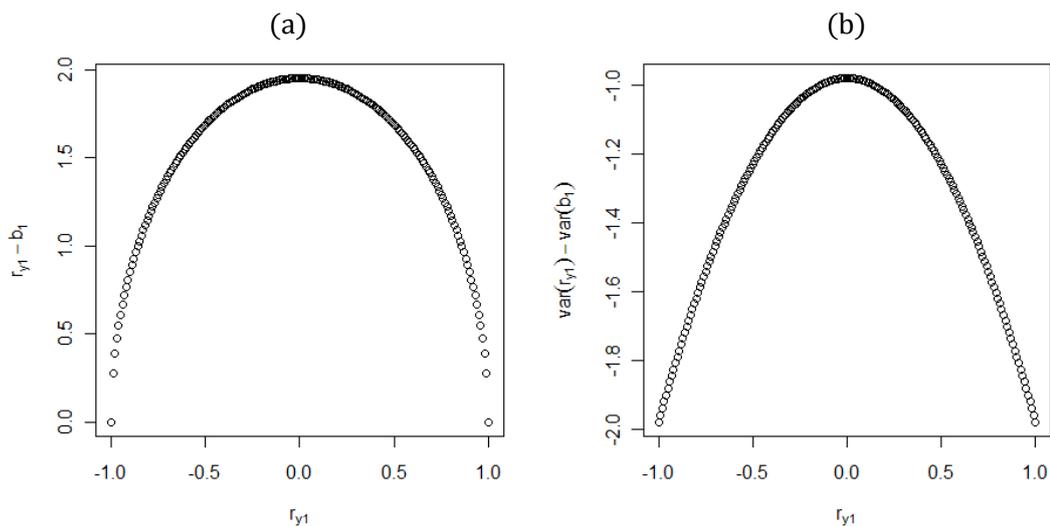


Figure 5.12: Simulation 5.1 - (a) Change in point estimate and (b) variance for b_1 .

The measures in Figure 5.12 display a similar pattern. When \vec{x}_1 and \vec{y} are orthogonal (i.e. $r_{y1} = 0$) the change in the point estimate reaches a maximum and the change in variance on the point estimates is minimized between r_{y1} and b_1 . This could reflect a disparity from ‘truth’ if reflected in the conceptual model and an increased precision of the estimates (reflected by a decreased variance) in the full regression model. However, the global D-index remains constant. The curves are not identical (as expected from point estimates and variances) but they provide an illustration of what the extended D-index should reflect in relation to the VIF·VDF equivalent.

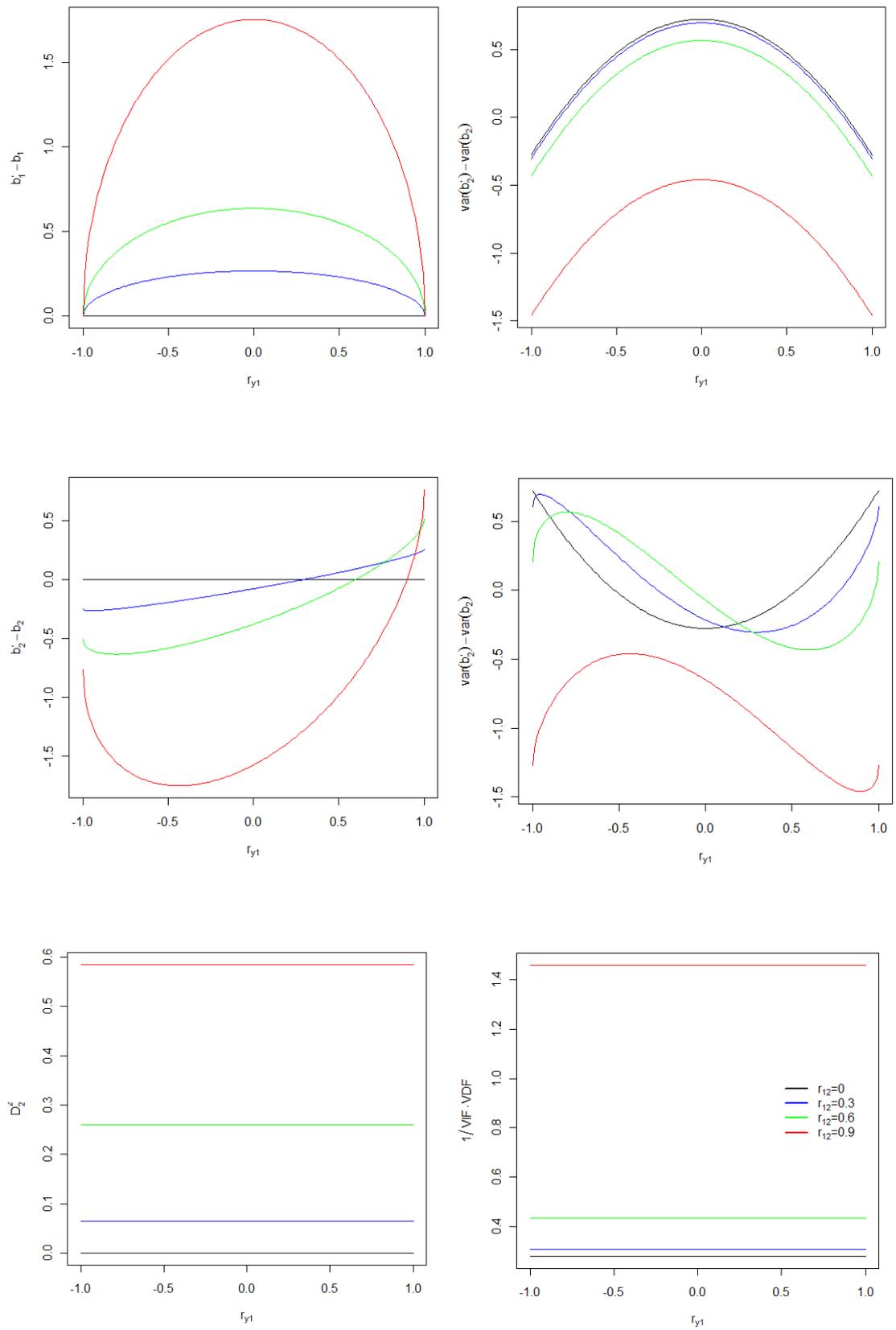


Figure 5.13: Simulation 5.2 – Variance, point estimates and index results.

In simulation 5.2 the value of R_y^2 in the weight + abdomen regression model is retained (0.72), whilst the correlation between the predictors is varied for $r_{12} = (0, 0.3, 0.6, 0.9)$. Clear links are observed between VIF·VDF and D_2^2 in that both represent an overall inflation (when the VIF·VDF is treated as a 'global' measure). The D_2 reflects the inflation in the point estimates, whilst the VIF·VDF is an inflation in variance. If this inflation were removed from the estimates, the curves would only be differentiated by the maxima of the b_1 and b_2 change. The reason for this feature is that the response is uni-dimensional. From a geometrical viewpoint, the 'global' effect of the response is dictated by the association between the univariable response and the regression space. The change in individual coefficients, in both point estimate and variance, is a reflection of the y variance explained individually by x_1 and x_2 , when r_{12} is held constant. This is the feature that will indicate the individual role of the predictors in producing the global 'impact'.

The optimum change on both point estimates shifts towards a similar maxima with increasing correlation (i.e. $r_{y1} = 0$). At this point, the correlation between the response and each predictor is low, resulting in a greater change in the point estimates. In comparison, the VIF of the bivariable model provides the maximum point of each variance change curve as it is comparing to a baseline of orthogonality in the univariable model. The deviation of the curve away from this horizontal line is entirely determined by the change in VDF, thus highlighting that the 'local' (or individual) change in either graph reflects the predictors correlation with the response. An optimum change is observed in each simulation at $r_{y1} = 0$ for x_1 and $r_{y1} = r_{12}$ for x_2 . At this point there exists the greatest difference between R_y^2 and the univariate r_{y1}^2 and r_{y2}^2 respectively.

The feature that is seen in the variance change plot is the suppression effect (see section 3.2.2). The suppression effect becomes apparent when the confidence change on the b_2 point estimate is considered alongside b_1 . For the orthogonal case (i.e. $r_{12} = 0$), the confidence curves are reflected (i.e. the black lines). When the change is maximal on the b_1 point estimate (i.e. \hat{y} is perfectly correlated with x_2), the confidence change is minimal on b_2 (i.e. no change in VIF or VDF). However, as the collinearity is increased, the optimal change on both coefficients shifts to a similar point. This gives the effect that entering a predictor with little or no correlation to the response will produce a bivariable model with R_y^2 that outweighs the combined r_{yj}^2 of the two univariable models. Any measure on the change in point estimates need only account for the correlation change and not R_y^2 . The inflation relating to the variance explained in the response is captured in the global D_2^2 .

5.5.2 Extending the Bivariate Index

Consider the components α_2 and β_2 that were formed in the construction of D_2 (see section 5.4.1). These were built to compute the D-index with α_2 parallel to $\bar{\mathbf{x}}_1$ and β_2 orthogonal to $\bar{\mathbf{x}}_1$. The axes were specified arbitrarily to understand the construction of the D_2 measure. Figure 5.14 demonstrates that the same components can be computed as a projection of the fitted response $\hat{\mathbf{y}}$.

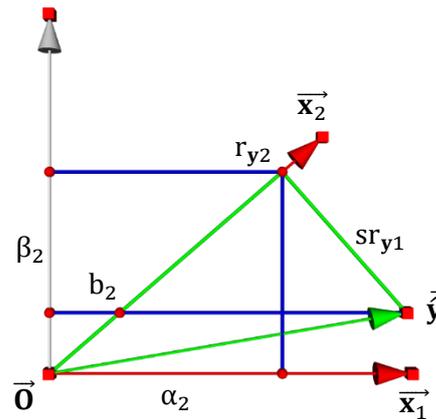


Figure 5.14: An alternative illustration of the computation of α_2 and β_2 .

Consider a simple regression with \mathbf{y} regressed on \mathbf{x}_1 . D_1 is defined to equal zero in any scenario of baseline orthogonality. This should be clear as there exists no collinearity in the model when only one predictor is entered. Next consider the addition of a second predictor \mathbf{x}_2 which generates an impact of collinearity reflected by a non-zero D_2 (unless \mathbf{x}_2 is orthogonal to \mathbf{x}_1 or either predictor explains no variance in the response). A simple regression of the r_{y2} component of $\hat{\mathbf{y}}$ onto \mathbf{x}_1 will generate α_2 (as shown in Figure 5.14). This demonstrates the degree of r_{y2} explained in the baseline model (i.e. containing only \mathbf{x}_1). For example, when \mathbf{x}_1 and \mathbf{x}_2 are perfectly correlated $D_2 = \alpha_2$, highlighting that all of the variance in the response explained by \mathbf{x}_2 is already explained by \mathbf{x}_1 . When they are orthogonal, $\alpha_2 = 0$. The second component β_2 is the semi-partial correlation of \mathbf{x}_1 with \mathbf{y} (whilst holding \mathbf{x}_2 constant), projected onto an axis orthogonal to $\bar{\mathbf{x}}_1$. As the semi-partial correlation sr_{y1} demonstrates in Figure 5.14, \mathbf{x}_2 cannot be held constant under the collinearity present (e.g. when $r_{12} = 0$, $\overline{sr_{y1}}$ would be parallel to $\bar{\mathbf{x}}_1$ and equal to r_{y1}). A portion of r_{y2} that would be unique to \mathbf{x}_2 under orthogonality is now confounded with the explained variance in \mathbf{x}_1 . Use of the axes specified in Figure 5.14 places an emphasis on r_{y2} and identifies the impact of collinearity after the addition of \mathbf{x}_2 to the model.

Recognising the importance of point estimate change alongside the ‘global’ measure, approaches were considered to incorporate this feature into the D-index. There is an important feature that has been largely ignored to this point – that of the angle of deviation. This is particularly evident in the examples displayed at the beginning of section 5.5 that D_2 was unable to distinguish. Consider the angles between the predictor vectors and the D_2 change (which is now considered in vector form - \vec{D}_2). The angles (that in turn provide correlations) perform a similar role to the individual changes demonstrated in the simulations by deconstructing the impact of collinearity. Figure 5.15 illustrates the calculation of these angles.

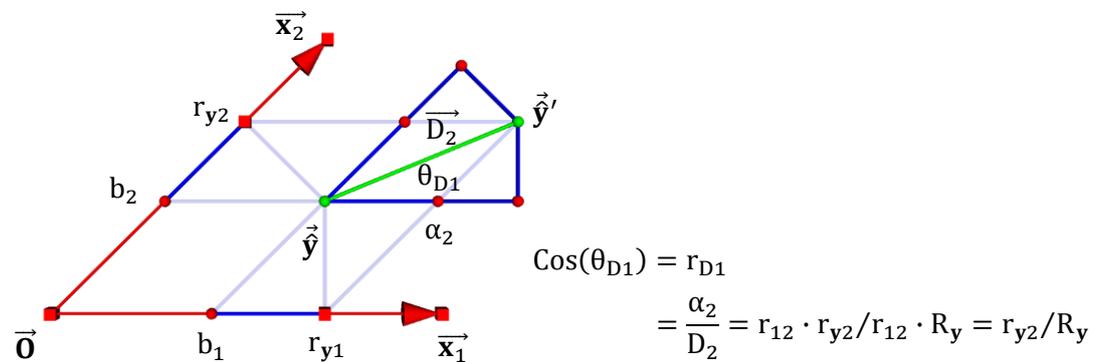


Figure 5.15: Components of the correlation between the \vec{D} and \vec{x}_1 .

The vector geometry in Figure 5.15 demonstrates that each correlation between the D vector and covariate vector (i.e. cosine of θ_{D1}) can be calculated in the bivariable model as the ratio of r_{yj} and R_y . This is the ratio of α_2 (shown in eqn(5.5)) to $|D_2|$. The component α_2 is redefined for the target variable with which we wish to identify its contribution to the collinearity impact. The correlation with \vec{D}_2 is computed by setting arbitrary axes parallel and orthogonal to the target covariate. Therefore, if \vec{x}_1 is added to the simple regression model consisting of the predictor \vec{x}_2 , the arbitrary axis would be formed parallel to \vec{x}_2 and represent the degree to which a proportion of r_{y1} is explained by \vec{x}_2 . Simulation 5.2 suggests that scaling by D would remove the inflation effect of collinearity, thus normalizing the quantity to place the estimate on a scale of 0-1. If the explained variance on each predictor in the univariable models is equal, then the correlations with \vec{D}_2 (in the bivariable case) will be equally split. However, if the ratio is larger on one covariate, then the covariate with the weaker correlation to the response will have a greater association with \vec{D}_2 . This dictates the direction of ‘global’ change to be greater in the direction of the covariate with the weaker correlation to the response.

Simulation 5.2 is now extended to consider the angles between \vec{D} and the predictors \vec{x}_1 and \vec{x}_2 . As always, the graphs in Figure 5.16 should be considered alongside the ‘global’ inflation in the point estimates captured by D_2^2 .

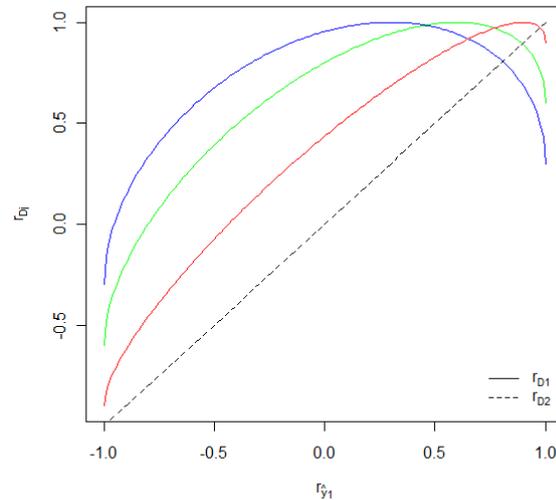


Figure 5.16: Correlations between predictor and \vec{D}_2 for simulations 5.2.

In Figure 5.16 the correlation between \vec{x}_2 and \vec{D}_2 (labelled r_{D2}) follows a linear relationship as r_{y1} is increased (i.e. the dashed line). The reason for this trend is seen from eqn(5.5) (as r_{12} and R_y are fixed and the correlations are adjusted by r_{y1}). The quantity r_{D2} is minimized when $r_{y1} = 0$ and maximized when $r_{y1} = 1$ or -1 . When r_{D2} is minimized, sr_{y2} is maximized (to maintain the constant D_2^2), so the addition of \vec{x}_2 to the model containing \vec{x}_1 is maximally beneficial to the estimation (from a statistical perspective). At the point in which r_{D2} is maximized, $sr_{y2} = 0$ and the addition of \vec{x}_2 adds nothing to the existing model in terms of variance explained. However, as D_2^2 demonstrates in Figure 5.13, adding \vec{x}_2 to the model increases the ‘global’ impact of collinearity on the point estimates.

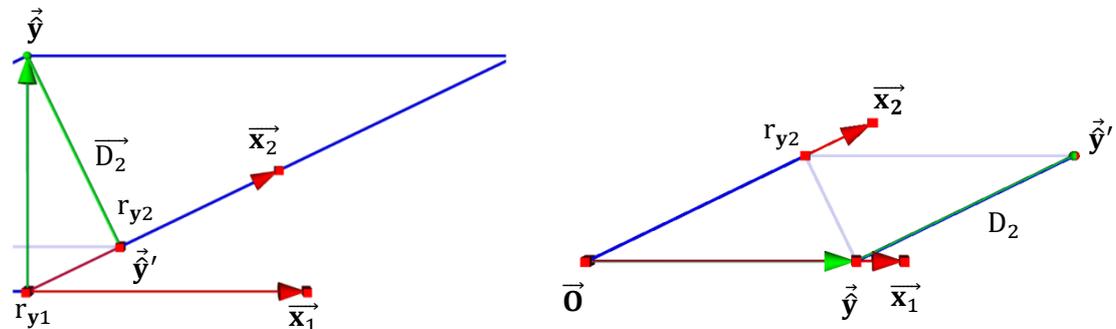


Figure 5.17: r_{D2} (a) minimized and (b) maximized for simulation 5.2.

The curve of r_{D_2} is not identical when the collinearity is varied. This is a result of the simulation procedure which remains on a scale of r_{y_1} . As the correlation between the predictors is increased, the maxima of the curves shift to the right (i.e. closer to $r_{y_1} = 1$). Therefore, when r_{y_1} is high, a high r_{12} will dictate that r_{y_2} is also high. However, the benefit to variance explained in y of adding the second predictor is reduced.

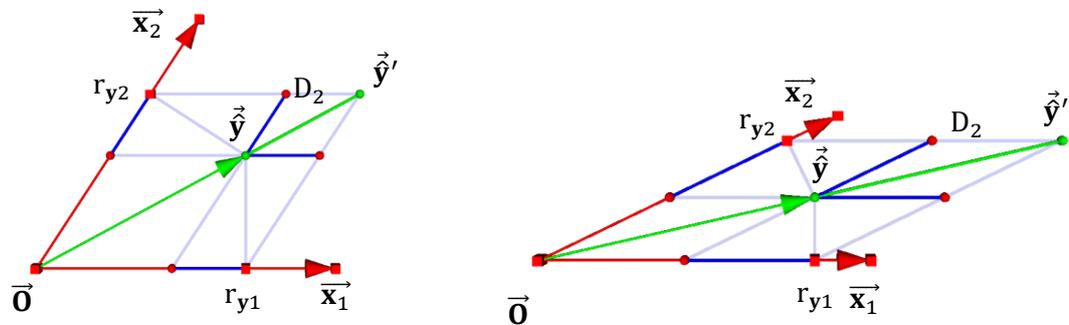


Figure 5.18: Equal correlations between covariate and D under differing r_{12} .

When r_{D_j} are equal, neither of the covariates would be preferable to retain in the model. However, if they are both equal and high, the impact of collinearity will also be high (indicated by an inflated D_2). In comparison, if they are equal and both low then they may both be contributing to explained variance in the full model and it may be valuable to retain both in the model. This would indicate that it is useful to consider R_y^2 in addition to D_2 , r_{D_1} and r_{D_2} . Whilst an increased R_y^2 would allow for a greater impact of the existing collinearity (inflating the D_2), a high R_y^2 coupled with a low D_2 would indicate an optimal model. This result is obvious for the bivariable example, but it may not be in models with a greater number of covariates.

If one of the correlations is substantially different in magnitude to the other then it would indicate a preference to retain one predictor. The predictor with the lower correlation to \vec{D}_2 would explain a greater portion of the variance in the response. It is the magnitude of one correlation in relation to the other that is particularly important. This may indicate that the use of a 'rule of thumb' would be misguided for this index. From an earlier discussion in chapter 3 this is perhaps a useful feature. It should encourage the conceptual understanding of the problem to drive the interpretation of this statistical index. It may also be useful to consider a combination of β_2 in addition to α_2 . The information of β_2 would already be contained in α_2 and D_2 (therefore, it would seem only sensible to consider 2 of the 3 measures), however when extending to higher dimensional examples, this partitioning may seem more natural to understanding the impact of adding a single new predictor to the existing model.

5.6 Measurement Error and Sampling Variation

If we compare a multivariable estimate to an *a priori* expectation, there is no 'problem' in the two sets of estimates. Rather there is an *a priori* assumption that brings an *a priori* expectation that is not satisfied due to sampling variation. The estimates are not 'wrong' nor 'biased' (if classical OLS assumptions are satisfied – see section 2.2.2). The 'problem' occurs (i.e. the reason there is a discrepancy between the sample estimates and the population estimates) due to sampling variation. Thus, what is true is the population estimate (for which there is no collinearity) and the 'problem' estimate is the sample estimate, which is strictly speaking unbiased, but not in agreement with the population estimate in this single sample case (i.e. the precision of the estimation).

For model selection two sets of sample estimates are compared. Rather than accounting only for the variation introduced in taking a single sample, we require a comparison of the uncertainty between model estimates. The D-index alone will provide a direct measure of change in the estimates of the two models. However, variation in the original univariable estimates will change the value of the index. Therefore, the uncertainty in the measure must be accounted for along with the uncertainty added in the new model (a further feature of 'model environment'). It may be that the 'guesstimate' of the population must be made within a confidence region. As a result, some measure of confidence is required for the index. This is a fresh challenge to existing 'correlation based' indices in that they only consider fixed covariates in the $\mathbf{X}^T\mathbf{X}$ matrix.

The challenge is to represent the variance for the two models in the geometry to illustrate and assess this change as part of the D-index. To this point the geometry has been presented in subject space rather than the traditional object space. Although the vectors are in n -dimensional space, it has been possible to present them in as few dimensions as there are covariates, as the remaining $n - k$ dimensions are effectively redundant. To now consider effects such as sampling variation, a greater understanding of the subject space is required. For instance, in each of the diagrams containing two covariates and a response, the geometry has been presented in 3D space, meaning that there are $n - 3$ redundant dimensions. When sampling variation is considered, the actual position in which the response vector lies in n -dimensional space may deviate in future samples (theoretically requiring more than the visually appealing 3 dimensions). A discussion of the error space was presented in section 4.1.2. This becomes essential to the work as the effects of sampling variation are considered on the index.

5.6.1 The Construction of a Confidence Interval

The work in this section considers how the error distribution appears in the vector geometry developed earlier and how the impact on the regression coefficients can be illustrated. The standard error of the estimated coefficient provides an indication of how far the point estimate is likely to deviate from the population coefficient. The standard error (SE) is calculated by taking the root of eqn(2.21) and substituting an unbiased estimate of σ_ε^2 (see O'Brien (2007)),

$$\begin{aligned}
 SE_{b_j} &= \sqrt{\frac{\sigma_\varepsilon^2}{(1 - R_{x_j}^2) \sum x_j^2}} = \sqrt{\frac{(1 - R_y^2) \sum y_j^2 / (n - k - 1)}{(1 - R_{x_j}^2) \sum x_j^2}} \\
 &= \sqrt{\frac{\sigma_y^2 (1 - R_y^2) (n - 1)}{\sigma_x^2 (1 - R_{x_j}^2) (n - 1) (n - k - 1)}} \quad (5.9) \\
 &= \sqrt{\text{VDF} \cdot \text{VIF} \cdot \frac{1}{(n - k - 1)} \left(\frac{\sigma_y^2}{\sigma_{x_j}^2} \right)}
 \end{aligned}$$

The standard error is used to construct confidence intervals for the partial regression coefficient. The 95% confidence interval for β_j is as follows (Wonnacott 1981),

$$\beta_j = b_j \pm t_{n-k-1}(0.025) SE_{b_j}$$

The impact of collinearity through the confidence interval can be demonstrated using vector geometry and standard OLS projections.

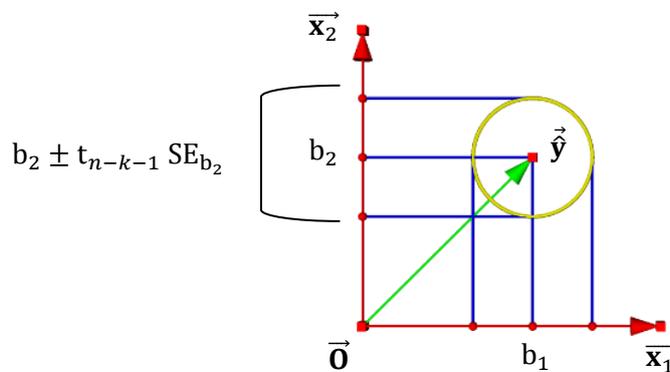


Figure 5.19: Confidence intervals of orthogonal predictors.

A circle is used to represent the bivariate normal distribution of a particular confidence region of $\vec{\hat{y}}$, centered on the point estimate (see section 4.1.1). The outermost points of the circle are projected along the covariates (analogous to obtaining a coefficient point estimate) to construct the individual confidence interval of b_1 and b_2 . This demonstrates the impact on the variance of the coefficient estimate by decreasing the angle between \vec{x}_1 and \vec{x}_2 ; and subsequently the projection of the error circle onto the predictors,

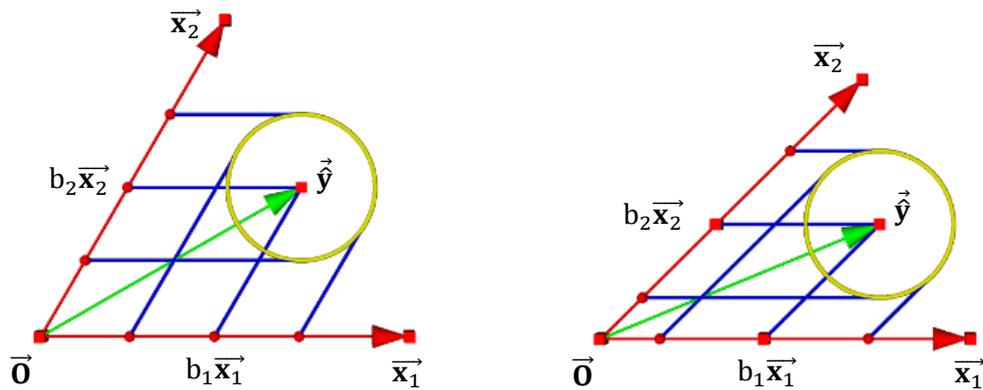


Figure 5.20: Confidence intervals for two sets of highly correlated predictors.

The distribution of the error is independent of the predictors in the model and so the radius of the confidence region is not influenced by a change in correlation. Due to the construction of the geometry, the size, shape and position of the confidence region will not change if the covariates are standardized, as the regions are scaled by the standard deviation of the predictor. By holding the VDF and the length of the covariate vectors constant, the impact of collinearity on the confidence intervals can be illustrated by adjusting the projection of the $\vec{\hat{y}}$ confidence region with respect to θ_{12} (see Figure 5.20).

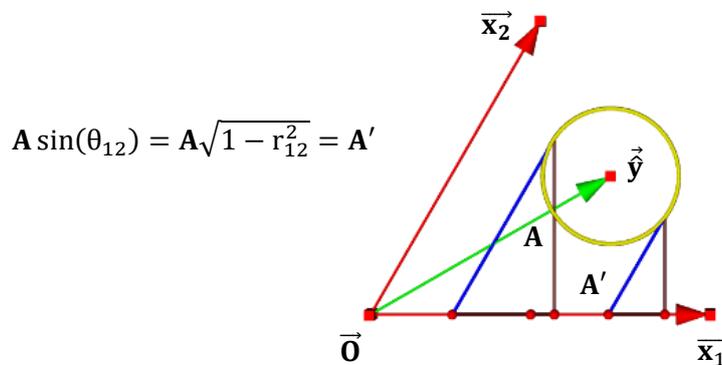


Figure 5.21: Demonstration of the construction of the confidence region.

The radius of the confidence circle can be determined by multiplying the bivariate standard error used in confidence interval (eqn(5.9)) by $\sqrt{1 - \cos^2(\theta_{12})}$ (i.e. $1/\sqrt{\text{VIF}}$). This highlights why the correlation has no effect on the confidence region of \hat{y} in the geometry, as the angle of the projection (determined by the collinearity) will cancel with the VIF in the confidence interval (see Figure 5.21). This nicely demonstrates the independence of error with the covariates. Therefore, the bivariate normal confidence region (i.e. the radius of the circle on the geometrical axes) is determined from eqn(5.10),

$$t_{0.025} \sqrt{\text{VDF} \times \frac{1}{(n - k - 1)} \times \left(\frac{\sigma_y^2}{\sigma_{x_j}^2} \right)} \quad (5.10)$$

This geometrical idea is extended to the three predictor example in Figure 5.22.

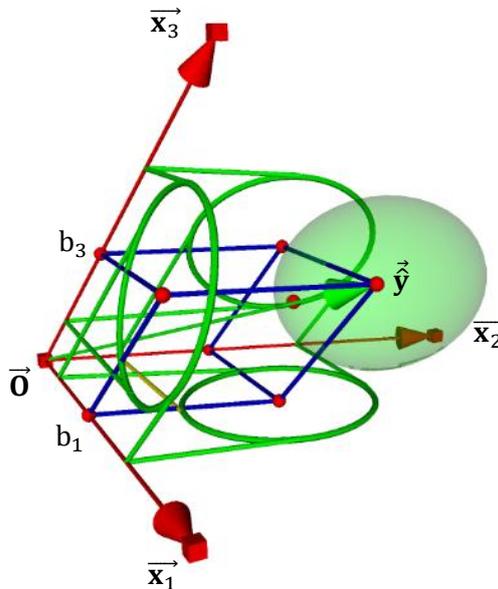


Figure 5.22: An illustration of the construction of the ellipsoid confidence region.

Confidence circles have been produced on each 2D regression plane formed by the pairs of predictors (i.e. the green circles). However, for the confidence region in 3 dimensions, the projection of each of these circles onto the 3D regression space spanned by \vec{x}_1 , \vec{x}_2 and \vec{x}_3 would produce an ellipsoid centered on $\vec{\hat{y}}$. This is a spherical error projected onto the regression space of the correlated axes. The representation of error in the geometrical example provides an illustration of uncertainty. The next step is to decide how to incorporate this impact as part of the D-index.

5.6.2 Confidence Interval of the D-Index

A focus was first placed on the probability of the estimate to produce a ‘symptom’ of collinear data relevant to epidemiology. This was initially based on the likelihood of a point estimate changing sign. The vector geometry was split into integration regions showing the probability of a change of sign under sampling variation (see Figure 5.23a).

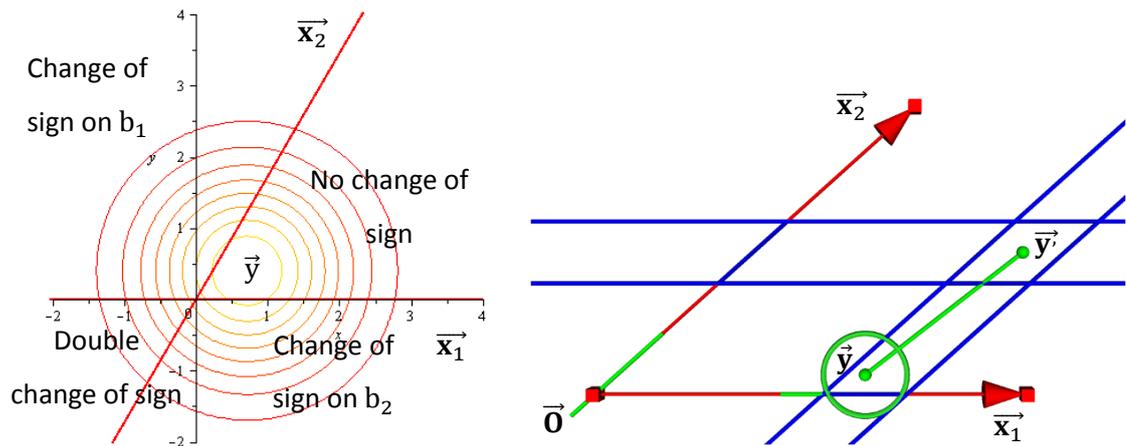


Figure 5.23: (a) Change of sign regions and (b) uncertainty region for a ‘guesstimate’.

The integration ideas were an attempt to incorporate effects of sample size, sampling variation and measurement error directly into the metric of the index. The problem was that the integration gave a greater percentage (indicating less movement) if the VDF and sample size were low (i.e. the result of a greater confidence region) and inflated the percentage with a smaller D-index (i.e. a positive result). Therefore, this method appeared to have conflicting results from a model building perspective when indicating the impact of collinearity.

The integration ideas would appear feasible for other purposes. For instance, a confidence region for a ‘guesstimate’ (see Figure 5.23b). A high percentage would suggest that a greater sample size, or correlation with the response is necessary for the sample estimates to be closer to the population parameter. However, for model building purposes a holistic index such as this confused the effects. A greater variance explained generates a greater movement in the point estimates. This would seem to suggest a greater problem, which works against our conceptual understanding of the beneficial role of the response from a variance perspective. The work returned to the D-index as a standalone measure and to provide a confidence interval to highlight uncertainty as an external feature.

The confidence region of \vec{y} can be mapped onto \vec{y}' using the components as illustrated in section 5.5.2. This relationship is scaled by the correlation between the predictors (r_{12}), which is a fixed effect as the predictors are assumed to be measured without error – see section 2.2.2.

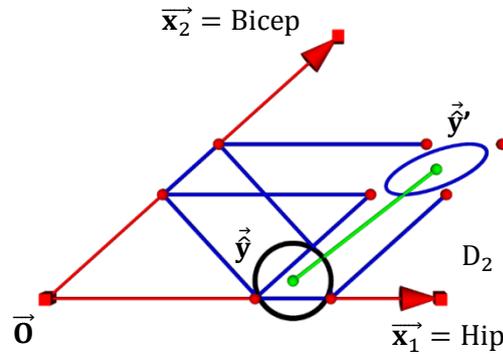


Figure 5.24: The translation of the confidence region from the bivariable models.

A confidence interval for R_y^2 is presented by Cohen (2003). The standard error is given by the following estimation,

$$SE_{R_y^2} = \sqrt{\frac{4R_y^2(1 - R_y^2)^2(n - k - 1)^2}{(n^2 - 1)(n + 3)}} \quad (5.11)$$

The confidence interval is calculated using the t distribution for a sample size >60.

$$R_y^2 \pm t(SE_{R_y^2}) \quad (5.12)$$

As the predictors are fixed, the r_{12}^2 is a scalar quantity. This is the reason for correlation based indices not requiring a confidence interval. Therefore, using eqn(5.12) the interval for D_2^2 is as follows,

$$r_{12}^2 \cdot R_y^2 \pm t(SE_{R_y^2}) \quad (5.13)$$

With the addition of a confidence interval, the index for the bivariable D_2^2 is now complete. The D-index and the angles could be developed in a different form in the future and so for now the confidence interval appears most useful as a separate element rather than a holistic measure (as demonstrated in the integration idea). The next stage in the development is to extend the concept to higher dimensional examples.

5.7 Diagnosis in Higher Dimensions

For the index to be of use in application it must be extended to models with $k > 2$ predictors. The work now assumes the impact in the bivariable model to be calculated by D_2 and looks to build on the metric developed earlier.

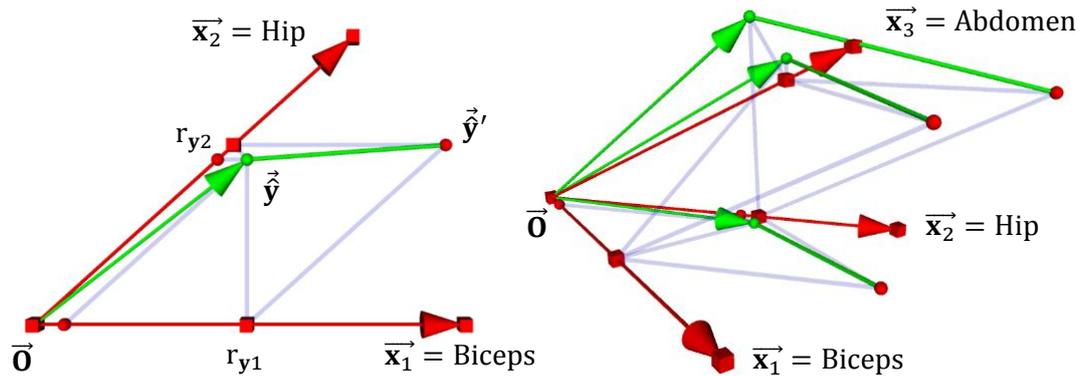


Figure 5.25: D_2 from the model including biceps, hip and abdomen.

D_2 can be calculated on a 2D plane spanned by any two predictors within the 3D regression space. The development of D_3 can be illustrated by building on the regression model containing \bar{x}_1 and \bar{x}_2 (biceps and hip respectively), with the 2D space for D_2 spanned by these covariate vectors (see Figure 5.25a). A third predictor \bar{x}_3 (abdomen) is then added to the model. The D_2 can be calculated for each of the three pairs of predictors in the new model (see Figure 5.25b and Table 5.3).

Model	R_y^2	r_{12}	D_2^2
x_1, x_2	0.39	0.74	0.21
x_1, x_3	0.67	0.68	0.31
x_2, x_3	0.69	0.87	0.53

Table 5.3: D_2 from the model including biceps, hip and abdomen.

Table 5.3 identifies the greatest D_2 with hip and abdomen entered as predictors. The second highest D_2 is seen in the model with biceps and abdomen. Whilst the correlation is weakest between these predictors, the higher variance explained in comparison to the model with biceps and abdomen has generated a greater inflation in the point estimates.

Two options are considered for extending D_2 to D_3 . The first is to calculate the additional impact on the existing bivariable model (including \mathbf{x}_1 and \mathbf{x}_2) of adding a third predictor \mathbf{x}_3 (labelled D_3), and second the impact of collinearity on a baseline model that assumes orthogonality amongst *all* of the predictors (labelled D_3). The difference between these approaches is the model assumed at baseline, however the metric remains the same. Both can be seen as a direct extension to D_2 in that the bivariable example can be viewed as an additional impact of adding \mathbf{x}_2 and also an impact to baseline orthogonality (as the univariable model has zero collinearity for a single predictor).

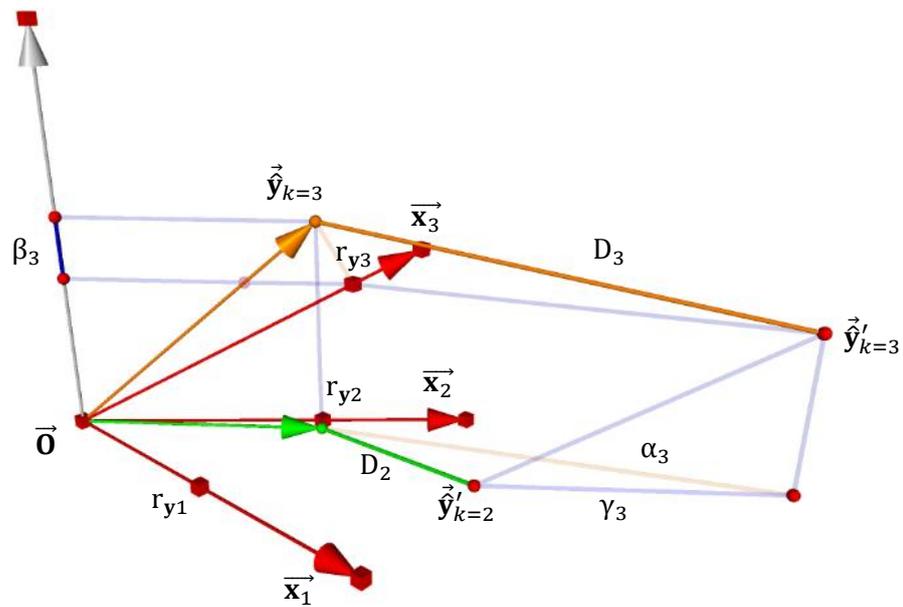


Figure 5.26: Illustration of the extension to D_2 with \mathbf{x}_3 included.

Figure 5.26 illustrates the vector geometry for the three predictor regression model. The fitted response $\hat{\mathbf{y}}_{k=3}$ is first projected orthogonally onto the covariate vectors $\bar{\mathbf{x}}_j$ to obtain the individual r_{yj} (i.e. regression coefficients from each univariable model). The r_{yj} are then projected along the plane formed by the remaining two predictors to identify $\hat{\mathbf{y}}'_{k=3}$ representing the baseline model of orthogonality amongst the covariates. The distance between $\hat{\mathbf{y}}_{k=3}$ and $\hat{\mathbf{y}}'_{k=3}$ forms the new D_3 (analogous to D_2 in computation). The fitted response in the three predictor model $\hat{\mathbf{y}}_{k=3}$ is an extension of $\hat{\mathbf{y}}_{k=2}$ in the direction orthogonal to the plane spanned by $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$. The orthogonality with the plane demonstrates that this extension represents a partial correlation between \mathbf{y} and \mathbf{x}_3 , holding \mathbf{x}_1 and \mathbf{x}_2 constant – labelled $\text{pr}_{y3.12}$. The $\hat{\mathbf{y}}'_{k=3}$ is an extension of $\hat{\mathbf{y}}'_{k=2}$ in the direction of $\bar{\mathbf{x}}_3$ with length (i.e. variance) equal to r_{y3} .

First consider the computation of an additional impact of adding \mathbf{x}_3 to the already assumed D_2 impact from the bivariable case. The vector $\overrightarrow{D_3}$ is projected onto the 2-dimensional plane spanned by the vectors $\overrightarrow{\mathbf{x}_1}$ and $\overrightarrow{\mathbf{x}_2}$. The projected $\overrightarrow{D_3}$ is demonstrating the overlap between \mathbf{x}_3 with variance r_{y_3} and the existing predictors in the model (i.e. $\overrightarrow{\mathbf{x}_1}$ and $\overrightarrow{\mathbf{x}_2}$) – analogous to α_2 in the bivariable index. This projection is labelled α_3 , which is composed of D_2 and a further component γ_3 (as illustrated in Figure 5.26). Following the previous construction of D_2 , γ_3 is computed as two components. The first is parallel to $\overrightarrow{\mathbf{x}_1}$ found by an orthogonal projection of \mathbf{x}_3 with variance r_{y_3} onto $\overrightarrow{\mathbf{x}_1}$,

$$\dot{\gamma}_3 = r_{y_3} \cdot r_{13} \quad (5.14)$$

This represents the overlap of \mathbf{x}_3 (of length r_{y_3}) and \mathbf{x}_1 . The second (labelled $\ddot{\gamma}_3$) is in the direction orthogonal to $\overrightarrow{\mathbf{x}_1}$ in the plane spanned by $\overrightarrow{\mathbf{x}_1}$ and $\overrightarrow{\mathbf{x}_2}$. This demonstrates that $\overrightarrow{\mathbf{x}_1}$ is held constant, thus defining a semi-partial correlation between $\overrightarrow{\mathbf{x}_3}$ and $\overrightarrow{\mathbf{x}_2}$, holding \mathbf{x}_1 constant (labelled sr_{23}).

$$\ddot{\gamma}_3 = r_{y_3} \cdot sr_{23} \quad (5.15)$$

Therefore, γ_3 is calculated as the squared sum of orthogonal components,

$$\gamma_3 = \sqrt{\dot{\gamma}_3^2 + \ddot{\gamma}_3^2} = \sqrt{(r_{y_3} r_{13})^2 + (r_{y_3} sr_{23})^2} \quad (5.16)$$

Eqn(5.16) is an extension to D_2 in the plane spanned by \mathbf{x}_1 and \mathbf{x}_2 after adding \mathbf{x}_3 to the model. Finally, there is an additional deviation that would represent the new β component. β_3 represents a deviation of the coefficients in the dimension orthogonal to the computation of D_2 . The geometry illustrates that this is a projection of the remaining explained variance of $\hat{\mathbf{y}}$ (i.e. the component of \mathbf{y} orthogonal to $\overrightarrow{\mathbf{x}_3}$) onto an arbitrary axis orthogonal to the plane spanned by $\overrightarrow{\mathbf{x}_1}$ and $\overrightarrow{\mathbf{x}_2}$. There is a residual from \mathbf{x}_3 (of length r_{y_3}) after regressing on \mathbf{x}_1 and \mathbf{x}_2 . This residual is composed of $pr_{y_3,12}$ and β_3 (analogous to the proof for D_2 with the residual of $sr_{y_2,1}$ and β_2). Therefore, β_3 is an impact of collinearity representing the explained variance of the original model confounded with \mathbf{x}_3 .

$$\beta_3 = R_{3,12} \sqrt{sr_{y_1,3}^2 + pr_{y_2,13}^2} \quad (5.17)$$

The index \dot{D}_3 can be calculated as the squared sum of the components γ_3 and β_3 ,

$$\dot{D}_3^2 = \gamma_3^2 + \beta_3^2 = [r_{y_3}^2 (r_{13}^2 + sr_{23_1}^2)] + [R_{3_{12}} \sqrt{sr_{y_{13}}^2 + pr_{y_{213}}^2}]^2 \quad (5.18)$$

Returning to the vector geometry, the computation of \dot{D}_3 can be summarized as follows. The response \hat{y} has been split into two components (r_{y_3} and $sr_{y_2} + pr_{y_{123}}$). The point r_{y_3} is projected onto the surface spanned by \vec{x}_1 and \vec{x}_2 (which would have zero correlation if it were uncorrelated with the baseline model) and $(sr_{y_2} + pr_{y_{123}})$ projected onto \vec{x}_3 (which would similarly be uncorrelated at baseline). However, if a correlation is present it will generate a deviation of the point estimates represented by a non-zero \dot{D}_3 . The advantage of using this measure (i.e. the bivariable model as baseline) is that the interpretation is much the same as the previous example for D_2 . There are again two components of the response to project, only in this example one component represents a baseline model with the explained variance of two predictors rather than one.

The second index D_3 is an impact of collinearity in moving from uncorrelated covariates at baseline to the three predictor model. In other words, if \mathbf{x}_3 had zero correlation with both \mathbf{x}_1 and \mathbf{x}_2 , \dot{D}_3 would always be zero. However, \mathbf{x}_1 and \mathbf{x}_2 could still be correlated and so an impact on the point estimates from baseline orthogonality would still be observed, but it would be represented solely in the D_2 . The index is now computed as an overall impact of collinearity to give D_2 and D_3 a common baseline. In this computation an emphasis is placed on \mathbf{x}_3 by considering the first component of \hat{y} to be r_{y_3} (followed by $sr_{y_{13}}$ and $pr_{y_{213}}$), however any construction of \hat{y} would produce the same 'global' result. This is only important for when the contributions of the individual predictors to the overall impact are considered using angles of deviation.

In the D_3 measure, α_3 is once again split into two components. The first deviation is the component parallel to \vec{x}_1 , which is the summation of α_2 and $\dot{\gamma}_3$. This represents the portion of explained variance from \mathbf{x}_2 and \mathbf{x}_3 overlapped with \mathbf{x}_1 . There is a second component of this impact in the plane spanned by \vec{x}_1 and \vec{x}_2 . This consists of the shared variance of r_{y_3} with \mathbf{x}_2 , whilst holding \mathbf{x}_1 constant. This is represented by the addition of $\dot{\gamma}_3$ and β_2 . The final component of D_3 is the deviation orthogonal to the plane spanned by \vec{x}_1 and \vec{x}_2 - this is the component β_3 - identical to that already computed for \dot{D}_3 (see eqn(5.17)). The D_3^2 demonstrating the overall impact from baseline of orthogonality can now be expressed as follows,

$$D_3^2 = [r_{y2}r_{12} + r_{y3}r_{13}]^2 + [r_{y3}sr_{23_1}]^2 + \left[r_{12} sr_{y1_2} + R_{3_{12}} \sqrt{sr_{y1_3}^2 + pr_{y2_{13}}^2} \right]^2 \quad (5.19)$$

In both measures, the components of the response are projected onto vectors which would have zero correlation at baseline. This deviation contributes to the global measure. The indices generated for the trivariable model can now be extended to the general case with k predictors as follows.

$$D_k^2 = \left[r_{yk}^2 \left(r_{k1}^2 + pr_{k2_1}^2 + \dots + pr_{k(k-1)_{1,\dots,(k-2)}}^2 \right) \right] + \left[R_{k_{1,2,\dots,(k-1)}} \sqrt{pr_{y1_k}^2 + pr_{y2_k}^2 + \dots + pr_{y(k-1)_k}^2} \right]^2 \quad (5.20)$$

$$D_k^2 = [r_{y2}r_{12} + r_{y3}r_{13} + \dots + r_{yk}r_{1k}]^2 + [r_{y3}sr_{23_1} + r_{y4}sr_{24_1} + \dots + r_{yk}sr_{2k_1}]^2 + \dots + \left[r_{yk} pr_{(k-1)k_{1,\dots,(k-2)}} \right]^2 + \left[R_{k_{1,2,\dots,(k-1)}} \sqrt{pr_{y1_k}^2 + pr_{y2_k}^2 + \dots + pr_{y(k-1)_k}^2} \right]^2 \quad (5.21)$$

It may be assumed that the D-index in higher dimensions represents a generalized form of $R_x \cdot R_y$. The first index D_k^2 measures the impact of adding a single predictor to a baseline model assumed as the model including $k - 1$ predictors. The second D_k^2 assumes the predictors to be uncorrelated at baseline and incorporates the previous D_{k-1}^2 impact as part of an overall impact.

5.7.1 Extending the Multivariate Index

The index D_k is restricted for model building purposes in that it is unable to identify the most influential predictors to generating the impact of collinearity (as was found for D_2). Once again the correlations can be considered between the covariate vectors and the deviation index \vec{D}_k (i.e. the direction of deviation). The calculation of any D-statistic is not dependent on the order in which the predictors are entered, however the order is important to compute the correlations. Figure 5.27 demonstrates that the angle between \vec{D}_3 and the predictor \vec{x}_j can be calculated by an orthogonal projection of \vec{D}_3 onto any 2D plane containing \vec{x}_j . The change in the direction parallel to \vec{x}_1 is calculated as $\alpha_2^2 + \gamma_3^2$.

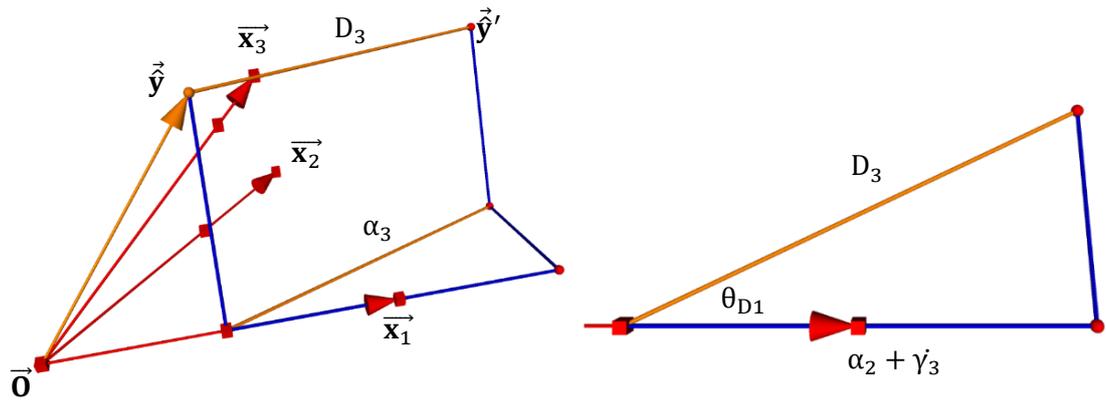


Figure 5.27: Orthogonal projection of \vec{D}_3 onto \vec{x}_1 .

The correlation is calculated as the ratio of $\alpha_2^2 + \gamma_3^2$ and D_k^2 for the target covariate. For example, the correlation of \vec{D}_3 with \vec{x}_1 (i.e. the target covariate) is as follows,

$$\cos(\theta_{D1}) = r_{D1} = (r_{y2}r_{12} + r_{y3}r_{13})/|\vec{D}_3| \quad (5.22)$$

For the general case, a similar orthogonal projection of D_k can be made onto any predictor. The correlation between \vec{D}_k and \vec{x}_1 would be as follows,

$$r_{D1} = (\alpha_2^2 + \gamma_3^2)/|\vec{D}_k| = [r_{y2}r_{12} + r_{y3}r_{13} + \dots + r_{yk}r_{1k}]/|\vec{D}_k| \quad (5.23)$$

The correlation between \vec{D}_k and any predictor can be found by placing the first arbitrary axis of the D_k calculation parallel to the target covariate. For D_3 , if the overlap of \mathbf{x}_2 and \mathbf{x}_3 with \mathbf{x}_1 is high in comparison to the overlap with either of the other predictors the deviation will be towards the \mathbf{x}_1 vector (i.e. will receive a greater weight in the direction vector). This will produce a greater r_{D1} and be flagged as an important contributor to the overall impact of collinearity on the point estimates. At this point the new index is complete for the general case.

5.8 Regression Study: Body Fat Analysis

The aim of the original study was to discover which subset of body circumference measurements could be used to represent body fat. In this example the focus is on four predictors in particular – \mathbf{x}_1 =wrist, \mathbf{x}_2 =neck, \mathbf{x}_3 =abdomen and \mathbf{x}_4 =biceps. A global measure of point estimate deviation is made using the index D_k (eqn(5.21)) and correlations between \mathbf{x}_k and D_k (eqn(5.23)) are included to highlight the role of the individual predictors in producing this impact (i.e. r_{Dj}).

Model	R_y^2	D_k^2	r_{D1}	r_{D2}	r_{D3}	r_{D4}
x_1, x_2	0.24	0.13 (0.08 to 0.18)	0.99	0.71	-	-
x_1, x_3	0.70	0.27 (0.25 to 0.29)	0.97	-	0.41	-
x_1, x_4	0.25	0.10 (0.06 to 0.13)	0.99	-	-	0.70
x_2, x_3	0.70	0.40 (0.36 to 0.43)	-	0.97	0.59	-
x_2, x_4	0.28	0.15 (0.10 to 0.20)	-	0.93	-	0.93
x_3, x_4	0.67	0.31 (0.28 to 0.34)	-	-	0.60	0.99
x_1, x_2, x_3	0.71	0.89 (0.65 to 1.16)	0.92	0.92	0.62	-
x_1, x_2, x_4	0.28	0.50 (0.32 to 0.74)	0.95	0.87	-	0.81
x_1, x_3, x_4	0.70	0.79 (0.54 to 1.05)	0.92	-	0.62	0.87
x_2, x_3, x_4	0.70	1.06 (0.80 to 1.32)	-	0.95	0.69	0.89
x_1, x_2, x_3, x_4	0.71	1.77 (1.28 to 2.33)	0.89	0.93	0.70	0.85

Table 5.4: Results from the D-index for the four predictor body fat example.

Considering first the collection of D_k^2 produced by each model, the greatest impact of collinearity is highlighted for the model involving x_2 and x_3 . This follows with the maximal correlation and subsequently the VIF ($r_{23} = 0.75$, $VIF = 2.29$). The r_{D2} demonstrates that the covariate with the greater correlation to the response is x_3 ($r_{y3} = 0.81$), highlighting that x_2 provides the greater contribution to the impact of collinearity on the point estimate. Studying the correlations between covariates indicates that the model involving x_2 and x_4 has a similarly high correlation ($r_{24} = 0.73$). For this example the variance in y explained by both predictors is low ($r_{y2} = 0.49$, $r_{y4} = 0.49$) and so the impact of collinearity on the model has been limited by the low model R_y^2 . However, both r_{Dj} are large indicating that the collinearity is high in this example. Therefore, the individual predictors can perform an important role in indicating a potential ‘problem’ even when the overall inflation is low.

In each bivariable model the covariate \bar{x}_1 had the strongest correlation with \bar{D}_2 . This is demonstrated by the low correlation with y (i.e. $r_{y1} = 0.35$). The covariate x_3 consistently had the lowest correlation with D_k (for any model) suggesting that it would be a useful predictor to include in the model due to its high explanatory power. A confidence interval for the bivariable model was formed using the eqn(5.13) discussed in section 5.6.2. The three and four predictor models were bootstrapped using a “leave one out” method

(Tukey 1958). The R_y^2 does not increase greatly beyond the two predictor model that included x_1 and x_3 (0.7). The correlations r_{D_j} indicate that x_1 was the main contributor to this impact of collinearity in the bivariable model. A very moderate increase in R_y^2 is found after including x_2 into this model (0.71). However, with this inclusion D^2 has increased from 0.27 to 0.89. The additional impact of adding the predictor x_2 to this model can be calculated as $\dot{D}_3^2 = 0.49$. This would appear high when viewed alongside other bivariable measures to attain a small increase in R_y^2 .

When x_3 is entered into the model along with x_1 and x_2 the r_{D_j} are equal for both x_1 and x_2 . In comparison, when x_4 is added to the model along with x_1 and x_2 , r_{D_1} is greater than r_{D_2} . This demonstrates how the role of each predictor changes dependent on other predictors entered into the model. In the full four predictor model the R_y^2 reaches 0.71, however the deviation peaks at 1.77. The collinearity structure of the four predictors and the variance explained suggests that x_2 is the greatest contributor to this impact of collinearity. This is a change to x_1 being consistently high in previous models. Observing model parsimony would seem to discount the four predictor model. Removing x_2 produces a model with a high R_y^2 (0.7) and moderately low D_3^2 (0.79). Excluding x_2 from the model wouldn't seem obvious from only observing the three predictor models (due to the consistently high r_{D_1}), however noticing the impact in the full model has highlighted the dependence of the predictor with others in the study.

From a model building perspective the bivariable model with x_1 and x_3 included as predictors would appear optimal. This model explained a high variance of the response and had a relatively low D^2 . Whilst moderately increasing R_y^2 , adding the predictor x_2 would generate a high impact on the existing model. Also, when considering the full set of predictors x_2 would seem to have the greatest impact. Therefore, if any predictor would be added to the bivariable model, x_4 would appear to be the best option. However, adding x_4 does not increase the explanatory power of the model and so this would not seem sensible. This example has been much simplified as the nature of any causal relationships amongst the covariates has not been considered. This would raise the complexity of the problem and our understanding of incorporating collinearity in the model. Instead this example has focussed on the statistical aspect of the D-index in application. However, it is possible to imagine how this tool could be adapted to a range of applications such as those discussed in section 5.2. Possible adjustments are considered in section 5.9 that would adapt the framework to such applications.

5.9 Discussion

When interpreting the D-index from a model building perspective it may appear conceptually appealing to assume a greater R_y^2 to be beneficial to the estimation. However, it is important to stress that the index is not necessarily measuring a 'problem' (as a 'correlation based' measure may be assumed to be doing). A high R_y^2 will inflate the point estimates and this is subsequently reflected in the index. If the user were comparing to a baseline *prior*, whether that be a zero correlation or some 'guesstimate' of a population correlation, then they would wish to know the deviation of the estimates away from the expectation. This is not representing a biased or 'wrong' estimate, but a population *prior* that is not reflected in the single sample case. Therefore, if a greater R_y^2 inflates the change in coefficients, then we would wish to know the degree of inflation resulting from the R_y^2 and the collinearity in the sample data.

When considering the analysis of the body fat data in section 5.8, the greatest change in impact from D_2 to D_3 is after adding \mathbf{x}_3 to the model including \mathbf{x}_2 and \mathbf{x}_4 . However, \mathbf{x}_3 contributes the greatest explained variance individually ($r_{y3} = 0.81$) and so including this predictor would appear to be positive in the model building process. It is labelled the most beneficial of the predictors by the correlations with \vec{D} , suggesting \mathbf{x}_1 and \mathbf{x}_2 contribute greatly to the impact of collinearity in this model. The high change in global impact could be misleading if it were interpreted as a measure of some collinearity 'problem'. This is why it would seem beneficial that any change in D between models be balanced with a change in R_y^2 . If little explained variance is gained by including an additional predictor in the model, but the point estimate inflation is high, then this should perhaps be viewed as a potential warning (based on the conceptual model employed) of the impact of collinearity on the parameter estimates.

The index could be developed in the future to produce a more natural interpretation for this purpose. Replacing explained variance with some reciprocal estimate may produce the desired deflation effect. This should certainly be considered in the future. The formation of a 'global' index is a novel aspect of this work and one that can play an important role in understanding the impact of collinearity in a range of applications. However, collinearity remains a complex feature in epidemiology and the development of a blanket statistical approach still requires a very careful conceptual understanding to be of benefit in application.

5.10 Conclusions

The aim of this chapter was to build a general apparatus for measuring the impact of collinearity. The D-index provides this tool and its development has been demonstrated from the basic two predictor model and extended to the general case. In some aspects, the construction of this statistic has provided new challenges in comparison to ‘correlation based’ indices, such as the VIF and CI, as the D-index incorporates information of the response. From general regression theory (shown in 2.2.2), the covariates are assumed to be measured without error. Subsequently, correlation based diagnostics are not given a confidence interval. However, considering the response incorporates error into the measure. A confidence interval was created for the bivariate example, however extension to the general case is left as a future development as our primary interest has been invested in demonstrating the utility of the index itself.

A useful feature of this measure is in its links to PLS. The D-index measures between a single and full component PLS (in any dimension). This allows for a useful interpretation in comparing between a baseline (either an uncorrelated *prior* or univariate estimates) and target model. This may be interpreted as the impact of collinearity from a single dimension to the correlated covariates. Interesting links were also highlighted to the VIF · VDF measure proposed in chapter 3. For the bivariate D_2^2 index, these links are clear in that the measures both utilize information from r_{12}^2 and R_y^2 . Similarities can also be seen in the extension to higher dimensions. However, the D-index assesses a ‘global’ impact, whilst the VIF · VDF will only provide indices on the individual covariates. Therefore, the collinearity is partitioned in ways to represent an $R_x \cdot R_y$ in the general case. A measure on the individual contribution of predictors to the global impact (such as that provided by the VIF · VDF) is achieved by correlations between \vec{D} and each covariate vector.

An achievement in the development of the D-index is in the use of vector geometry to create the measure and also to understand it. One of the reasons for proposing the geometric alternative to the VIF · VDF is that it allows a flexibility to incorporate different *a priori* knowledge as to the original aims of the index. This may be extended to assessing an optimal number of PLS components to retain (i.e. finding a balance between signal and incorporating collinearity). Another development would be to develop machinery to process the information from the index to create a dependency structure. The matroid method developed in chapter 6 could provide an ideal vehicle to achieve this aim to produce a structure that incorporates the y information.

6. Approaches to Unravel the Latent Structure within Metabolic Syndrome

In previous chapters a focus has been placed on understanding the relationship between a response variable and a set of correlated predictors. In chapter 6 the focus is on the covariates to identify common mechanisms that influence them. The aim is to analyze the dependence structure amongst manifest variables, to discover the existence and structure of unobserved latent constructs. Techniques such as EFA and PCA are frequently used for this purpose (see section 4.2). Whilst a single dataset can generate a range of heuristic models from either technique, a considered methodological approach based on biological and statistical theory is required to identify factors that reflect 'real life' constructs. The balance between clinical and statistical evidence must be carefully understood to generate consistent and considered reasoning behind their application.

The widespread use and software availability of EFA methods has contributed to it becoming a traditional and accepted means of applied research across a number of diverse fields. However, the results obtained are often highly subjective. The methodological decisions made during their application are rarely given appropriate consideration in their context and this often leads to hypotheses and conclusions based on questionable evidence. Research into the existence and structure of metabolic syndrome is one such application. The use of exploratory and confirmatory factor analysis to analyze the structure of risk factors appears almost compulsory in this field, however the rationale behind the application leaves the analysis open to misuse and misinterpretation. These issues will be discussed along with potential alternative methodology with roots in cluster analysis. In contrast to EFA, variable clustering takes a very different approach to achieve similar structural goals. The aim of this study is to encourage a consistently high contextual validity, without the need to increase the complexity in the application of statistical methods. Two advanced techniques of clustering in the VARCLUS and matroid methods are developed and applied to MetS data, with the results compared to the traditional EFA.

6.1 Example Application

Metabolic syndrome (MetS) defines a clustering of risk factors that act as an indicator for chronic diseases such as kidney disease, cardiovascular disease (CVD) and type-2 diabetes (T2-DM) (Gami et al. 2007;McNeill et al. 2006;Wilson et al. 2005), however components of MetS are still controversial. In recent literature, EFA and CFA have been used to generate and test a latent structure amongst MetS components and regression modelling is used to test the relation between chronic diseases and MetS components (Mannucci et al. 2007;Pladevall et al. 2006). Proposed risk factors, such as obesity, hypertension, insulin resistance and lipid metabolism, are typically correlated and clustered, and this poses a challenge for statistical modelling. Collinearity amongst these components can have serious implications in regression analysis if it is not identified and treated with care. Whilst some exploratory analyses can provide an insight into the structure of the data, the results are often difficult to interpret and methodological decisions are rarely justified.

A cross-sectional study by Shen et al. (2003) is considered in this work. Data were collected from 847 men aged between 21 and 81 years in the ‘Normative aging study’. The study based in Boston, Massachusetts included a total of 2,280 predominantly white community-dwelling males (with a mean age of 61 years). The subjects were selected from an original 6,000 applicants who were screened at entry for existing health conditions. Those suffering from known chronic diseases, such as CVD and T2-DM, were excluded from the study. The subjects selected for the application, were those examined between 1987 and 1991 who provided complete data for the following covariates: fasting insulin (Ins), postchallenge insulin (PCIns), fasting glucose (Glu), postchallenge glucose (PCGlu), body mass index (BMI), waist/hip ratio (WHR), high density lipoprotein cholesterol (HDL), triglycerides (Trig), systolic blood pressure (SBP) and diastolic blood pressure (DBP).

	Ins	PCIns	Glu	PCGlu	BMI	WHR	HDL	Trig	SBP
PCIns	0.65								
Glu	0.27	0.26							
PCGlu	0.29	0.55	0.53						
BMI	0.45	0.37	0.21	0.21					
WHR	0.33	0.29	0.11	0.15	0.46				
HDL	-0.22	-0.21	-0.13	-0.1	-0.27	-0.2			
Trig	0.29	0.36	0.16	0.22	0.24	0.26	0.47		
SBP	0.17	0.23	0.10	0.21	0.10	0.12	-0.02	0.15	
DBP	0.17	0.17	-0.01	0.09	0.19	0.11	-0.01	0.12	0.57

Table 6.1: Pearson correlations for the 10 metabolic risk factors.

6.2 Defining Metabolic Syndrome

The concept of MetS was first proposed by Eskil Kylin (1923) as a syndrome involving hypertension, hyperglycaemia and hyperuricemia (Alberti 2005). The structures to define MetS have frequently changed - more recently with an increased attention to grouping risk factors for T2-DM and CVD. Criteria for the diagnosis of MetS have been proposed by a number of leading health bodies. Two of the most commonly accepted are those of the World Health Organization (WHO) and the National Cholesterol Education Adult Treatment Panel III (ATP III) (Darsow et al. 2006). A study by Ford and Giles (2003) compared the prevalence of MetS using these two definitions. In a nationally representative sample of 8,608 Americans they found disagreement on 13.8% of the subjects classified as suffering from MetS. The variance in definitions demonstrates an uncertainty in the underlying mechanisms. If we are unable to identify the core disease that we are trying to characterize, then it is impossible to form a precise definition of such a disease or even to justify its existence. Clustering amongst the risk factors should stimulate research into an understanding of the relationships, but the use of existing definitions should be implemented with caution (Kahn 2007). Further research is required to provide consistency and reproducibility across studies to form a sound evidence base for the existence and structure of a MetS construct.

The study by Ford and Giles highlighted substantial differences amongst subgroups of the population (e.g. 16.5% of African-American men were diagnosed as suffering from MetS using the ATP III criteria, whilst 24.9% were diagnosed using the WHO criteria). Evidence suggests that the form of the hypothesized syndrome may not be consistent across members of populations (Ford and Giles 2003; Shen et al. 2006). Therefore, statistically methodology is required that is flexible to accommodate this change, but remains able to identify consistency when it is present across subgroups. The clinical relevance should be the primary aim for selecting this methodology. If the same dataset were given to ten different researchers to produce an EFA, they could conceivably return ten different hypotheses. This does not demonstrate consistency in either the method or the interpretation. If it is difficult to achieve this on a single dataset, the task is only magnified across a range of studies and study populations. Discussion regarding the misuse of factor analysis in psychological research is quite common (Fabrigar et al. 1999; Ford et al. 1986; Streiner 1994), however many of the same issues are rarely noted in the clinical and epidemiological literature.

6.3 Factor Analysis and Clustering Procedures

The work of Galton and Pearson in developing the correlation coefficient provided the stimulus for both factor analysis and clustering methodologies to develop. The first to arise was a form of factor analysis in psychology and the study of intelligence. The concept of “intelligence” is a subject of great debate in psychology. A number of theories have been proposed in trying to identify the structure and define such a construct. Charles Spearman (1904) developed the “two factor theory” to represent human ability. The “model of intelligence theory” was developed upon the results of an EFA on the scores from psychometric tests (Spearman 1904). Spearman noted that all the test scores measuring mental ability were positively correlated. He postulated that if this is the case, there must be some underlying factor that generates this positive relationship amongst the scores. Using a simple form of factor analysis Spearman developed the two factor theory, that suggested the presence of a ‘g-factor’ representing an individual’s “general intelligence” and a unique factor specific to the abilities required for each individual test.

The evidence proposed by Spearman for the two factor theory was heavily criticized for being overly simplistic by fellow researchers Edward Thorndyke and Louis Thurstone (1934). They believed that the single g-factor was a spurious finding bourn from limitations of the early methodology used by Spearman. Factor analysis was later developed by Cyril Burt (1909) to extract multiple factors. In 1948, Thurstone published a different structure that suggested the presence of seven “primary mental abilities”. These were labelled (1) verbal comprehension, (2) reasoning, (3) perceptual speed, (4) numerical ability, (5) word fluency, (6) associative memory and (7) spatial visualization (Thurstone 1948). A major difference between Spearman’s and Thurstone’s theories is in the inter-correlations of the factors (Goodman 1943). Spearman believed that the factors must be uncorrelated, whereas this was not deemed important as part of Thurstone’s theory. A later study analyzed test data from subjects with a similar IQ which again supported the seven factor structure. However, when this process was repeated on a sample of children with seemingly similar mental ability, Thurstone’s structure was less clear, appearing to instead support Spearman’s g-factor approach. Instead of the two theories contradicting one another, it is possible to consider them as part of a common hierarchical structure. The g-factor would overarch the seven primary mental abilities. This example in the early developments of factor analysis encompasses the hierarchical structures common to many statistical models representing ‘real world’ concepts.

Although Spearman was limited by technology of the time, the results of the factoring approach highlighted the usefulness of the methodology in a much broader range of fields. Robert Tryon (1935) recognised the use of factor analysis from this evidence, but the complex factor structures and having to perform the analysis by hand drove him toward a goal of improving the practicality of the method.

"The price of such mathematical elegance was one's not knowing exactly what one had measured"

(Tryon and Bailey 1970)

Tryon pictured observations in geometry – observing 'natural' formations and clusters. He proposed the idea of cluster analysis, citing that the structure of objects can be generated using 'objective' methodology, without the need to impute complex latent variables underlying those observed.

Paralleled with the development of personal computers in the 1960's the number of automated clustering methods has substantially increased since Tryon's initial work. The range of methodology is now at a stage in which cluster analysis can be considered an independent scientific discipline in itself (Kaufman and Rosseeuw 1990). Cluster analysis was originally conceived as an alternative to factor analysis for theory development. This meant employing one methodology over another would benefit from some advantages, whilst compromising on the useful features of the other. However, the development of the clustering algorithms has seen the lines between the methods fade. For instance, there are techniques such as "partitioning" (e.g. k -means clustering) in which the number of clusters to be extracted is specified by the user prior to the analysis and "fuzzy clustering" in which predictors are allowed to appear in more than one cluster (similar to the factors generated in a factor analysis) – see section 4.3.

The discussion regarding cluster analysis must be fairly vague as it encompasses a number of algorithms and definitions. The sheer range of options provided by clustering methodology often leads to criticism in applied work, with the dissimilarity in the results being cited as too 'subjective'. However, the range of options should be viewed as beneficial to research. It is the responsibility of the user to decide the appropriateness of the methodology to the research question. The discussion will begin with considering the advantages and limitations of the 'traditional' approaches in factor analysis and cluster analysis. It will then be considered how to tailor the methodology of both to the potential benefit of MetS research. This will encourage the development of a novel technique designed for this specific application in section 6.7.

6.4 PCA vs. EFA in Applied Research

A range of EFA and PCA methods have developed from the early work of Thurstone and Burt. They are extensively used techniques in the social, business and biological sciences. Both EFA and PCA provide tools for dimension reduction and are similar in procedure, which commonly leads to the terms being interchanged in texts and statistical packages (e.g. SPSS/STATA label both techniques as forms of “factor analysis”). However, subtle methodological differences differentiate the methods. Conceptual differences in the techniques should ensure that the two procedures are not readily interchanged in definition or application (Costello and Osborne 2005; Fabrigar et al. 1999).

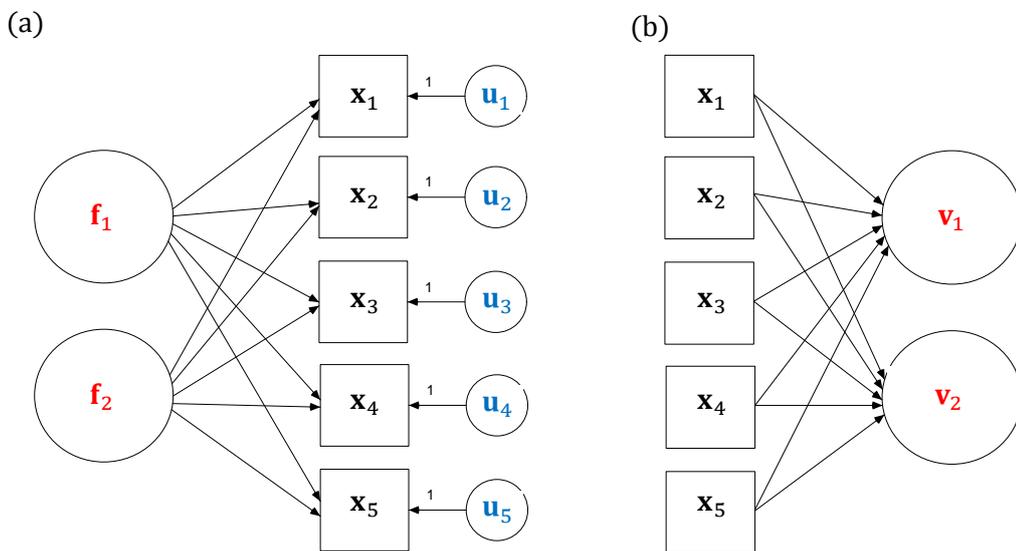


Figure 6.1: Model assumed by applying (a) EFA and (b) PCA.

The preferred use of an EFA or PCA is dependent on the hypothesis the user intends to assess. There is an important conceptual difference between factors and components. The model structures in Figure 6.1 demonstrate how the theoretical direction of causality between variables and latent constructs is reversed. Factor analysis assumes an underlying causal structure – that the covariation amongst the predictors is being *caused* by one or more latent factors - i.e. a reflective model. If the user believes latent factors to be exerting some influence on the manifest variables, then an EFA can be employed to discover and interpret those factors. For example, in the MetS data the observed variables are entered as “symptoms” exhibited by the patient. When an EFA is performed on the data, the researcher intends the factors produced to represent a “syndrome” as collectively they characterize some unobserved condition (Child 1990).

PCA is intended to maximize total variation in the first few components. It makes no distinction between unique and shared variance. The components are linear combinations of the original variables (i.e. influenced by the observed variables) – i.e. a formative model. The technique has been developed only for dimension reduction and makes no causal assumptions. For this reason PCA should not be considered a form of factor analysis, but an alternative technique with a different interpretation. The issue with PCA in many practical applications arrives when trying to apply the results to “real life” concepts (such as “syndromes” and “symptoms”). The PC’s generated are statistical constructs. Therefore, conceptually PCA should not be considered a form of factor analysis and freely interchanged by the user.

6.4.1 Decision Making in an EFA Study

A considered use of EFA can provide an insight into the underlying model structure of a set of manifest variables. However, the use of the technique requires careful decision making and an understanding of the potential consequences of each decision. Preacher and MacCallum (2003) define three major methodological decisions for employing EFA:

1. *the method used to extract the factors;*
2. *the number of factors to retain; and*
3. *the method of rotation to obtain an interpretable factor solution.*

In addition, the following methodological challenges can have a major effect on the interpretation of the results:

4. *the approach used to determine a ‘significant’ loading;*
5. *the sample size;*
6. *violation of the linearity assumption; and*
7. *using too few variables to construct a factor.*

These decisions define the nature of EFA and should be carefully considered by the user.

Factor Extraction

Although EFA and PCA should be labelled separate entities due to conceptual differences, this is not the practice of many researchers and statistical software. They are frequently described under the umbrella of “factor analysis” and seen as different methods of “factor extraction”. To discuss the mindset of researchers in employing a factor analysis and to avoid confusion over these techniques, this discussion will reluctantly include PCA as an

option in employing a “factor analysis”. In addition to PCA, methods of extraction include maximum likelihood (ML), principal axis factoring (PAF), iterative principal (IP) factor analysis and alpha factoring. With the exception of PCA, they each follow the same common factor model (as illustrated in Figure 6.1a), but proceed with different methods of computing the loadings. Each approach has advantages and disadvantages. For instance, ML estimation allows for the formation of confidence intervals, significance testing of factor loadings and goodness of fit statistics to be performed on the resulting model (Cudeck and Odell 1994). However, the procedure assumes multivariate normality (which is not an assumption of principal factor methods). Therefore, when the data does not adhere to the distributional assumption, principal factor methods are preferable – but with these, many of the statistics available to ML factoring are forfeited.

Factor Retention

If strong assumptions cannot be made about the number of factors, methods such as Kaiser’s criterion (in which factors are retained if the associated eigenvalues are greater than one) or retaining enough factors to explain a fixed variance percentage are often popular techniques. They can be implemented with no user intervention, thus appearing to give objectivity to the analysis. However, such an arbitrary threshold is not without limitations. There is often very little solid reasoning for choosing one limit over another. Why should eigenvalues greater than one indicate ‘important’ factors? Why should this threshold be transferable across subject fields? A number of studies have been produced to test the performance of Kaiser’s criterion and few appear to demonstrate support for the index in application (Fabrigar et al. 1999).

Another popular method used to decide the number of factors to retain is the scree test (Cattell 1983). This method involves plotting the eigenvalues of the correlation matrix in decreasing order and identifying a ‘substantial’ drop between factors. The number of factors included before the drop is then adopted for the analysis. In contrast to Kaiser’s criterion, this method is often criticized for being too subjective – it is left to the user to decide what represents a ‘substantial’ drop (if one exists). A final and perhaps more promising method is that of parallel analysis. Parallel analysis compares the eigenvalues obtained for the actual data against those of a random dataset matching the same criteria (i.e. equal sample size and variable number). This method appears to provide a balance between the ‘objective’ Kaiser’s criterion and the ‘subjective’ scree test. Simulation studies have shown this technique to perform well in practice (Humphreys and Montanelli 1975).

The objective of a factor analysis is to obtain a parsimonious model that adequately reflects a 'real life' construct. A considered EFA should encourage a combination of methods for factor selection to see if there is any agreement between them. It is generally thought to be more dangerous to choose too few factors (i.e. underfactoring) rather than too many (i.e. overfactoring) (Wood et al. 1996). When too few are retained, variables that would load on factors left out of the analysis can falsely load on those left in. In addition to this, factors that are independent may become combined, thus hindering the interpretation of the true factor structure. In contrast, by ignoring model parsimony and 'overfactoring' we can give strength to factors that demonstrate little evidence of any clinical importance.

Factor Rotation

The results of direct solutions from extraction methods such as PCA, PAF, ML etc. will follow the desired mathematical structure, but the loadings do not have to present a clear and interpretable solution in application. This is just one view of an almost infinite number of mathematically equivalent models. Thurstone (1940) proposed that the "best" structure (i.e. a structure that is interpretable, unique and replicable) is the "simple" structure – each factor defined by a small number of variables that load highly in relation to others and each variable loading 'highly' on a small number of factors (preferably only one).

- 1. Each row of the loadings matrix should contain at least one zero loading.*
- 2. Each column of the loadings matrix should contain at least k zero loadings.*
- 3. For every pair of columns of the loadings matrix, there should contain factors with a zero in one column and a high loading in the other.*
- 4. If more than four factors are extracted, then every pair of columns should have several rows with zeros in both columns.*
- 5. Every pair of columns should contain few rows with nonzero loadings in both columns.*

The direct solution will rarely realize this form. The variables will often have 'high' loadings on a number of factors, subsequently making interpretation difficult.

A factor rotation can be employed to obtain an approximate 'simple' structure, whilst still accounting for the same level of variation as the original (i.e. a derived solution). If the factors are imagined as axes (such as in the geometrical illustrations in Figure 4.6), a factor rotation can be demonstrated as a rotation of these axes about the origin (i.e. a

transformation of the factor loadings). The two main forms are an orthogonal rotation (which maintains an uncorrelated structure) and an oblique rotation (which allows correlation amongst factors). The varimax rotation developed by Kaiser (1958) has become the standard procedure for orthogonal rotation. It is considered that whilst an orthogonal structure is more generalizable across studies, an oblique rotation is often more likely to achieve the criteria of a 'simple' structure.

The strengths of the oblique and orthogonal rotation methods can be demonstrated on the MetS dataset. A series of bi-plot's are used to demonstrate the first two factors attained from a PCA extraction with various rotation methods employed,

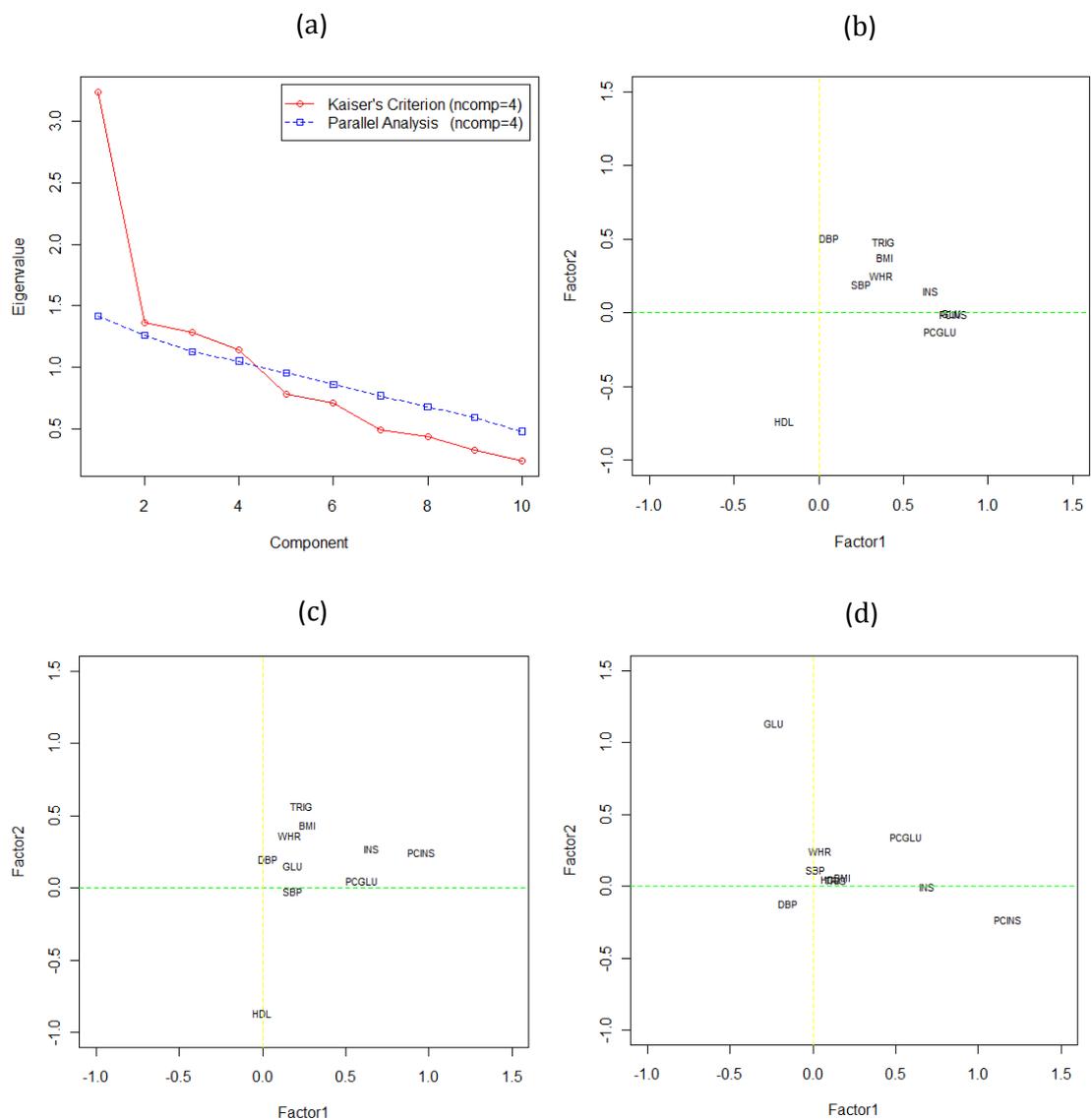


Figure 6.2: (a) Scree plot and bi-plot's of (b) raw, (c) varimax, (d) promax solutions.

Kaiser's criterion and parallel analysis are used to decide the number of factors to retain. Both methods recommend retaining 4 factors (see Figure 6.2a). In the first example, no rotation has been used (Figure 6.2b). Moderate loadings are observed on all variables with most shared evenly between factor 1 and factor 2. The third and fourth examples use a varimax (i.e. orthogonal) and promax (i.e. oblique) rotation (Cureton and Mulaik 1975) respectively. The factors retain their orthogonal structure in the varimax rotation, however in the promax rotation the variables are more clearly defined as belonging to a single factor or to none at all (i.e. clustering around the origin).

"The reason for using uncorrelated reference traits can be understood but it cannot be justified"

(Thurstone 1947)

The quotation above is Thurstone's opinion speaking as a psychologist, but this view should be transferable to most studies analyzing biological mechanisms. In utilizing an oblique rotation the complexity of the interpretations will increase as the user must consider the correlation of the factors along with interpretation of the loadings. However, in many cases correlated factors will be more likely to reflect structures closer to 'real life' biological systems (Child 1990). Obtaining a solution that is interpretive and meaningful should be the ultimate goal of exploratory research.

A Further Consideration

The three methodological decisions listed by Preacher and MacCallum define the computational options available to the researcher in an EFA study. Along with these, there are equally important interpretational challenges. A major decision is deciding what constitutes a 'significant' factor loading. Child (1990) proposes an arbitrary cut-point of ≥ 0.3 for data with sample size equal to or greater than 100 (further alternative criteria have been suggested for different sample sizes). Another suggestion by Burt and Bank's (1952) is that from the first factor onwards the levels to which the loadings become significant should increase, to compensate for the lesser variance explained by the factor. However, the criterion employed is still a rather subjective decision that the researcher must make and justify. As yet, an adequate method of testing the significance of a loading does not exist (Leng and Wang 2009). There should always be application specific knowledge used when interpreting the components as opposed to arbitrary measures and so application of such techniques for our purposes would require greater thought.

6.4.2 Limitations of Factor Analysis

In 1967 Armstrong published an article discussing the utility of factor analysis in applied research. He used a sample dataset to demonstrate the pitfalls of ‘automatic’ decision making in the use of EFA. The example contained 11 measurements of 63 right angled parallelepipeds of various sizes. The hypothetical situation defined that the user had no previous experience in these particular objects, but intended to use factor analysis to define a classification system. The 11 measurements were as follows,

- | | | | |
|------------|-------------------|-------------------------|------------------------|
| 1. Length | 4. Volume | 7. Total surface area | 10. Thickness |
| 2. Width | 5. Cost per Pound | 8. Cross-sectional area | 11. Length of internal |
| 3. Density | 6. Weight | 9. Total edge length | diagonal |

The reasoning behind the user employing EFA was that the measures presented a high degree of collinearity and they aimed to reduce the number of covariates to produce a more economical classification system. Unknown to the user, the actual structure is very simple. The characteristics of the object can be completely defined by measures 1-5 and the remaining six are combinations of the initial five. A PCA was performed with the user adopting ‘popular’ methodology, which included retaining factors with eigenvalue greater than one (i.e. Kaiser’s criterion) and using a varimax rotation method (see Armstrong (1967) for original PCA results). They found that the approach suggested 3 factors with an explained variance of 90.7%. Armstrong suggests that these factors may be describing (I) “compactness” (II) “intensity” and (III) “shortness”. However, knowing the structure, it is frustrating that PCA produced a complicated structure that only explains a portion of the variance, when it can be fully explained by a very simple model.

As is often the case, the supervisor has failed to provide (or has) valid reasoning behind the methodological decisions taken in the analysis. The decisions have led to an uncomfortable set of results that leave the user forming sketchy descriptions of abstract factors, which are unrealistic of the population model (i.e. a heuristic interpretation). The supervisor may have arrived at these decisions for any number of reasons; (1) they are ill-informed of the procedure, the alternatives, or the effects of the decisions; (2) A certain approach appears traditional (e.g. the literature suggests an optimum method, or by employing the same procedure it enables a comparison of results); (3) the ‘default’ options of the statistical software make the decisions for the user (Fabrigar et al. 1999). Another

researcher could conceivably repeat the analysis with more appropriate methodology. For instance, employing *prior* knowledge of the components which traditionally make up similar classification systems would have aided the user in deciding the number of components to retain. Both models (as well as numerous others) are statistically acceptable, however employing reliable *a priori* knowledge is likely to lead to a result that is representative of the clinical hypothesis the user intends to form.

When the conclusions of an explorative study are heavily dependent on the application of the method, the reasoning behind each decision must be made on stronger grounds than those in this illustrative example. The decision making process and the interpretation of the components make EFA a very subjective approach. The ease and speed of performing an EFA on modern statistical software has encouraged widespread use of the methodology, but this should only serve to heighten the caution adopted with the results. Despite repeated attempts to warn against the dangers of misguided decision making in factor analysis (Fabrigar et al. 1999; Floyd and Widaman 1995), they are still too commonly found in the clinical literature. A further danger is that the decisions (and the reasons for them) are rarely commented on in the literature to allow the reader to make a judgement on the validity of the analysis. The analysis can be easily swayed in a particular direction without solid reasoning for doing so.

"At the present time, factor analysis still maintains the flavour of an art, and no single strategy should yet be chiselled into stone"

(Johnson and Wichern 2002)

There are very few guidelines for researchers undertaking an EFA in applied research. To add further confusion, many of the same methods exist under different terminology across statistical packages (Costello and Osborne 2005). Often researchers choose to employ the 'default' options in the software, which gives strength to methods such as PCA, Kaiser's criterion and the varimax rotation. These decisions (along with the others in an EFA) should be appropriate to the data under consideration and the hypothetical questions being analyzed (such as the concept of MetS). The example used by Armstrong was in fact very simple. However, this is rarely the case in many applications. The absolute model (if one exists) will often be more complex in clinical examples and there will be less background theory to guide the researcher. Fabregar et al. (1999) suggests that it is not the utility of the method that we are concerned about, but rather the way in which it is employed. However, the utility of the method must be questioned when it is so heavily dependent on its application.

6.5 The Relationship of Factor Analysis and Metabolic Syndrome

Returning to the concept of heredity discussed in section 2.1, an argument that has been regularly levelled toward evolutionary theory is Plato's concept of 'essentialism' (Dawkins 2009). Plato saw the universe as truly perfect and that as humans we can only view a poor interpretation of this perfect state (i.e. the perceived universe).

"...although they make use of the visible forms and reason about them, they are thinking not of these, but of the ideals which they resemble "

(Plato)

The essentialist viewpoint implies that everything we perceive is simply a deformed copy of a perfect state. For instance, the species label of a 'homo sapien' describes an unchanging universal form that defines a human. This theory however counters the concept of evolution and natural selection – that any two living things have evolved from a common ancestor at some point in the evolutionary chain i.e. "population thinking" (Mayr 1970). The idea of a perfect human would appear counterintuitive of this notion and that the species label is just an artificial descriptive term. The essentialist viewpoint is often one taken when defining MetS. If we view the concept of MetS as an essentialist we consider an absolute model never changing. However, striving for an absolute model of the mechanisms (and subsequently the concept) may explain some of the difficulties encountered in the study of MetS and defining such a construct.

There is growing evidence to suggest that factors defining socio-economic position (SEP) such as income, education and occupation can have a 'significant' impact on the prevalence of T2-DM and CVD. Therefore, it is highly likely that the study environment in which the subject exists can be having a substantial effect on the structure of MetS (Ford and Giles 2003; Loucks et al. 2007). Along with SEP, factors such as smoking, alcohol and sleep levels have all been proposed as risk factors for T2-DM and CVD (Kaplan and Keil 1993; Lawlor et al. 2002). From this evidence it appears dangerous to define MetS as though there exists an absolute model across populations and population individuals (i.e. as an essentialist). If we accept that the form of the construct can change then this should motivate us to understand how and why it changes. For this, methodology is required that can identify consistency as well as change across constructs. Whilst the use of CFA is increasing in popularity in the study of MetS, the use of EFA is essential to generate hypotheses across populations and analyze new risk factors (Tang et al. 2005).

6.5.1 Factor Analysis and MetS

In the MetS example discussed in section 6.1, Shen (2003) considers evidence from a range of exploratory studies to construct three hypothetical models of the structure of MetS. The evidence was gained from the use of EFA and in particular PCA. The subjective nature of factor analysis as an exploratory technique is highlighted by Shen. The series of factor structures underline the range of potential hypotheses and heuristic interpretations. Instead, a CFA is employed based on the results of previous EFA studies and biological knowledge. The use of CFA is repeated in Shen (2006) to examine the structure of MetS across sex and ethnic groups, citing the conflicting and inconsistent results of EFA studies as the motivation for this approach. Combined with sound *prior* knowledge, a CFA can be used effectively to validate potentially complex structures; allowing the testing of specific questions about the nature of underlying mechanisms (Lawlor et al. 2004).

The difficulty with employing CFA in the MetS example is the sheer range of potential structures suggested by previous studies (as commented on by Shen). In MetS analysis there are many hypothetical structures available that can be tested, but little definitive evidence (as illustrated by the numerous definitions). A primary reason for the varying structures is that researchers can bias an EFA, using particular thresholds or rotational methods, to prize out a preconceived structure. Additionally, the (potentially default) orthogonal rotations and PCA extraction used in many applications of factor analysis for MetS data will be unlikely to represent the concept of an underlying construct and appear contradictory to the notion of a single unified syndrome (Shen et al. 2003).

To gain a solution using EFA a principal axis factoring (PAF) method is selected for the Shen data (incidentally, the ML solution of this data generated an ultra-Heywood case – i.e. a negative variance on a factor, which invalidates the result - see Khatree and Naik (2000)). As with any 'true' factor analysis method (i.e. excluding PCA), PAF extraction is based on the common factor model. The intention of employing this model is to capture the clinical notion of an underlying construct amongst the manifest variables (as discussed in section 6.4). The scree test and parallel analysis are used in addition to the arbitrary cut value employed by Kaiser's criterion. A substantial drop is not clearly defined, however each suggest the use of a four factor structure. Studies analysing similar risk factors of MetS have also proposed a four factor structure and so this will form the basis of the EFA model (Lafortuna et al. 2008; Shah et al. 2006).

Variable	Loadings				Communality h ²
	Factor I	Factor II	Factor III	Factor IV	
Ins	0.69	-0.07	-0.07	0.25	0.55
PCIns	0.76	-0.05	0.12	0.14	0.62
Glu	0.43	-0.18	0.37	-0.12	0.36
PCGlu	0.58	-0.08	0.48	-0.10	0.58
BMI	0.56	-0.06	-0.24	0.16	0.40
WHR	0.46	-0.05	-0.26	0.11	0.29
Trig	0.37	0.20	0.30	-0.30	0.36
HDL	-0.50	-0.09	-0.23	0.32	0.41
SBP	0.35	0.58	0.06	-0.10	0.48
DBP	0.30	0.61	-0.07	-0.03	0.47
Variance Expl.	2.71	0.81	0.65	0.34	4.51

Table 6.2: PAF extraction (Loadings > .3 are in bold to indicate significance).

Using the arbitrary 0.3 threshold suggested in Child (1990), it is clear that the direct solution does not adhere to the criteria for Thurstone's 'simple' structure (i.e. 'significant' loadings highlighted in bold). For instance, Trig demonstrates a 'significant' loading on three of the four factors.

Variable	Loadings				Communality h ²
	Factor I	Factor II	Factor III	Factor IV	
Ins	0.66	0.30	0.12	0.13	0.55
PCIns	0.57	0.50	0.17	0.14	0.62
Glu	0.11	0.58	-0.01	0.11	0.36
PCGlu	0.17	0.73	0.12	0.08	0.58
BMI	0.57	0.11	0.09	0.23	0.40
WHR	0.47	0.03	0.08	0.23	0.29
Trig	-0.19	-0.06	0.03	-0.56	0.36
HDL	0.23	0.17	0.11	0.56	0.41
SBP	0.08	0.13	0.67	0.04	0.48
DBP	0.14	-0.01	0.67	0.03	0.47
Variance Expl.	1.47	1.26	0.98	0.79	4.51

Table 6.3: A Varimax rotated factor pattern for the PAF extraction.

Variable	Loadings				Communality h ²
	Factor I	Factor II	Factor III	Factor IV	
Ins	0.70	0.11	-0.01	-0.05	0.55
PCIns	0.52	0.37	0.04	-0.03	0.62
Glu	-0.06	0.63	-0.08	0.06	0.36
PCGlu	-0.04	0.77	0.04	0.00	0.58
BMI	0.63	-0.08	0.00	0.09	0.40
WHR	0.53	-0.13	0.00	0.12	0.29
Trig	-0.08	0.00	0.06	-0.57	0.36
HDL	0.07	0.09	0.08	0.55	0.41
SBP	-0.06	0.08	0.69	0.01	0.48
DBP	0.06	-0.11	0.69	-0.01	0.47
Variance Expl.	2.34	1.89	1.27	1.28	4.51

Table 6.4: Promax factor solution (Loadings > .3 are in bold to indicate significance).

The oblique solution is very similar to the orthogonal, with generally higher loadings increasing further and smaller loadings reducing further. The loadings are approaching a ‘simple’ structure. Evidence of this can be seen with Ins becoming insignificant on factor II after oblique rotation. This risk factor is only ‘significantly’ loaded on factor I. However, it is important to remember that oblique factors are correlated as demonstrated in Table 6.5.

	Factor I	Factor II	Factor III
Factor I			
Factor II	0.54		
Factor III	0.36	0.27	
Factor IV	0.49	0.27	0.13

Table 6.5: Correlations between factors in the ‘promax’ solution.

Factor I and factor II display the largest correlation, with factor I and factor IV also moderately high. It appears from the loadings that Ins, PCIns, BMI and WHR could form part of a system along with Glu and PCGlu. Lipid factors are clearly marked as Factor IV. There is evidence to suggest that blood pressure (i.e. factor III) is independent of the other factors in this study, but that obesity may still be associated with this factor. This alone would not prove (or equally discount) the potential existence of a single unified factor, such as that hypothesized for MetS.

An orthogonal solution is regularly promoted as easier to generalize amongst studies. The interpretation is often simpler (with no correlations amongst the factors) and there is a clear favourite rotation method amongst the orthogonal solutions - i.e. the 'varimax' rotation. Also, the loadings produced by an orthogonal and oblique rotation are usually similar (as seen in our example). However, by opting for the simpler interpretation, the user is sacrificing the 'simple structure' of the factors to the data that an oblique rotation can achieve. This is the view of Thurstone (1947) and Cattell (1978) who suggested that "in half of these cases it is evidently done in ignorance of the issue rather than by deliberate intent". The issue appears to be that the analysis has to be simplified to achieve a consistency in the decision making across studies at the expense of the machinery available. When poor decision making in exploratory analysis is spread over a large scale of studies (such as the study of MetS), they can be of great hindrance to a researcher intending to make valid conclusions from the findings. If to achieve a greater consistency in MetS study we have to choose unsuitable options then it appears necessary that the field considers alternative methodology better suited to the application.

6.5.2 Why is Factor Analysis Chosen for MetS?

A major reason that some decisions have become popular amongst researchers is due to the complexity of the methods and interpretation. Employing Kaiser's criterion requires no user input and an orthogonal rotation may experience a simpler interpretation. However, as highlighted by Shen (2003), orthogonal factors would appear counterintuitive of an underlying syndrome. If the factors happen to be orthogonal, an oblique rotation will result in orthogonal factors anyway (and provide counter evidence to the hypothesis of a single unified syndrome). This demonstrates interesting parallels to the comparison of the Spearman and Thurstone models of intelligence. The decision was directly related to the hypothesis, which is also true in MetS application. The scree test or parallel analysis demonstrate a considered reasoning by the user for the number of factors to retain. Unfortunately, 'default' choices in statistical packages give strength to options such as PCA and Kaiser's criterion without particular reason.

When biological or statistical evidence is available for a researcher to confidently propose a single or small number of plausible models, then a CFA can be an effective approach. It will narrow the decision making automatically for the user by determining features of the model. Also, as opposed to the primarily 'data driven' nature of exploratory

analysis, the assumptions made mean that the user will be less likely to obtain 'chance' results from the dataset. When the researcher cannot be certain of the form of the model, EFA should become the preferred choice. The form of the model is open to ensure that plausible models are not ruled out of the analysis. This is particularly useful when investigating data from different populations or including lesser studied risk factors in the analysis. It appears that for an EFA to provide consistent and valid reasoning to be employed for a MetS study we often require a greater complexity than the 'default' options used in statistical software. This is a problem highlighted by many researchers such as Armstrong, Fabregar and MacCallum in the psychological literature.

It is important to emphasize that the use of CFA should not be considered invalid or redundant for MetS analysis from this discussion. Various tests and measures of model fit exist to validate the structure of MetS and identify when a hypothetical structure does not fit the study data well. However, are we confident that the state of knowledge in proposing the MetS structures is sufficient for the use of CFA? The answer to this question is likely to differ greatly amongst researchers in the field. Current evidence, definitions and debate regarding the MetS structure and its existence suggests that EFA still has an important role to play in our understanding of this construct. The use of either methodology rests at what point the user believes the current knowledge to exist on the continuum from EFA to CFA. Can a complete model be suggested, or is there sufficient uncertainty that an explorative approach can relieve? These methods are not separate entities; they are instead a reflection of the confidence in the *a priori* knowledge. A considered and justified decision making process for EFA research can provide a powerful tool in developing an understanding of the MetS construct in partnership with CFA.

The reasons suggested in section 6.4.2 for the misuse of EFA in such applications all appear to stem from a similar cause - the user does not fully understand the factor analysis procedure and the impact of their decisions. This has to be expected as in the majority of cases non-statisticians will not wish to search through complex mathematical articles to understand a procedure when computer software performs the task with minimum effort. It appears unrealistic to expect a great number of applied researchers to readily employ and understand the reasoning for this more complex methodology when simpler methodological options would seem to provide similar results. Instead, alternative techniques may be proposed that could provide consistent and easily interpretable results intended for MetS application without the need for such complex decision making. The discussion returns to cluster analysis to provide motivation for these methods.

6.6 Cluster Analysis

The intention of this work is to improve the consistency in results and interpretation across MetS studies. Although many algorithms exist for clustering objects, the results of ‘hard’ clustering (i.e. clustering observed covariates) and in particular hierarchical clustering methods can bring substantial improvements in these areas. One of the major motivations behind employing a hierarchical cluster analysis is the visual element to aid with interpretation. The user is able to observe the natural groupings in the data and see where the relationships originate from. The visual element of hierarchical clustering is seen as particularly beneficial for small datasets (such as the MetS example).

“Cluster analysis relies much more upon subjective judgement and much less on statistical analysis than factor analysis”

(Child 1990)

A main criticism by Child is the number of choices available in performing a cluster analysis, with the algorithm and proximity criteria demonstrating little relevance to ‘real world’ concepts. A similar ‘subjectivity’ of EFA has been highlighted regarding the decision making and interpretation, and so to balance the argument these issues will be discussed for clustering methodology in the remainder of the chapter.

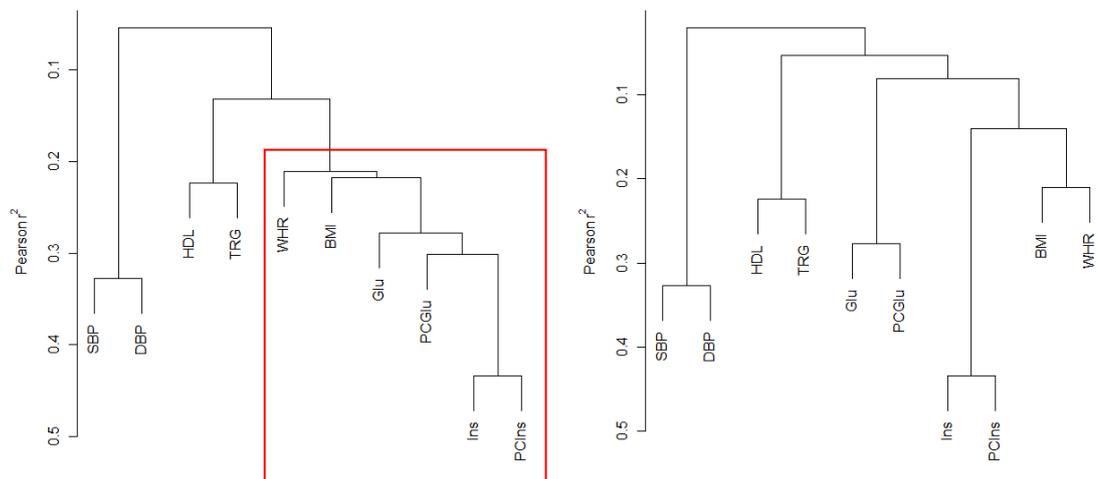


Figure 6.3: Cluster analysis using Pearson r^2 with (a) single and (b) average linkage.

The hierarchical cluster analysis in Figure 6.3 demonstrates the results of a variable clustering of the MetS dataset (i.e. R-mode). The two examples in Figure 6.3 appear different in their interpretation of the data. This difference can be expected as algorithms are designed to extract clusters with different features. For instance, single linkage will often produce chains of variables (e.g. highlighted in red in Figure 6.3a), whilst average linkage will generally produce 'compact' clusters. The analysis of observed covariates rather than abstract factors should make clustering techniques an attractive option in applied research. The problems associated with a heuristic reading of components in factor analysis are much simplified by considering distinct clusters, allowing for datasets with a large numbers of variables to be analyzed with substantially less difficulty and improved consistency. As the clusters do not overlap (as with factor analysis), hierarchical clustering allows for images to be constructed to aid with interpretation. Also, the techniques make no assumptions about the underlying distribution of the data (i.e. violation of the linearity assumption is avoided).

An issue that hinders cluster analysis as a technique to identifying latent structures is that the analysis is based on pair-wise near dependencies. This means that underlying relationships amongst covariates may not be identified (similar to using correlations to diagnose collinearity) - for example, a variable z can be approximated as a function of x and y , but none of the variables are involved in a pair-wise near dependency. This will be problematic for any method dependent on a distance or similarity metric. Another issue is that agglomerative hierarchical methods do not have any procedure to rectify wrongly placed variables in later results. Once a variable forms part of a cluster, it cannot move to another if it appears more applicable in lower thresholds. Some researchers also argue that in an applied setting, clustering methods suffer in a similar subjective manner to EFA - different similarity and distance measures can be used to confirm the prior hypotheses of the researcher (Anderberg 1973). Differences can also occur due to fluctuations in the data (often leading to criticism of purely data driven methodology).

6.6.1 The VARCLUS Procedure

An alternative approach to clustering variables using similarity/dissimilarity measures is to utilize PCA in a hierarchical clustering framework - labelled the VARCLUS approach. The process builds clusters around latent components. The technique computes the first PC of each cluster (beginning at a cluster containing all the covariates) and iteratively splits them

into two separate clusters based on some predefined criteria. For example, the user may employ Kaiser's criterion to suggest that any cluster with an eigenvalue greater than one demonstrates evidence of an additional dimension or the user can pre-define the number of clusters to extract. The variables are assigned to the cluster in which they demonstrate the highest squared correlation (R_x^2) and later reassigned if the variance explained increases by including it in another cluster. Each cluster is formed from a different overhanging set of variables to any other, and so the orthogonality assumption of PCA is relaxed. The components obtained are naturally oblique and are subsequently referred to as "cluster components" rather than principal components.

The process compromises on the maximal variance extraction of a traditional PCA to maintain intuitive advantages of clustering observed variables. In addition, VARCLUS provides an R_x^2 for each variable within its own cluster (similar to a VIF - labelled R_{own}^2), and also with the nearest cluster in which it demonstrates the greatest R_x^2 (labelled $R_{nearest}^2$). If the clusters are well defined (i.e. the degree of association is maximal for variables within the same cluster and minimal to those in others) the value will be high for their own cluster and low for the nearest. PROC VARCLUS in SAS provides a ratio of the two values, calculated as $(1 - R_{own}^2)/(1 - R_{nearest}^2)$. Low values indicate 'strong' clustering. These measures are particularly useful when considered with the limitations of the ordinary 'hard' clustering procedures. Whilst the results of a VARCLUS are of the form of a hard clustering method, the R_x^2 ratio gives an indication of how 'fuzzy' the clusters are. The R_x^2 results can be interpreted in a similar way to component/factor loadings in an EFA.

The exploratory methods discussed in this chapter are based on different statistical approaches and so agreement cannot be guaranteed between the methodologies. Methods of EFA optimize the fit of the data to a model in which the "common components" of the observed variables are expressed in terms of a k -dimensional collection of "common factors" for some selected k . Only in a secondary step are rotation methods used to provide some indication of how the original variables cluster together, and the identification of such clusters is generally *ad-hoc* and not incorporated in the original model fitting. By contrast, the VARCLUS approach sacrifices the goal of providing a best fitting k -dimensional representation, but instead is designed to directly identify clusters of observed variables. This achieves maximal simplicity by postulating only 1-dimensional clusters of mutually pair-wise correlated variables. The various interpretations given by these procedures can be used to strengthen research, by reducing the subjectivity of relying solely on one methodology.

6.6.2 Application of VARCLUS

Two examples of the VARCLUS procedure are performed on the MetS dataset. The first uses PCA to define the cluster components and a maximum eigenvalue of 1 (analogous to Kaiser's criterion) to define the point at which the clusters are split.

	Members	Variance Explained	Proportion	2 nd Eigenvalue
Cluster I	4	1.95	0.49	0.98
Cluster II	2	1.57	0.79	0.43
Cluster III	4	2.29	0.57	0.93
Tot. Variance:		5.81		

Table 6.6: A summary of the cluster components generated.

The cluster summary in Table 6.6 highlights that three cluster components were formed from the options specified. These explain 58% of the total variation in the dataset, with cluster III explaining the largest variation (these values are similar to variance explained in an EFA). The 'proportion' column represents the variance of the covariates in the cluster divided by variance explained.

	Variable	R^2_{own}	R^2_{nearest}	$1 - (R^2_{\text{own}}/R^2_{\text{nearest}})$
Cluster I	BMI	0.50	0.17	0.60
	WHR	0.47	0.09	0.59
	Trig	0.48	0.05	0.54
	HDL	0.50	0.12	0.57
Cluster II	SBP	0.79	0.06	0.23
	DBP	0.79	0.02	0.22
Cluster III	Ins	0.54	0.21	0.58
	PCIns	0.70	0.20	0.37
	Glu	0.42	0.05	0.61
	PCGlu	0.62	0.06	0.40

Table 6.7: R^2 measures demonstrating the quality of each cluster component.

Table 6.7 lists the members of each cluster. Cluster II represents a hypertension component and appears well defined (i.e. high R^2_{own} and low R^2_{nearest}). Cluster III appears to be describing insulin resistance and cluster I combining obesity and lipid metabolism measures. This cluster structure suggests a potential association between lipid metabolism and obesity.

	Cluster I	Cluster II	Cluster III
Ins	0.46	0.19	0.74
PCIns	0.44	0.23	0.84
Glu	0.22	0.05	0.65
PCGlu	0.24	0.17	0.79
BMI	0.71	0.16	0.41
WHR	0.68	0.13	0.30
Trig	-0.70	-0.02	-0.22
HDL	0.71	0.15	0.35
SBP	0.14	0.89	0.24
DBP	0.16	0.89	0.15

Table 6.8: A cluster summary of variable loadings on each cluster component.

	Cluster I	Cluster II
Cluster I		
Cluster II	0.17	
Cluster III	0.46	0.22

Table 6.9: The inter-cluster correlations.

The cluster structure has the same interpretation as factor loadings in a factor analysis. The inter-cluster correlations are also analogous to inter-factor correlations of an oblique solution. They add further strength to the independence of a blood pressure factor. The 0.46 correlation between components 1 and 3, along with the high R_x^2 ratios appear to demonstrate the presence of an additional component.

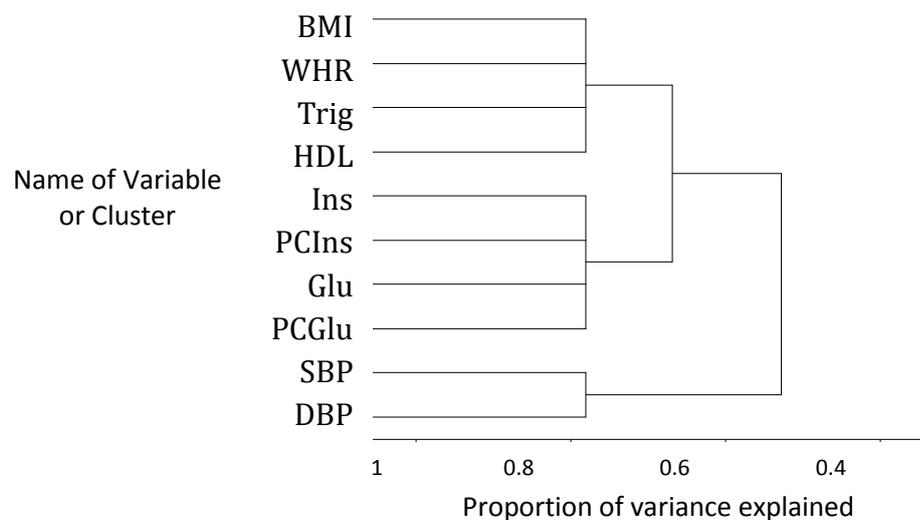


Figure 6.4: A dendrogram summary of the VARCLUS procedure with maxeigen=1.

The purpose of the following VARCLUS analysis is to investigate the strength of a four-cluster structure. A maximum cluster option of four is specified, whilst PCA is once again used to construct the cluster components.

	Members	Variance Explained	Proportion	2 nd Eigenvalue
Cluster I	2	1.47	0.79	0.53
Cluster II	2	1.57	0.79	0.43
Cluster III	4	2.29	0.57	0.93
Cluster IV	2	1.46	0.73	0.54
Tot Variance:		6.79	0.68	

Table 6.10: Cluster components generated for the four cluster solution.

Cluster	Variable	R ² _{own}	R ² _{nearest}	1 - (R ² _{own} /R ² _{nearest})
Cluster I	Trig	0.74	0.08	0.29
	HDL	0.74	0.12	0.30
Cluster II	SBP	0.79	0.06	0.23
	DBP	0.79	0.03	0.22
Cluster III	Ins	0.54	0.21	0.58
	PCIns	0.70	0.15	0.35
	Glu	0.42	0.04	0.60
	PCGlu	0.62	0.04	0.39
Cluster IV	BMI	0.73	0.17	0.33
	WHR	0.73	0.09	0.30

Table 6.11: R² measures demonstrating the quality of each cluster component.

Variable	Cluster I	Cluster II	Cluster III	Cluster IV
Ins	-0.30	0.19	0.74	0.46
PCIns	-0.33	0.23	0.84	0.39
Glu	-0.17	0.05	0.65	0.19
PCGlu	-0.19	0.17	0.79	0.21
BMI	-0.30	0.16	0.41	0.85
WHR	-0.27	0.13	0.30	0.85
Trig	0.86	-0.02	-0.22	-0.28
HDL	-0.86	0.15	0.35	0.29
SBP	-0.10	0.89	0.24	0.13
DBP	-0.08	0.89	0.15	0.18

Table 6.12: Cluster loadings of the 4 cluster model.

	Cluster I	Cluster II	Cluster III
Cluster I			
Cluster II	-0.10		
Cluster III	-0.33	0.22	
Cluster IV	-0.33	0.17	0.41

Table 6.13: Inter-correlations of the cluster components.

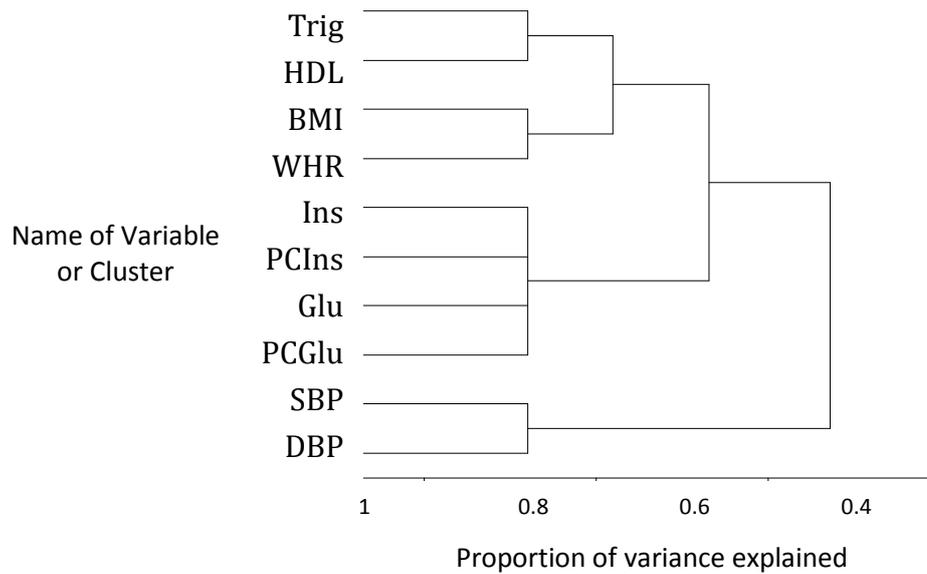


Figure 6.5: Dendrogram of the cluster components for the second VARCLUS analysis.

The cluster components of this analysis appear to relate to lipid metabolism (cluster I), blood pressure (cluster II), insulin resistance (cluster III) and obesity (cluster IV). This is analogous to model 1 and model 3 investigated in Shen (2003). Low R_x^2 ratios for clusters I, II and IV indicate that the components are ‘well formed’. However, cluster III exhibits high R_x^2 ratios for insulin and glucose predictors in particular. The cluster structure analysis confirms that Ins loads highly on cluster IV and relatively highly on cluster I. This is further evidenced by the correlation between cluster III and cluster IV.

The cluster structure explains 68% of the total variation and reduces the dimension of the variables from 10 to 4. This simple example has allowed us to gain an immediate insight into the cluster structure, whilst still observing that the variables are likely to be involved in multiple mechanisms. The cluster structure and R_x^2 statistics indicate which covariates appear ‘least comfortable’ within the clusters and with which others they are associated with. For instance, Glu has a high R_x^2 ratio (0.6), but it is not highly associated

with another cluster (i.e. low R^2_{own}) - the variable itself is not explained well within its own cluster. This adds further strength to the involvement of Ins and PCIns in other dependencies; namely, a relationship between BMI and insulin risk factors (as suggested in the EFA analysis). Also, HDL again demonstrates a high loading on the 'insulin resistance' cluster component (i.e. cluster III). In addition, the analysis provides further evidence as to the independence of a 'blood pressure' component (i.e. cluster IV).

It is a particular advantage of the VARCLUS methodology that a non-orthogonal structure can be achieved without needing to raise the complexity of the methodology or the interpretation substantially. The method has provided a structure for a CFA if required, but also indicated how stable the clustering is. The VARCLUS approach is a very effective method in balancing the sophistication of a PCA with the interpretive power of a cluster analysis. With this approach the user can choose an eigenvalue or variance explained threshold to base the analysis on, whilst providing a complete summary of the overall factor structure. The methodological decisions are limited, which provides an objective and consistent reasoning for the model they choose in MetS analysis. There are also direct links to features of an EFA, such as variance explained, factor loadings and inter-factor correlations. The R^2_x ratio provides a very useful tool for variable selection and again has a simple interpretation for the user. However, whilst the iterative reassignment of variables makes some attempt to rectify wrongly placed variables, this process is still somewhat hindered by the need to maintain a hierarchical structure in the computation.

A recent ISI literature search identified only 12 references of the use of VARCLUS in practice. The intention of statistical software such as SPSS is to make statistical methods available and simple to use to the non-statistician. The 'drop down' menus and default options mean that a PCA/EFA can be produced in a matter of seconds with no manual coding required. There is no doubt that this has a major impact on the use and popularity of these techniques. EFA and PCA have been developed over a long period and discussed at length in numerous articles. This has led to an accepted use of such techniques in a number of fields. It also potentially prejudices caution towards modern techniques such as VARCLUS, which has only recently been introduced through statistical software packages. The importance of such a technique in practice is that the statistics can be easily digested by the non-statistician, whilst also reaching closer to biologically justifiable decisions of a considered EFA application. It is this balance in disciplines of statistics and clinical relevance that should be encouraged and provides the motivation for the next methodology in this discussion.

6.7 Variable Clustering using Matroids

Although the method chosen to perform variable clustering still has a subjective nature, there are fewer methodological decisions and the ability to picture the form of a complicated structure is of great benefit in interpreting the results. However, in developing disjoint clusters, it appears necessary of each technique to ensure that the results are consistent and do not overlap. This is achieved by a continuous process of joining or splitting clusters based on a single criterion. At each threshold, the results build upon those attained in the previous to ensure continuity. The issue with this approach is that the results have to be reduced to only their uni-dimensional relationships, at the potential cost of missing higher dimensional relationships.

It is realistic for the user not to expect predictors in a complicated structure (such as MetS) to form 'neat' hierarchical groups. Rather this is forced with 'hard' clustering techniques (i.e. distinct clusters). This form of clustering is useful because of the benefits to interpretation it brings, but with it we bypass some of the subtleties in the relations that EFA attempts to identify (thus adding to an often difficult interpretation). In a complicated structure such as MetS it would seem likely that the predictors may be involved in multiple dependencies. It may therefore be useful to run both a VARCLUS and EFA with oblique rotation for comparison. The results can indicate how 'fuzzy' the clusters are and how much is compromised by gaining the intuitive advantages of the hierarchical VARCLUS procedure.

A novel methodology is instead proposed, labelled the matroid approach that could provide a compromise to this, whilst retaining the general 'hard' structure of the clusters. A matroid does not describe a method in itself, but rather a structure that adheres to an axiom system. In applying certain axioms that define the structure, we can look to avoid the consistency limitations of an ordinary variable clustering procedure. If the axioms of the matroid can be maintained within a dataset, the consistency of the clusters can be ensured, whilst retaining the subtle dependence structures present in the dataset. A matroid is a combinatorial structure that captures some notion of "independence". Hassler Whitney (1935) developed this notion of independence in a graph theory setting. Whitney identified similarities between this and the concept of linear independence in linear algebra. He also developed a notion of rank in graph theory which relates to the idea of dimensionality in linear algebra. Thus, a matroid can be viewed as a useful tool that generalizes the idea of independence across different fields in mathematics.

A matroid independence structure $\mathbf{M} = (\mathbf{S}, \mathbf{I})$ is defined such that the following axioms (I1) - (I3) are satisfied:

(I1) $\emptyset \in \mathbf{I}$.

(I2) If $\mathbf{x} \in \mathbf{I}$ and $\mathbf{x}' \subseteq \mathbf{x}$ then $\mathbf{x}' \in \mathbf{I}$.

(I3) If $\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{I}$ and $|\mathbf{x}_1| < |\mathbf{x}_2|$, then there exists $\mathbf{x} \in \mathbf{x}_1 \setminus \mathbf{x}_2$ such that $\mathbf{x}_2 \cup \mathbf{x} \in \mathbf{I}$.

(Welsh 1976)

These axioms provide the framework for a number of properties relating to bases, ranks and circuits of subsets. They can even be used as axiom systems in their own right to define a matroid. The result that is of particular interest to our application is that any subset of \mathbf{M} not labelled linearly independent is subsequently dependent.

6.7.1 Matroids to Identify Latent Structures

The use of matroids to identify linear dependencies is a recent development in the context of forming a dependency structure. Suggested by Greene (1990), the method draws from existing successful ideas in the field of collinearity diagnosis and cluster analysis, whilst also introducing favourable properties of matroids, which have previously been confined to theoretical work. The intention of employing this method is to produce a detailed mapping of the linear dependencies throughout a dataset, not only to provide an intricate picture of their structure, but to indicate how we may look to handle 'problematic' dependencies in future regression analyses. In this approach the general concept of 'hard clustering' is retained, in that the variables are considered in groups of a hierarchical form (although not of a traditional dendrogram), but effectively a dependency structure is produced at each threshold. The matroid approach seeks to identify not only 1-dimensional clusters of mutually correlated variable, but also higher dimensional near dependencies in which collections of the observed variables are identified as falling close to lower dimensional subspaces. Due to the flexibility of applying matroid axioms, it also allows the user to apply their own measure of linear dependence (i.e. a collinearity index) within the coding.

The Matroid Approach

The matroid approach works on the variables as subsets, rather than considering the entire set at once. Initially, data are divided into all the possible permutations of covariates. Therefore, for the MetS data, there are $2^{10} - 1 = 1023$ possible combinations of

covariates. These permutations are then assigned to either a ‘dependent’ or ‘independent’ group using a suitable collinearity index. For example, a ‘dependent’ subset may be defined by the smallest eigenvalue being lower than a particular threshold; any remaining subsets are subsequently labelled ‘independent’ (i.e. similar to the CI).

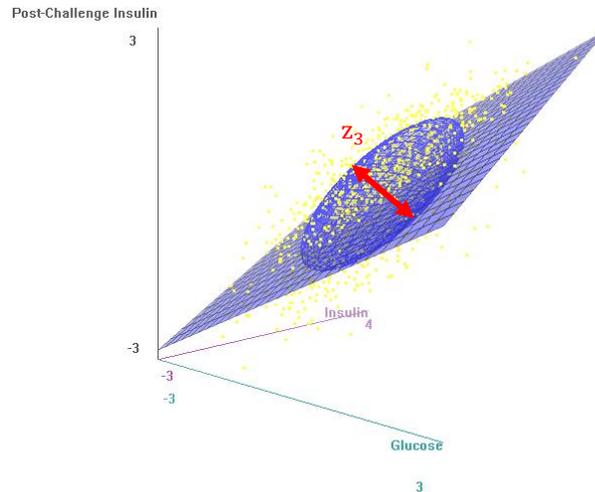


Figure 6.6: Plot of PCIns, Ins and PCGlu, indicating the smallest eigenvalue.

The plot in Figure 6.6 illustrates the subset of PCIns, Ins and PCGlu with the minimum eigenvalue indicated by the red arrow. The cardinality of this subset is 3, however if the selection criteria assigns it to the dependent set, the clustering may be displayed as a 2-dimensional plane. It is clear from the theory of PCA that using the minimum eigenvalue will give a measure of the variance of the final PC (see section 4.2.1). If this value falls below a chosen threshold, then a redundant dimension can be defined and therefore a linear dependency. If the threshold used to assign the subsets is varied, we capture dependencies at multiple levels in the dataset, which are used to build a detailed mapping of the linear dependencies.

The challenge with the matroid technique is how to convey the information of all the dependent subsets in the simplest form to the user. Greene suggests extracting a combinatorial group of permutations known as flats. A rank- j flat is a maximal set of covariates that can be represented by a j -dimensional projection (Greene 1991). Therefore, the flats ensure that every covariate involved in a dependency is identified. If we are unable to add another covariate (\mathbf{x}_j) to the subset without increasing its rank, then it is labelled a flat (see eqn(6.1)).

$$\text{rank}(\mathbf{X} + \mathbf{x}_j) = \text{rank}(\mathbf{X}) \quad (6.1)$$

The Labelled Hasse Diagram

The illustration proposed by Greene to display the flats is termed a labelled Hasse diagram (LHD). It is a hierarchical graphing of the dependencies in the dataset, in which the rank (i.e. the lowest dimensional projection that accurately approximates the subset) is illustrated by the horizontal level it occupies. In order to present data in hierarchical form it is vital that the subsets conform to the basic axioms of linear dependence (see Whitney (1935)). For instance, if there exists three covariates in which two pairs can be described by a uni-dimensional projection, then the third should similarly be described by a uni-dimensional projection. It is very possible that in real data, the subsets will be inconsistent with these laws. Matroids do not suffer these problems and by converting the dependent set into a matroid the subsets demonstrate a combinatorial structure corresponding to linear relationships amongst a collection of variables.

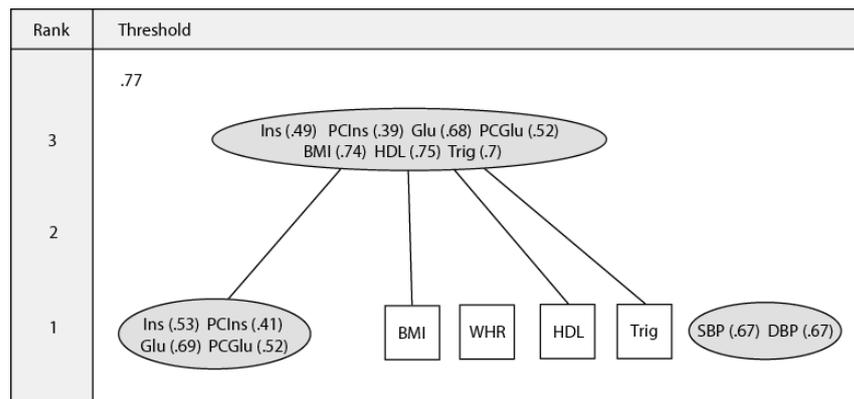


Figure 6.7: An example of an LHD depiction.

The LHD demonstrates distinct differences to the dendrogram used in a hierarchical cluster analysis. They both share a hierarchical structure and link groups of covariates (i.e. flats/clusters). However, the LHD differs in that each threshold produces its own hierarchical structure containing dependencies of any dimension. The flats are displayed as ellipses and those presenting no dependency as squares. The rank of each subset is illustrated on the left of the LHD. The flats joined with lines are to show the sources of any dependency. A multiple R_x^2 (in brackets alongside each variable) in the form of a VIF (see eqn(3.21)) has been added to the LHD diagram suggested in the original paper by Greene. The motivation is from the R_x^2 measures present in the VARCLUS analysis which provided a useful measure of the ‘fuzzy’ nature of a ‘hard’ structure.

Selection of the Matroid Criteria

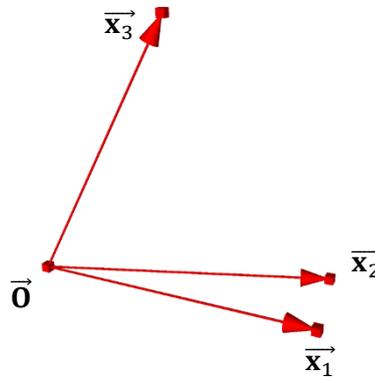


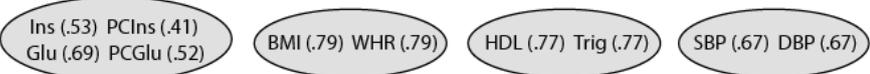
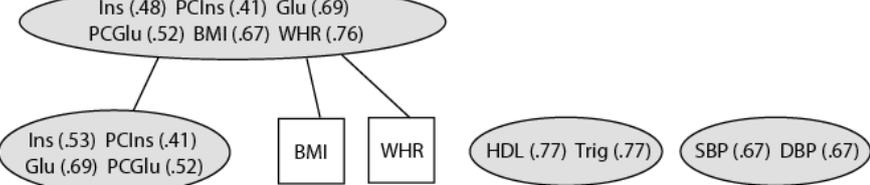
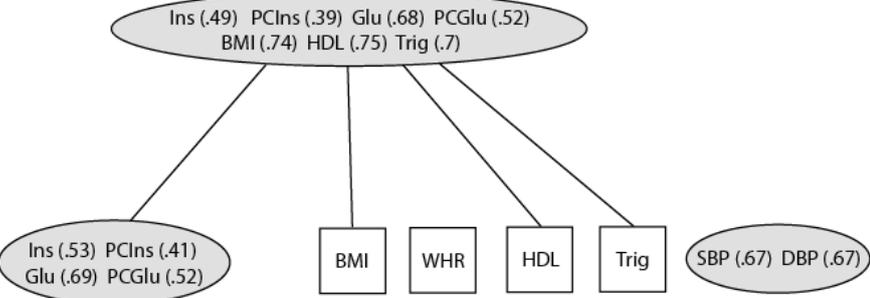
Figure 6.8: Problematic situation using the eigenvalue selection criteria.

The minimum eigenvalue is a selection criteria suggested by Greene, however it has limitations as part of the process. The definition of linear dependence is that each column should be dependent on the other columns in the subset (Greene 1990). Figure 6.8 illustrates a situation that should not be termed linearly dependent. However, the minimum eigenvalue approach would generally identify this subset as dependent due to the high correlation between x_1 and x_2 . If the intention is to identify flats to show higher ranked dependencies between uni-dimensional flats, then selecting this subset as dependent would be an issue as \bar{x}_3 is almost orthogonal to the other two covariates.

The advantage in this work is that existing measures of collinearity can be employed. Any ‘correlation based’ index could in theory be used in the matroid framework (when no response is defined). If it is difficult to justify different linkage methods in cluster analysis as rigorous approaches to identifying dependencies amongst the predictors, then this should be considered an improvement in using matroids. However, as demonstrated by the minimum eigenvalue criteria, the measure must be chosen carefully within the context of the method. They will adopt a different meaning when used as part of a collective analysis. One criterion that it would be of interest to test is the D-index developed in chapter 5. It would seem ideal to place a D-statistic on each of the subsets, use component measures to indicate covariate involvement in each dependency and analyze the consistency of the measure as part of the matroid framework. However, this index would not be suitable for the current application of MetS as there is no response variable. Instead, the following analysis takes a step back and employs a criterion based on the VIF. This would appear more suitable for the current application.

MetS Results

The matroid procedure was coded using R (version 2.8.1). The method has been applied to the Shen dataset using a ‘minimum VIF’ criteria - if a subset displays a VIF higher than the given threshold it is assigned dependent. In this example the threshold has been modified to represent the inverse of the minimum VIF to give the LHD a finite scale of 0-1 (i.e. the tolerance).

Rank	Threshold
1	.88 
1	.8-.87 
1	.79 
2	.78 
3	.77 

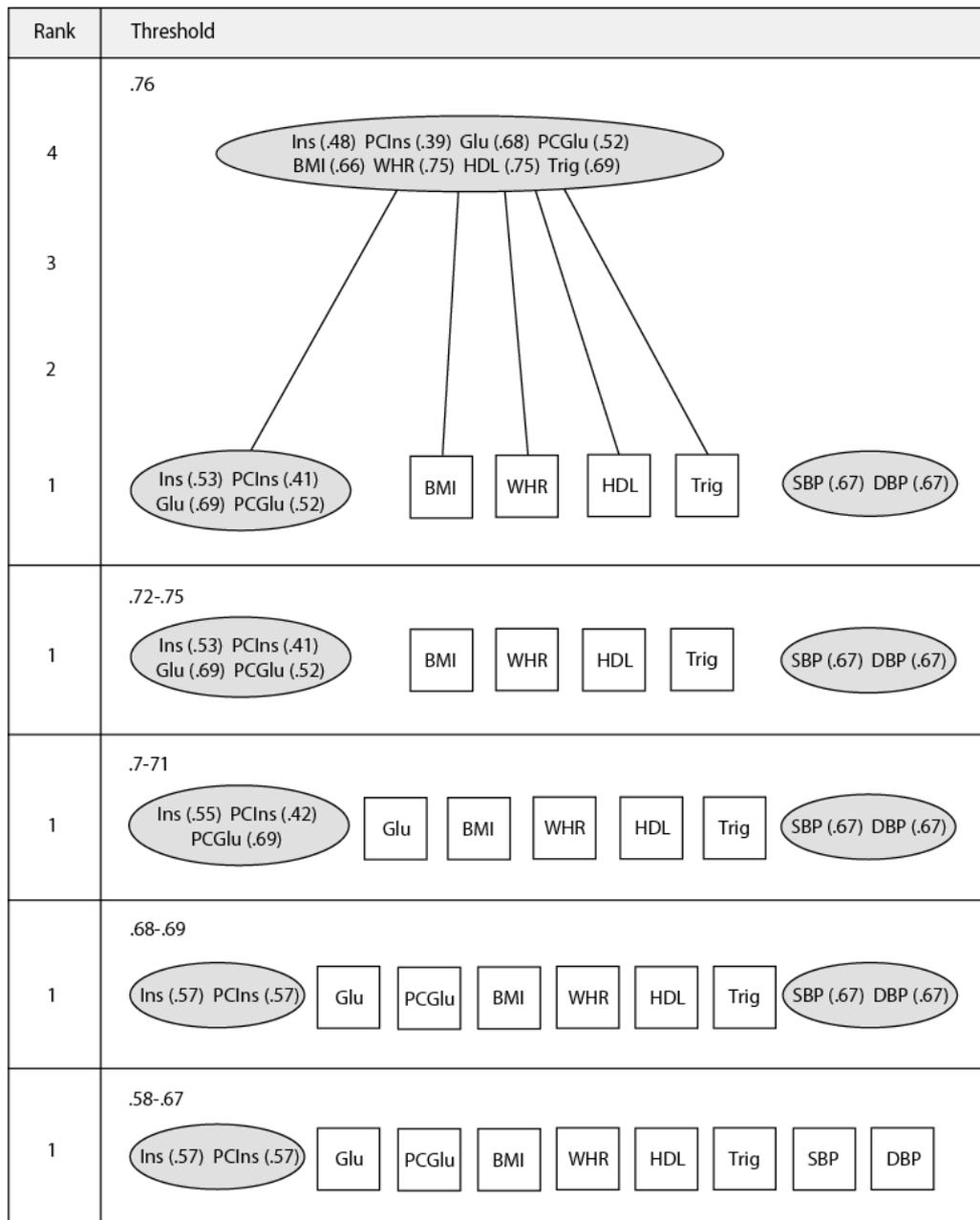


Figure 6.9: Matroid analysis of the MetS data using an inverse VIF criterion.

The ‘uncorrelated’ factors identified by Shen using PCA are consistent with the 0.79 threshold level of the matroid depiction. The flats have been summarized as ‘insulin resistance’, ‘obesity’, ‘lipid metabolism’ and ‘hypertension’. They follow the results of similar studies analysing comparable risk factors of MetS (Lafortuna et al. 2008; Shah et al. 2006). This construction is also consistent with the four factor solution identified in the VARCLUS analysis (see section 6.6.2). In comparison, the matroid technique has shown to be in relative agreement over the MetS example, but each provides a different perspective on the cluster formation. In the 0.78 threshold an overarching flat of rank-2 is observed

that links BMI and WHR with the insulin resistance flat. This was hypothesised in the second VARCLUS in observing a high correlation between these dependencies and a high loading of the Insulin predictor on the obesity cluster component. Also, in the 0.77 threshold BMI, HDL and Trig are linked with insulin resistance, however WHR is not. This agrees with cluster I of the four cluster VARCLUS solution.

6.7.2 Matroid Analysis in MetS Research

The reason that matroids can provide a different perspective is that they are based on a unique philosophy to identifying dependencies. The uni-dimensional (i.e. rank 1) flats can be considered equivalent to the clusters produced in a traditional hierarchical clustering technique, however there are important differences. With the matroid procedure, the 'reallocation' of misplaced clusters is allowed. This is vital as early results in a hierarchical clustering can typically affect those later in the analysis. Each threshold is constructed individually, meaning that previous results and the order in which the variables are entered will have no influence on the structure at each threshold. The number of flats to extract is not decided prior to the analysis, but the structure presented at a range of dependencies. It is subsequently possible to observe how the dependencies develop at different thresholds and consider which, if any, are in agreement with *prior* evidence as opposed to employing some arbitrary thresholding.

An important feature of the matroid technique is found with the higher ranked subsets extracted at particular thresholds. For instance, at the 0.77 threshold there is a rank-3 flat containing {Ins, PCIns, Glu, PCGlu, BMI, HDL, Trig} that is not identified elsewhere in the clustering. This may indicate an underlying mechanism amongst the variables. The advantage is that we retain the interpretive benefits of 'hard' clustering whilst identifying relationships potentially masked by stronger dependencies at higher thresholds. The flats in this example appear to demonstrate that the risk factors (aside from SBP and DBP) are 'fuzzy' in nature and overhanging dependencies of higher rank could be viewed as evidence of a structure such as MetS. A common intention of dimension reduction is in variable selection and matroid analysis has its benefits in this area. The rank that follows each of the flats gives an indication of the approximate dimensionality of the subset. Therefore, if a factor analysis or PCA were applied to any of these flats, the user would have strong reasoning for the number of factors/components to retain to represent the subset.

6.8 Identifying Consistency across MetS Studies

This section considers a second example of MetS data using a longitudinal study by Maison et al. (2001). The study appeals as an illustrative example because it provides a similar set of variables to the Shen data, but on a different study population. It also stratifies the population by sex and treatment group. This provides an interesting examination of the consistency in the statistical methods amongst study populations. The motivation of this study is to identify a common underlying pathway for CVD and T2-DM. This may be in the form of an insulin resistance component or as proposed in the original paper centered on overall/central obesity. Maison suggests that the definition of MetS and the assumptions made about the structure are easily accepted, whilst the evidence is still unconvincing. Many researchers use this ‘established’ theory as a reason to automatically choose a CFA study based on *prior* hypotheses. The incentive for Maison is to identify a centrally dominant factor (such as obesity), without imposing a *prior* structure on the analysis. For this aim, selecting an EFA approach seems justified by Maison.

The paper considers a sample of 937 subjects aged between 40-65 years in Cambridgeshire, England. The subjects were recruited as part of the “Ely study” (Forouhi et al. 2007; Sandhu et al. 2002). Each took part in a medical examination between 1990-1992 and once again between 1994-1996 (i.e. a 4.5 year follow-up). At both examinations, patients undertook a glucose tolerance test and were measured on 9 risk factors associated with MetS. An additional subset of 471 patients was also considered who did not receive treatment for hypertension and dyslipidemia. The correlation structure displayed in Table 6.14 is for the change in risk factors amongst the female subjects in the study (n = 543 – overall, n = 277 – subgroup).

	BMI	WHR	DBP	SBP	Ins	PG0	PG120	HDL	Trig
BMI	1	0.25	0.15	0.16	0.19	0.21	0.22	-0.07	0.29
WHR	0.21	1	0.06	0.00	0.12	0.03	0.11	-0.03	0.09
DBP	0.18	0.11	1	0.75	0.05	0.08	0.07	-0.10	0.06
SBP	0.19	0.08	0.78	1	0.06	-0.03	0.01	-0.14	0.09
Ins	0.23	0.13	0.08	0.07	1	0.1	0.16	-0.11	0.07
PG0	0.21	0.01	0.07	0.05	0.12	1	0.27	0.04	0.10
PG120	0.21	-0.02	0.01	-0.02	0.09	0.33	1	-0.14	0.16
HDL	-0.09	-0.01	-0.03	-0.05	-0.06	0.01	-0.09	1	-0.11
Trig	0.25	0.05	0.05	0.06	0.16	0.14	0.19	-0.09	1

Table 6.14: Female study population - coefficients for the entire population of the Ely study are displayed in regular typeface, coefficients in bold are for subgroup patients.

PG0 and PG120 represent fasting plasma glucose and plasma glucose 120min respectively, whilst the remaining abbreviations are the same as the Shen example. The procedure employed by Maison uses PCA (although labelled factor analysis) to identify independent ‘factors’ for both sexes. Kaiser’s criterion ($\lambda_j > 1$) is used to identify ‘significant’ factors. In addition, an orthogonal varimax rotation is used and any loadings > 0.25 are labelled ‘significant’. It is possible to suggest why these methodological decisions were made. The main finding of this study is that BMI is a central component of the MetS structure, demonstrated by the ‘significant’ loadings. However, this is only observed by a loading of 0.28 on factor I for the male cohort (see original paper for factor loadings). It seems unlikely that the 0.25 cut point has been chosen by chance from this finding. These results appear instead as a reflection of the author’s hypothesis regarding the role of BMI. The methodological decisions such as the use of PCA, Kaiser’s criterion and orthogonal rotation experience no justification in the text and so we are left to assume that they are either default or unreliable decisions. They appear neither suited to the original hypothesis, nor an attempt to identify a *simple* solution. A CFA model based on a similar population could provide strength to the findings; however an EFA with justified methodological decisions could also uncover an alternative structure with a greater biological relevance.

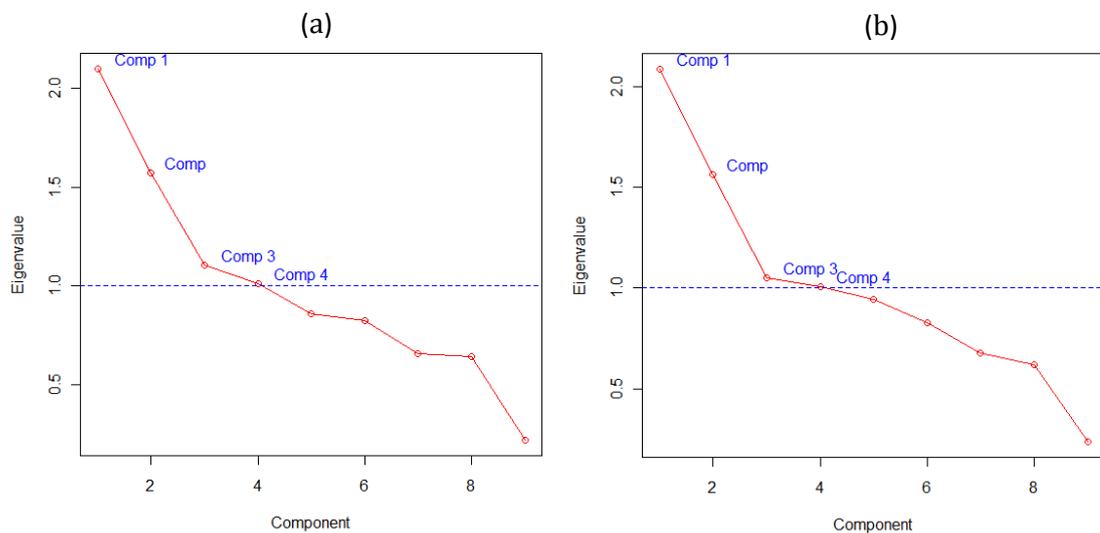


Figure 6.10: Scree plots of (a) the study population and (b) the untreated subgroup.

Four components were identified using Kaiser’s criterion and they explain 64.4% and 63.4% of the overall variation respectively. However, the scree plot demonstrates a clear elbow in the graph at three components for both samples in Figure 6.10. This solution would instead explain 53.2% and 52.2% of the variation respectively.

	Study Population				Untreated Subgroup			
	I	II	III	IV	I	II	III	IV
SBP	0.94	0	0.06	-0.04	0.93	0.05	-0.04	-0.09
DBP	0.94	0.03	0.09	0.00	0.93	0.07	0.07	-0.01
PG0	0.07	0.77	0.07	0.18	0.03	0.05	0.84	0.16
PG120	-0.04	0.77	0.01	-0.14	-0.02	0.15	0.65	-0.33
HDL	-0.03	0.03	-0.02	0.93	-0.10	0.03	0.07	0.89
Trig	0.00	0.38	0.34*	-0.33	0.06	0.42	0.21	-0.25
Ins	0.00	0.13	0.61	-0.10	-0.01	0.32	0.17	-0.39
BMI	0.19	0.35	0.59	-0.12	0.17	0.67	0.31	-0.06
WHR	0.05	-0.22	0.75	0.15	-0.04	0.81	-0.17	0.07

Table 6.15: Factor loadings for females using the original methodological decisions.

Maison identifies the four components of the full sample as (1) hypertension { SBP, DBP }, (2) lipid metabolism { HDL, Trig }, (3) glucose metabolism { PG0, PG120, Trig, BMI } and (4) a component including the risk factors { BMI, WHR, Ins, Trig*} (*Trig was found non-significant in reported loadings). BMI (i.e. the focus of the study) loads ‘significantly’ on the glucose metabolism component and another relating to Ins, Trig and WHR. In the untreated subgroup, similar components for blood pressure and the additional component are found. However, Trig is found insignificant in the glucose metabolism factor, whilst Ins and PG120 produce significant loadings on the lipid metabolism factor. Also, Trig is found only just significant using the original 0.25 threshold.

Maison’s solution is complex as risk factors load significantly on multiple factors in both analyses. The component structure also appears substantially different between the full sample and the sub-sample. Is this a biological difference, or simply a statistical artefact borne from the methodological decisions employed in Maison’s study? The notion that BMI is the central component in the mechanism is difficult to justify when a variable such as Trig experiences similarly high loadings across multiple factors. This analysis is repeated using a ‘true’ EFA, VARCLUS and matroid methodology to examine the consistency across the subsets and compare to the biological interpretation of Maison. The analysis will stay true to Maison’s original motivation to generate a structure free from *prior* assumptions about the number and structure of the factors. The common factor model of a PAF extraction is applied.

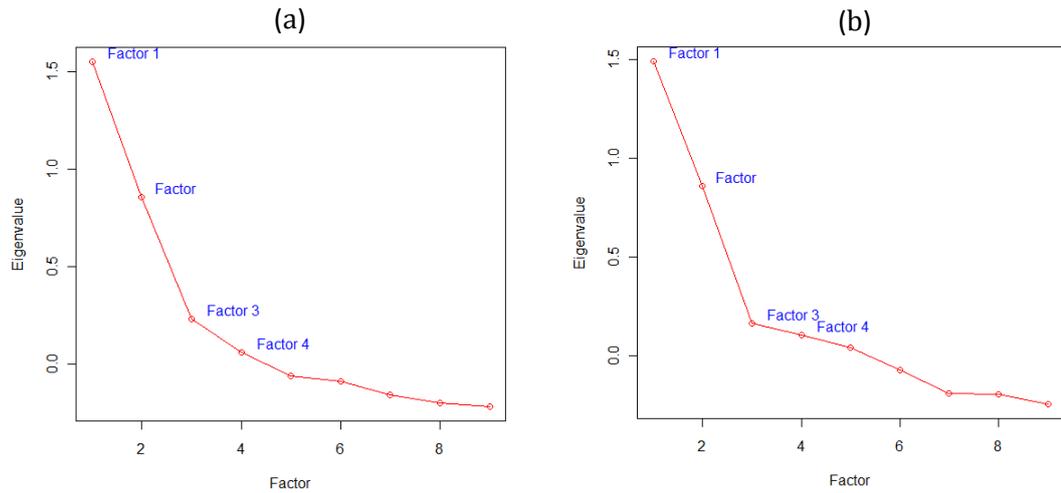


Figure 6.11: Scree plot for the eigenvalues of the factors in each group.

	Study Population				Untreated Subgroup			
	I	II	III	h ²	I	II	III	h ²
SBP	0.84	-0.01	0.00	0.70	0.82	0.05	-0.08	0.69
DBP	0.83	0.01	0.00	0.70	0.82	-0.05	0.09	0.66
PG0	0.04	0.49	-0.04	0.23	0.00	-0.09	0.51	0.21
PG120	-0.04	0.53	-0.04	0.26	-0.02	0.16	0.35	0.22
HDL	0.00	-0.07	-0.11	0.02	-0.08	-0.29	0.12	0.07
Trig	-0.03	0.25	0.22	0.16	0.00	0.34	0.07	0.15
Ins	-0.03	0.08	0.34	0.14	-0.01	0.27	0.08	0.10
BMI	0.06	0.22	0.39	0.30	0.06	0.40	0.19	0.31
WHR	0.00	-0.14	0.39	0.12	-0.06	0.35	0.00	0.11
VAR	1.48	0.87	0.87	2.64	1.41	0.97	0.80	2.52

Table 6.16: PAF extraction with promax rotation (loadings > .3 highlighted in bold).

	Factor I	Factor II	Factor I	Factor II
Factor I				
Factor II	0.10		0.26	
Factor III	0.34	0.47	0.10	0.60

Table 6.17: Inter-factor correlations.

The scree plots for both groups identify an elbow at 3 factors and so this is the initial decision for factor retention (see Figure 6.11). Similar to BMI, Trig loads highly on factor II and factor III, but with a higher loading on the former. Employing the PAF with oblique rotation, less factors and a higher threshold has produced a *simpler* structure than Maison's results, with factors clearly marked as Hypertension (Factor I), Glucose (Factor II) and Obesity/Insulin Resistance (Factor III) in the full sample. In the subset, the only change to this structure is the drop in significance of Ins in the obesity factor. Instead, HDL almost reaches significance on this factor, but still has the lowest communality estimate of all the risk factors. The communality on HDL indicates that it is not explained well by the factors in either model. This suggests that it may well be independent of the other risk factors.

The VARCLUS analysis is next considered to try to uncover the role of Trig, HDL and BMI risk factors as part of an underlying mechanism. The decision of how many clusters to extract is aided by the previous EFA in noting the variance explained by three factors and the potential independence of the HDL risk factor. Therefore, a four cluster solution is considered.

Cluster	Members	Variance Explained	Proportion	2 nd Eigenvalue
Cluster I	2	1.78	0.89	0.22
Cluster II	2	1.33	0.67	0.67
Cluster III	4	1.53	0.38	0.95
Cluster IV	1	1	1	
Tot Variance:		5.64	Proportion: 62.67%	

Table 6.18: A summary of the cluster components from a four cluster solution.

Cluster	Variable	R^2_{own}	R^2_{nearest}	$1 - (R^2_{\text{own}}/R^2_{\text{nearest}})$
Cluster I	DBP	0.89	0.03	0.11
	SBP	0.89	0.03	0.11
Cluster II	PG0	0.67	0.04	0.35
	PG120	0.67	0.04	0.35
Cluster III	BMI	0.55	0.07	0.48
	WHR	0.25	0.01	0.76
	Ins	0.39	0.02	0.62
Cluster III	Trig	0.34	0.04	0.69
	Cluster IV	HDL	1	0.01

Table 6.19: Cluster structure with R^2 measures to indicate the stability of each cluster.

HDL is independent in cluster IV and so automatically has a perfect correlation with itself. The useful statistic is that the correlation is low with any other cluster and so evidence for the independence of this covariate appears strong (contrary to Maison's PCA analysis). SBP and DBP are highly correlated within their own cluster and have small correlations with any other. Therefore, the overall R_x^2 ratio is low. Clusters II and III appear less well defined. In the original PCA, { BMI, WHR, Ins, Trig } formed an individual factor whilst each loaded significantly along with glucose. BMI and Trig in particular have a high correlation with their nearest cluster. However, BMI is explained well in its own cluster and also has a high correlation with the nearest cluster.

Variable	Cluster I	Cluster II	Cluster III	Cluster IV
BMI	0.20	0.26	0.74	-0.09
WHR	0.10	0.00	0.50	-0.01
DBP	0.94	0.05	0.17	-0.03
SBP	0.94	0.02	0.17	-0.05
Ins	0.08	0.13	0.62	-0.06
PG0	0.06	0.82	0.21	0.01
PG120	0.00	0.82	0.20	-0.09
HDL	-0.04	-0.05	-0.11	1.00
Trig	0.06	0.20	0.58	-0.09

Table 6.20: Loadings on each cluster.

	Cluster I	Cluster II	Cluster III
Cluster I			
Cluster II	0.04		
Cluster III	0.18	0.25	
Cluster IV	-0.04	-0.05	-0.11

Table 6.21: Inter-correlations of the clusters.

Notice that BMI also loads highly on the blood pressure cluster component (cluster I) and again suggests a possible overhanging relation. Trig loads highly on both cluster II and cluster III whilst not being explained well by either. The role of Trig appears weaker than BMI but of a similar nature. The 'small' loadings on cluster IV add further strength to the independence of HDL. As expected, factors II and III are highly correlated (particularly as they are essentially defined as independent factors - i.e. 0.25). Therefore, it appears that both BMI and Trig play a role in glucose and insulin factors. The analysis using the matroid technique is illustrated in Figure 6.12.

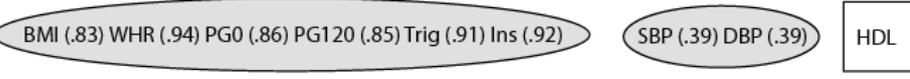
Rank	Threshold
1	.99 
1	.97-.98 
1	.96 
1	.95 
1	.94 
1	.93 
1	.9 

Figure 6.12: Matroid analysis using minimum VIF criteria for the complete sample.

The matroid analysis highlights the independence of the HDL risk factor and the blood pressure factor until high thresholds are reached. The remaining risk factors display a similar pattern to what has been previously observed. The glucose flat is independent until the 0.96 threshold, whilst BMI and Trig form a separate flat with Ins and finally WHR forming. WHR displayed the lowest R_{own}^2 in the VARCLUS analysis and this result is

demonstrated in the matroid analysis. It appears that the four factor structure is best defined from the exploratory analysis, with BMI overhanging separate glucose and insulin factors in a hierarchical model. The notion of BMI as a central component to all factors appeared weak from the initial PCA analysis, however the EFA, VARCLUS and matroid analyses have given this hypothesis a greater strength. The untreated members of the female population are next considered. The low communality of Ins, WHR and HDL from the EFA in Table 6.16 may suggest a 5 or 6 cluster structure.

(a)

Cluster	Members	Variance Explained	Proportion	2 nd Eigenvalue
I	3	1.43	0.48	0.91
II	2	1.75	0.88	0.25
III	2	1.27	0.64	0.73
IV	1	1	1	
V	1	1	1	
Tot Variance:		6.45	Proportion: 72.67%	

(b)

Cluster	Variable	R ² _{own}	R ² _{nearest}	1 - (R ² _{own} /R ² _{nearest})
I	BMI	0.63	0.07	0.4
	WHR	0.36	0.01	0.65
	Trig	0.44	0.03	0.58
II	DBP	0.88	0.02	0.13
	SBP	0.88	0.02	0.13
III	PG0	0.64	0.03	0.38
	PG120	0.64	0.06	0.39
IV	HDL	1	0.02	0
V	Ins	1	0.04	0

(c)

Variable	Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V
BMI	0.79	0.17	0.27	-0.07	0.19
WHR	0.60	0.03	0.09	-0.03	0.12
DBP	0.14	0.94	0.09	-0.10	0.05
SBP	0.13	0.94	-0.01	-0.14	0.06
Ins	0.19	0.06	0.16	-0.11	1
PG0	0.18	0.03	0.80	0.04	0.10
PG120	0.24	0.04	0.80	-0.14	0.16
HDL	-0.10	-0.13	-0.06	1	-0.11
Trig	0.66	0.08	0.16	-0.11	0.07

Table 6.22: Five factors - (a) cluster summary, (b) R² measures and (c) loadings.

	Cluster I	Cluster II	Cluster III	Cluster IV
Cluster I				
Cluster II	0.14			
Cluster III	0.26	0.04		
Cluster IV	-0.10	-0.13	-0.06	
Cluster V	0.19	0.06	0.16	-0.11

Table 6.23: Inter-correlations of the clusters from the five cluster solution.

The 2nd eigenvalue on cluster I suggests that the cluster is not stable. There is also a decrease in the stability of the glucose cluster. The cluster summary illustrates that PG120 also has a close correlation with another cluster. Ins is clustered independently of the remaining risk factors, however it has a relatively high $R^2_{nearest}$, suggesting that it may not be independent. The cluster structure in general is similar to the previous VARCLUS on the full sample. The lower stability of the glucose cluster is due to a high correlation of PG120 with the BMI cluster. However, this is still well formed with a low $R^2_{overall}$. BMI has high loadings on clusters I and III, whilst also to a lesser extent clusters II and V. This is further demonstrated by the inter-correlations of cluster I with the remaining clusters. Whilst insulin forms its own cluster it still has relatively high correlations with clusters I and III. HDL once again appears independent.

Rank	Threshold
1	.99 
1	.98 
1	.97 
1	.96 

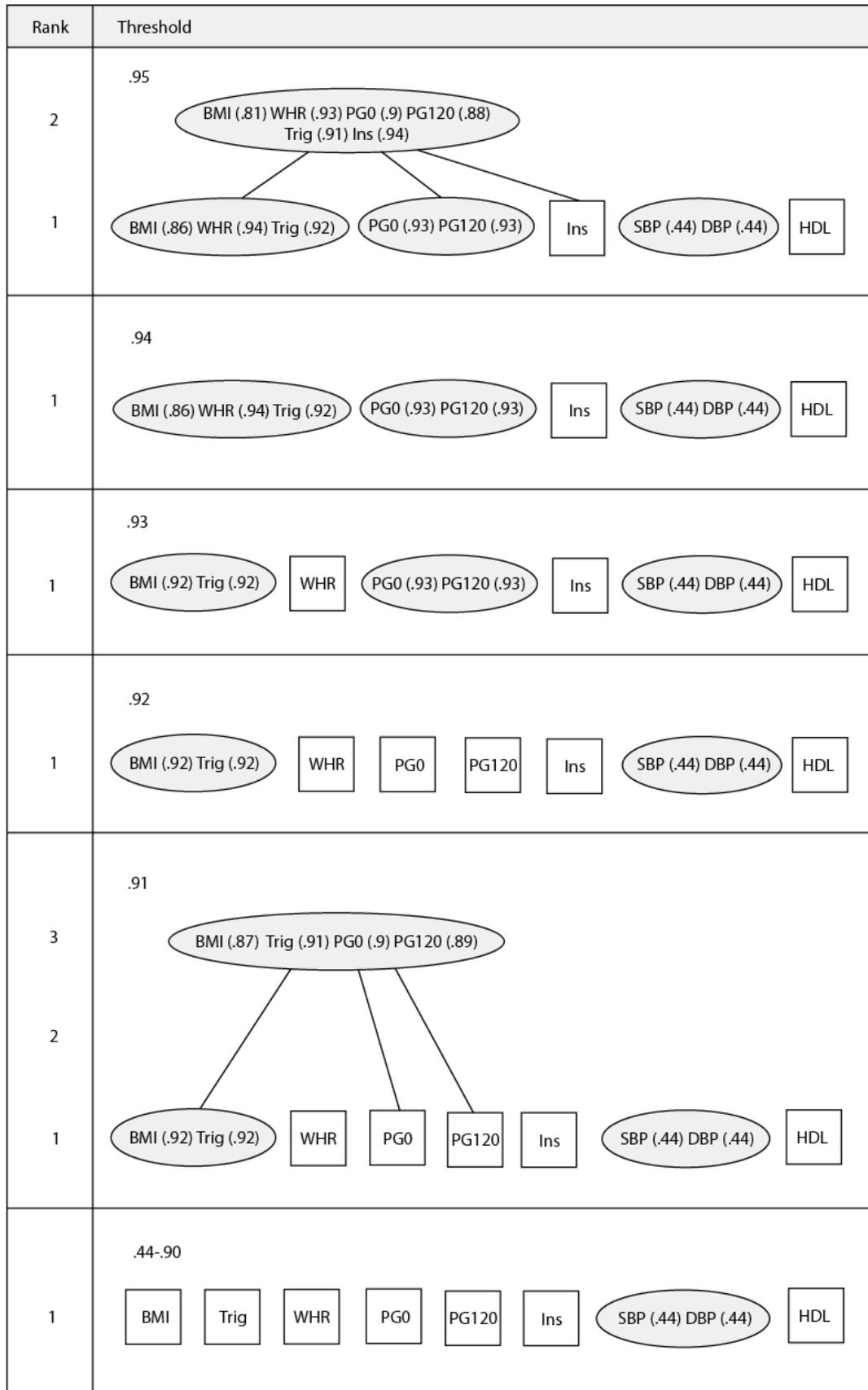


Figure 6.13: Matroid analysis of the female subset cohort

The matroid analysis is in agreement with the findings from EFA and VARCLUS. The HDL risk factor is independent until the very highest threshold is reached. The hypertension flat is once again independent, until the 0.98 threshold. Seemingly this flat joins a larger dependency due to the association with BMI. Ins remains independent until 0.97, in which a higher ranked flat demonstrates a link to glucose and BMI/Trig flats.

Discussion

Recall the structure from Maison's paper self-labelled (1) hypertension { SBP, DBP }, (2) lipid metabolism { HDL, Trig }, (3) glucose metabolism { PG0, PG120, Trig, BMI } and (4) a component including { BMI, WHR, Ins, Trig }. It appeared as though the model changed substantially between the complete data and the untreated subgroup, with Trig becoming insignificant in the glucose metabolism factor, whilst Ins and PG120 became significant on the lipid metabolism factor. Also, Trig was only just significant on lipids by the 0.25 threshold employed by Maison. The analyses produced from EFA, VARCLUS and matroid methods demonstrate an alternative structure. HDL appears independent in women for both samples. There is some effect on blood pressure between groups, but not to change its overall position in the construct, just the strength of the association with general obesity. Each methodology has identified a separate glucose factor, with a close association to obesity. However, there also appears to be a pathway through Trig. In the complete sample WHR appears uncomfortable as part of an Ins/Trig clustering and in the untreated sample Ins appears independent. This structure is reflected in all analyses and the original correlations. The lipid lowering drugs may have had some effect on this part of the structure. An association could occur more naturally between physical exercise and triglycerides in the untreated group.

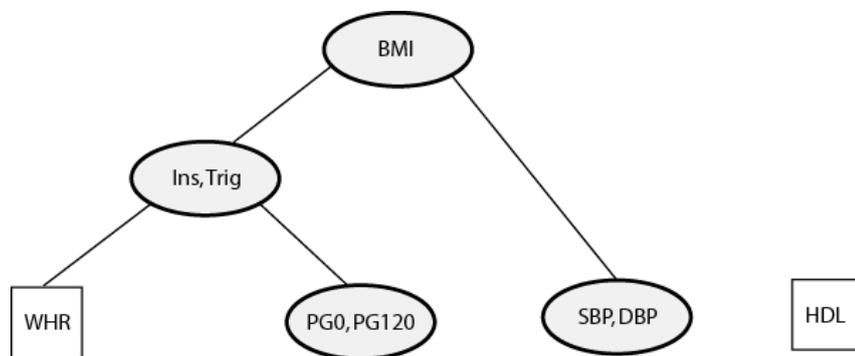


Figure 6.14: Potential factor structure from complete data

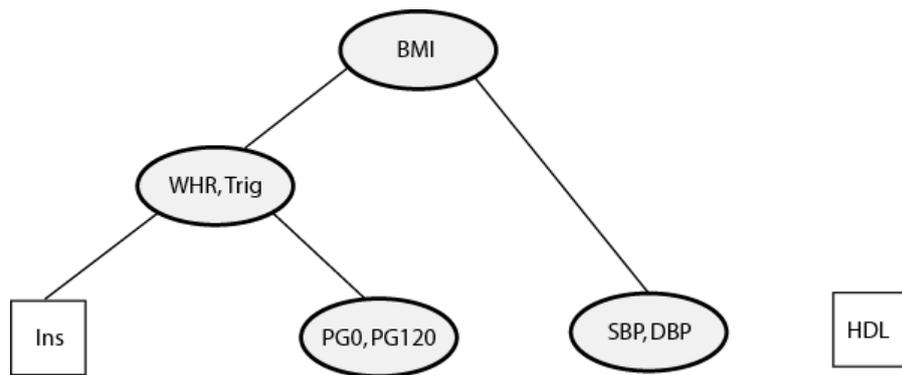


Figure 6.15: Potential factor structure from untreated sample.

The analyses employed in this example have reached a similar conclusion to the Maison paper in labelling BMI as a central construct to CVD and T2-DM. This example has demonstrated that whilst the ‘simple’ default options in performing an EFA may appear to give an easily interpretable structure, they are also likely to complicate the analysis and interpretation. Choosing suitable options for the MetS hypothesis is likely to produce the consistency required to identify subtle differences in the structure amongst study populations. Whilst the paper chose to use Kaiser’s criterion to automatically select four components, a combination of judgement in the scree plot, analyzing which factors load together in the EFA and identifying independent risk factors lead us to selecting a three factor model. This was followed by a four cluster VARCLUS to test the presence of the three factors in addition to the independence of the HDL risk factor. The agreement between these methods increases the confidence in the overall model.

The difficulty with MetS is that the structure is likely to be hierarchical in nature (at least from a statistical perspective). A PCA with ‘default’ methodological decisions is unsuitable to match the complexity of the MetS construct. It may be that a hierarchical or second order factor analysis could provide an appropriate tool for MetS analysis (with the intention to separate ‘broad’ factors from ‘narrow’ factors). The promax rotation in an EFA, correlated cluster components and matroid flats represent methods to identify oblique factors of different weight. However, it is important to remember the context in which these methods are to be used. A likely reason that an oblique EFA or a hierarchical factor analysis are rarely used in practice is due to the statistical complexity. Therefore, it is important to remain mindful of this when promoting methodology such as VARCLUS and matroid approaches to retain a simpler interpretation, whilst raising the consistency and appropriateness of the decision making in MetS study.

6.9 Conclusions

An EFA of a MetS dataset can generate an enormous range of hypotheses and interpretations. It is clear that for biologically meaningful structures to be obtained, the methodological decisions should be application specific and justified by the user. The use of EFA requires a great number of decisions that can have a substantial influence on the results obtained in an explorative study. In many practical applications, these decisions are made with little consideration of the potential impact on the results and conclusions formed. When poor decision making is repeated over a large number of studies the full effects of these decisions are observed through the range of potential model structures suggested. The methodological variation can cloud any deviation due to population differences and the lack of reasoning provided for these decisions in many studies does little to encourage the reader of the usefulness of the study findings.

A considered approach to selecting and applying methodology can provide consistent and meaningful results, but many of these options increase the complexity of the procedure substantially. To some researchers this is an acceptable compromise to obtain clinically relevant results. However, default software options appear to give weight to simpler methodologies (on the surface), but these may not be appropriate to an application such as MetS. Alternative methodology must be provided and be suited to the context under examination. Two such techniques have been presented in the VARCLUS and matroid approaches. The aspect that has been focussed on is the use of visual image and 'hard' clustering to simplify interpretation. These proposed methods can be used in tandem with EFA to unravel the latent structure of MetS.

The criteria for MetS, such as those proposed by the WHO and ATP III, have been developed to diagnose subjects, whereas the methods presented in this chapter are not intended to form such criteria. However, the continued use of explorative techniques is of great importance. If methods such as PCA or EFA fail to reveal an underlying latent structure, the very existence of MetS becomes questionable. The intention of developing methodology such as the VARCLUS and matroid approaches is primarily to encourage consistency and reproducibility across MetS studies. It is not possible to judge from the explorative methods which will provide the 'correct' structure, and there may never be such a structure. Exploratory approaches should instead be valued on which yield the more useful results in terms of understanding the complex inter-relationships amongst the metabolic risk factors.

7. Identifying the Critical Phase of Growth in the Lifecourse

The aims of this project have focussed on collinearity amongst covariates in applied clinical and epidemiological research. In chapter 7 a unique case is considered in which the predictors present an exact linear dependency - labelled 'perfect collinearity' (see section 3.1). An example is considered from lifecourse epidemiology. The predictors that are considered in this study – birth size, growth and current size – are perfectly collinear (i.e. birth size + growth = current size). Whilst the individual effects of these covariates on various health outcomes have been studied extensively, entering all of the covariates simultaneously into a regression model presents a challenge to the researcher in both computation and interpretation of the model coefficients. The challenge in the lifecourse study is to identify a critical phase of growth for health outcomes in later life. This could potentially allow for earlier interventions prior to the onset of chronic disease.

The dependence amongst covariates in the lifecourse model means that it suffers from a well known identification problem - the design matrix is ill-defined. This prevents the inversion of the $\mathbf{X}^T\mathbf{X}$ matrix by a regular inverse which prevents the computation of OLS parameter estimates for the full set of predictors. To obtain coefficient estimates for each of the predictors, a constraint is required in the estimation of $\boldsymbol{\beta}$. Statistical software often proceeds by removing one of the predictors from the regression model. This is effectively constraining the regression coefficient of the removed covariate to zero. The choice of such a constraint will naturally impact on the usefulness of the estimation in practice. Shrinkage methods such as PCR and PLS can introduce a bias into the estimation that allows for the inversion of the $\mathbf{X}^T\mathbf{X}$ matrix (or even remain unbiased if the constraint employed is correctly specified). However, an investigation of the benefits of these estimates in the perfect collinearity case has not been fully explored. In this chapter a conceptual link is provided between these methods using analytical and geometrical methods to aid with the interpretation of the results in the lifecourse problem.

7.1 Example Application

The example for this chapter has been chosen from a prospective study of 3,080 subjects based in the Philippines. The invited participants included all the pregnant residents of 33 randomly selected communities in Metro Cebu during the year 1983. Height and weight measures were collected for each of the children at six intervals from birth to 19yrs – (at 0yrs, 1yrs, 2yrs, 8yrs, 15yrs, 19yrs). In addition, measurements of average systolic blood pressure (SBP) and average diastolic blood pressure (DBP) were taken for each of the children involved in the study (in mm/Hg). These outcome measures are based on the average of three measurements taken at 19yrs. The focus will be on SBP for each analysis in this chapter as an indicator of hypertension. The data was attained from the website of the University of North Carolina Population Centre (<http://www.cpc.unc.edu/projects/cebu/datasets.html>) and further detail of the study can be found on this website and in selected papers (Adair et al. 2011; Adair and Popkin 2001; Tudor-Locke et al. 2003).

The data used in this chapter comprises of the 960 boys with complete measurements for weight, height and blood pressure. To demonstrate different approaches to the perfect collinearity problem the discussion will begin with a 3 predictor example – i.e. BMI at birth (BMI_0), current BMI (BMI_{19}) and change in BMI (BMI_{0-19}) (where $BMI = \text{Weight}/\text{Height}^2$). The correlation matrix for the variables in this study is illustrated in Table 7.1,

	BMI_0	BMI_{0-19}	BMI_{19}
BMI_0			
BMI_{0-19}	-0.32		
BMI_{19}	0.11	0.91	
SBP	0.01	0.33	0.35

Table 7.1: Pearson correlations of the BMI predictors and SBP in Metro Cebu.

The analysis is extended in section 7.8 with increased information through including multiple body size measurements over the lifecourse (i.e. six interval measures). The BMI change will be defined by the difference between neighbouring BMI values (i.e. $BMI_{0-1} = BMI_1 - BMI_0$). Note that regardless of how the data is partitioned, the perfect collinearity problem remains. The common motivation is to identify a critical phase of growth for the subjects, determining lasting effects on hypertension that are experienced in their late teenage years – known as ‘biological programming’ (Ben-Shlomo and Kuh 2002).

7.2 The Foetal and Developmental Origins of Disease

The ‘foetal origins of adult disease’ (FOAD) or ‘Barker’s hypothesis’ states that an individual’s later life health outcomes may have been influenced much earlier in the subjects’ life (Barker 1992). In particular, a low birth weight due to a factor such as malnutrition prior to birth has been suggested to have an inverse relationship to many chronic diseases in later life (Henriksen and Clausen 2002). It has been theorized that the trajectory of growth is defined in the very early stages to benefit the child in its immediate early years. However, this trajectory could have adverse effects to health in later life (Gluckman et al. 2005). This theory has been met with great debate in epidemiology over the past two decades. Recent evidence has been presented to suggest that a low birth weight, coupled with a rapid compensatory growth, may increase the risk of adverse health outcomes such as CVD, hypertension and T2-DM. This theory has been labelled the ‘developmental origins of health and disease’ (DOHaD) hypothesis (Barker 2004). Whichever theory the researcher chooses to believe, the evidence is generally accepted that nutrition at some critical period in the individual’s lifecourse can dictate the onset of chronic disease in later life (Adair and Prentice 2004).

7.3 Adjustment for Lifecourse Variables

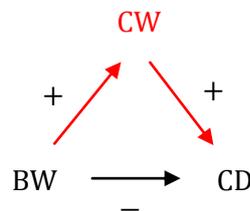


Figure 7.1: A path diagram of a hypothesized system in the lifecourse model.

A number of studies have reported an inverse relationship between birth weight (BW) and chronic diseases (CD) such as CVD (Barker et al. 1993; Rich-Edwards et al. 1997), non-insulin dependent diabetes (Lindsay et al. 2000; Lithell et al. 1996) and hypertension (Law and Shiell 1996). The unadjusted association is usually non-significant. Upon adjusting for current weight (CW), the association between BW and CD enhances the negative coefficient, which often becomes significant (e.g. see Figure 3.5b). This has been suggested as evidence for the DOHaD hypothesis. However, the statistical practice has experienced much criticism (Huxley et al. 2002; Lucas et al. 1999).

The variable CW is usually reported to have a positive association with both BW and CD. The inclusion of this predictor in the model is often justified by labelling CW as a confounder and it is seen to be ‘dampening’ the unadjusted association between BW and CD (see section 3.2.2). However, considering the notion of causality, it is questionable whether CW is a ‘true’ confounder. Adopting the hypothetical model in Figure 7.1, CW lies on the causal pathway, which defines the covariate instead as a mediator. Two pathways are observed from exposure to disease; A ‘direct’ path from BW to CD and an ‘indirect’ path from BW to CD (with CW on the causal pathway). From the perspective of the path model the analyst would seem to have two options to proceed. Either consider the total effect between BW and CD, or the association between CW and CD, with BW included as a confounder. To adjust for CW would bring about a suppression effect. The coefficient of BW would become more negative and the positive coefficient for CW would be enhanced (as observed by a number of studies). The adjustment for a ‘false’ confounder is a primary criticism of the statistical evidence of the FOAD hypothesis.

Change in weight (WC) now takes on a potential significance. It has been thought that the inverse association between BW and CD could signify the importance of WC over BW (primarily due to the non-significant correlation between BW and CD). A small BW coupled with a rapid compensatory “catch-up” growth could be a strong predictor of later life chronic disease (Ong et al. 2000). The statistical evidence to support this hypothesis is generally no stronger than that already discussed for the model in Figure 7.1.

Variable	Univariable	Bivariable Regression	Bivariable	Bivariable
	Regression	I	Regression II	Regression III
	Coefficient (95% CI)	Coefficient (95% CI)	Coefficient (95% CI)	Coefficient (95% CI)
Unstandardized				
BMI ₀	0.08 (-0.47 to 0.63)	1.09 (0.55 to 1.63)	-0.26 (-0.77 to 0.26)	
BMI ₀₋₁₉	1.2 (0.98 to 1.42)	1.35 (1.12 to 1.58)		0.26 (-0.26 to 0.77)
BMI ₁₉	1.33 (1.1 to 1.56)		1.35 (1.12 to 1.58)	1.09 (0.55 to 1.63)
Standardized				
BMI ₀	0.1 (-0.57 to 0.77)	1.34 (0.67 to 2.01)	-0.32 (-0.95 to 0.32)	
BMI ₀₋₁₉	3.44 (2.81 to 4.08)	3.87 (3.21 to 4.54)		0.74 (-0.75 to 2.22)
BMI ₁₉	3.66 (3.03 to 4.29)		3.69 (3.06 to 4.33)	2.99 (1.5 to 4.47)
R², %		11.85	11.85	11.85

Table 7.2: Unstandardized and standardized OLS coefficients of the Cebu data.

For the univariable estimates, BMI_0 had no association with SBP, whilst the coefficients for BMI_{0-19} and BMI_{19} had a positive association with SBP. When BMI_{0-19} is entered into the bivariable model along with BMI_0 , the coefficient for BMI_0 remains positive and becomes significant. In comparison, entering BMI_{19} along with BMI_0 reverses the sign on BMI_0 but remains non-significant. These relationships are linked and can be explained using a simple relationship inherent in the data.

$$y_{SBP} = b_{BMI_0} X_{BMI_0} + b_{BMI_{19}} X_{BMI_{19}} \quad (7.1)$$

The coefficients rest on the intrinsic relationship defined by the lifecourse problem – i.e. $BMI_0 + BMI_{0-19} = BMI_{19}$. The estimates from any of the three combinations of BMI_0 , BMI_{0-19} and BMI_{19} can be calculated using this model as follows,

$$\begin{aligned} y_{SBP} &= b_{BMI_0} X_{BMI_0} + b_{BMI_{19}} (X_{BMI_0} + X_{BMI_{0-19}}) \\ &= (b_{BMI_0} + b_{BMI_{19}}) X_{BMI_0} + b_{BMI_{19}} X_{BMI_{0-19}} \end{aligned} \quad (7.2)$$

$$\begin{aligned} y_{SBP} &= b_{BMI_0} (X_{BMI_{19}} - X_{BMI_{0-19}}) + b_{BMI_{19}} X_{BMI_{19}} \\ &= (b_{BMI_0} + b_{BMI_{19}}) X_{BMI_{19}} - b_{BMI_0} X_{BMI_{0-19}} \end{aligned} \quad (7.3)$$

Regardless of the model the researcher chooses to analyze, the estimates are all intrinsically linked by the coefficients and each have the same explained variance (i.e. R_y^2). One model would appear equally plausible to any other from a statistical perspective. The differences in estimated coefficients make the epidemiological interpretation of the coefficients from any one model difficult to justify.

The statistical problem is defined by the causal model adopted. If $X_{BMI_{19}}$ is not believed to lie on the causal pathway, then controlling for it in the model may be justified. This is reflected in the literature, with the proposal of a range of hypothetical models. However, assuming the causal model illustrated in Figure 7.1, care should be taken regarding which coefficients to interpret. Ordinarily a researcher would look to include all three covariates in a multiple regression model to analyze the ‘importance’ of each phase of growth. However, this introduces new problems in regression analysis. The reason that such relationships in eqn(7.2) and eqn(7.3) can be found from eqn(7.1) is that any two covariates can be used to find the remaining one. It becomes impossible to partition the unique contribution of each covariate to explained variance using least squares estimation. This is labelled the identification problem.

7.4 The identification Problem

The restriction to identifying a critical phase of growth in the lifecourse is in the failure of OLS to provide estimates when all of the predictors are entered simultaneously. The presence of perfectly collinear covariates violates one of the classic assumptions of OLS regression and prevents computation of the estimated coefficients (see **A3** in section 2.2.2). The problem lies with the inversion of a singular $\mathbf{X}^T\mathbf{X}$ matrix in OLS computation. In general, the identification problem ensures that there will exist more than one solution to the estimated coefficients that produce an equally high explained variance (i.e. R_y^2). The challenge is to obtain a justifiable solution out of those available.

7.4.1 SVD of the Ill-Defined Matrix

To demonstrate the statistical problem of the ill-defined matrix, it is useful once again to consider the SVD formulation presented in section 3.3.2. The $\mathbf{X}^T\mathbf{X}$ matrix is constructed of an orthogonal matrix \mathbf{V} containing the columns of eigenvectors \mathbf{v}_j and a diagonal matrix $\mathbf{\Lambda}$ of eigenvalues λ_j .

$$\text{SVD}(\mathbf{X}^T\mathbf{X}) = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \quad (7.4)$$

As the matrix \mathbf{V} is orthonormal (i.e. $\mathbf{V}^T = \mathbf{V}^{-1}$), the inverse is calculated as follows,

$$\begin{aligned} (\mathbf{X}^T\mathbf{X})^{-1} &= (\mathbf{V}^T)^{-1}\mathbf{\Lambda}^{-1}(\mathbf{V})^{-1} \\ &= \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T \end{aligned} \quad (7.5)$$

The OLS coefficient \mathbf{b}_{OLS} can be represented in the following form,

$$\mathbf{b}_{\text{OLS}} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T\mathbf{X}^T\mathbf{y} \quad (7.6)$$

When perfectly collinear predictors are entered into the model, the $\mathbf{X}^T\mathbf{X}$ matrix is rank deficient. This will result in at least one of the eigenvalues in the diagonal of the matrix $\mathbf{\Lambda}$ to be zero. Principal components maximize variance, therefore the zero eigenvalue illustrates a redundant dimension (i.e. a null space of \mathbf{X}). The matrix $\mathbf{\Lambda}$ is defined as singular, which prevents the inversion of $\mathbf{\Lambda}$ by regular full rank regression methods. Therefore, at least one covariate must be removed from the model to attain an estimate from OLS. A useful illustration of this problem can be provided by the vector geometry.

7.4.2 Vector Geometry of the Lifecourse

Vector geometry can provide a conceptually appealing illustration of the lifecourse problem. The relationship between the variables in the lifecourse study is uniquely defined by eqn(7.7),

$$\mathbf{X}_{\text{BMI}_0} + \mathbf{X}_{\text{BMI}_{0-19}} - \mathbf{X}_{\text{BMI}_{19}} = 0 \quad (7.7)$$

The Metro Cebu data demonstrates the identification problem in that each of the three predictors lies in a common 2-dimensional regression plane,

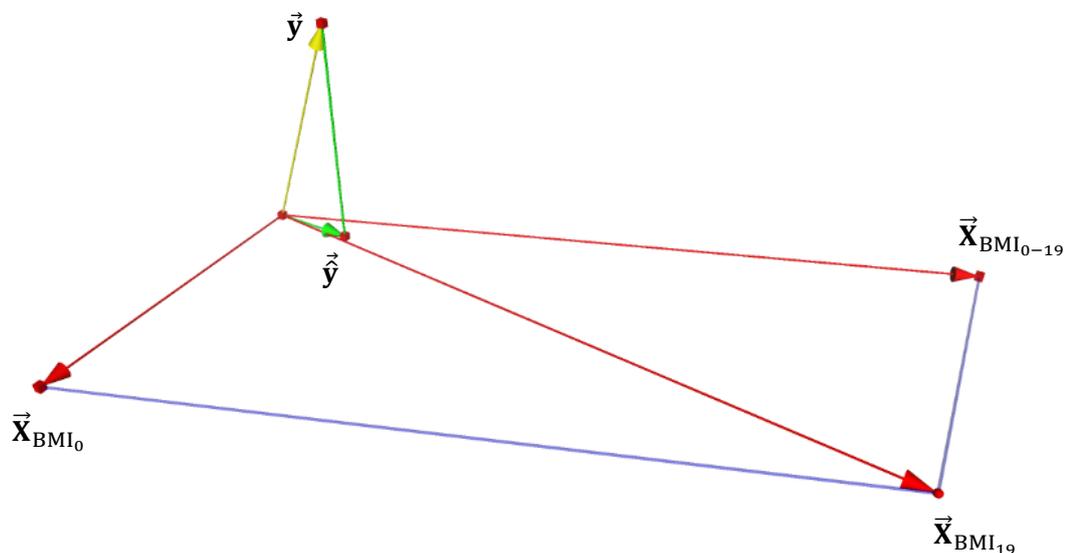


Figure 7.2: Perfectly collinear standardized predictors in the Cebu study.

To obtain the partial regression coefficient for BMI_0 , the OLS solution would project \hat{y} along the plane formed by BMI_{0-19} and BMI_{19} (see section 4.1.3). However, BMI_0 already lies on this plane and so the projection would not produce a unique solution. Therefore, the effect of BMI_0 is considered to be completely confounded by BMI_{0-19} and BMI_{19} . There are infinitely many solutions to the identification problem (i.e. parallel planes). Therefore, some form of constraint must be placed on the estimation to produce a unique solution. The challenge is to choose the solution with the most appropriate justification to the hypothesis. To add further confusion, a great number of solutions have been proposed (out of the infinitely many) for similar problems in the epidemiology, sociology and bioinformatics literature. It is not always clear what assumptions these techniques are making, with some authors choosing not to justify them. This ultimately defines the utility of the method in application.

7.4.3 Should the Variables be Standardized?

Before considering potential ‘solutions’ to the identification problem, it is important to discuss the decision of whether to enter standardized or unstandardized predictors into the regression. In the Metro Cebu example (and typically in studies considering body weight in general), BMI_0 (1.63) has a smaller variance than BMI_{19} (2.74) and BMI_{0-19} (2.88). Many researchers choose to scale the covariates to unit variance to prevent penalizing the estimated coefficients. The vector geometry of the unstandardized predictors was demonstrated in Figure 7.2, with the standardized covariates presented in Figure 7.3.

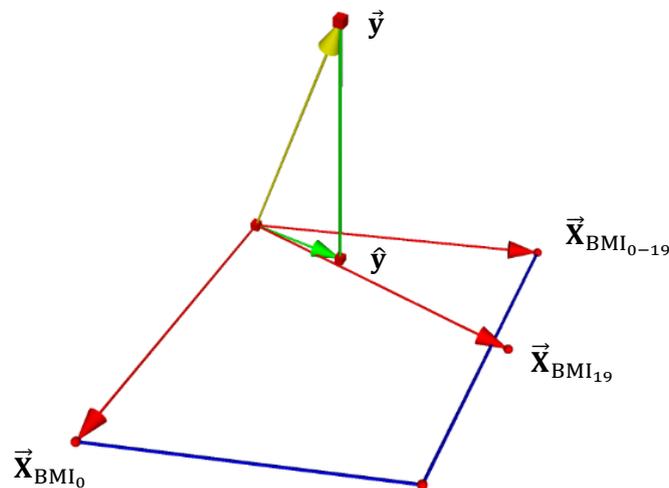


Figure 7.3: Vector Geometry of the Standardized Predictors.

When the variables are scaled, the direct intrinsic relationship will not remain (eqn(7.7)). However, perfect collinearity will still exist.

$$w_{BMI_0} \beta_{BMI_0} + w_{BMI_{0-19}} \beta_{BMI_{0-19}} = w_{BMI_{19}} \beta_{BMI_{19}} \quad (7.8)$$

The user must decide whether it is appropriate to standardize the variables under study. Should BMI_0 be given a lesser weighting as it is furthest from SBP at 19 years, or is the smaller variance valid as BMI_0 has a lesser impact? In the Metro Cebu example it could be argued that both stances are justified if correctly interpreted. Understanding the weighting for different estimation methods is vital. The discussion will return to the weights shown in eqn(7.8) when discussing alternative estimators to OLS.

7.5 Tracing the Lifecourse

A growth trajectory plot (Figure 7.4a) is often presented as evidence for the DOHaD and FOAD hypotheses. This approach is demonstrated on the Metro Cebu data,

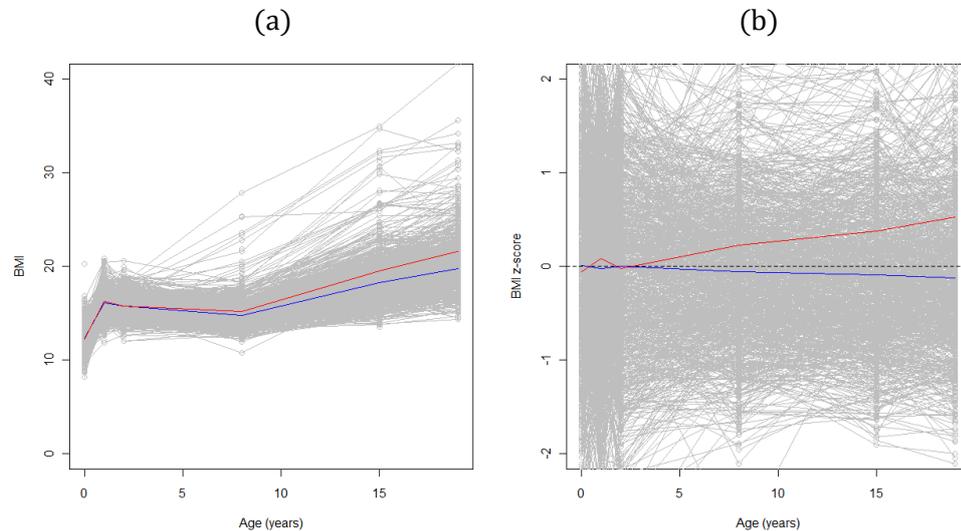


Figure 7.4: Growth trajectory plot of (a) the raw data and (b) z-scores.

Studies have used these plots to identify a distinct growth trajectory for subjects who go on to develop adult chronic diseases. Tracing the raw data in the Metro Cebu example (Figure 7.4a) shows that the mean trajectories follow a similar path from birth to age 8. Beyond this age, the hypertensive group (i.e. SBP > 115 mmHg - shown in red) demonstrates a steeper growth trajectory towards BMI₁₉ than the 'normal' group (shown in blue). To examine the differences in trajectories between groups, studies have used mean z-score growth trajectories (Figure 7.4b). These are calculated by centring and standardizing the observations in each group. It can be observed from the plot that the BMI₀ of the hypertensive group begins lower than that of the 'normal' group (i.e. a lower than average BMI₀). Although the lines meet at age 1 (indicating no difference between groups), the general trend of the 'hypertensive' group is a rapid increase in BMI in comparison to the 'normal' group (i.e. a postnatal 'catch up' in BMI). When this reaches the late teenage years, BMI appears much greater than the average for their age. These results would seem to give strength to the DOHaD hypothesis for hypertension (Barker et al. 2005; Eriksson et al. 2000). However, whilst interpreting the plots we must remind ourselves of a statistical phenomenon highlighted earlier using Galton's data (see Figure 2.2) – labelled "regression to the mean".

If the data are centered for z-scores, the overall mean (including both groups) becomes zero by definition. Therefore, when one group demonstrates a mean below zero (i.e. the dashed line), the other will automatically be greater than zero - a reflection if the number of subjects are equal in both groups. At the two points in which the lines cross, the plot implies that there is no association between BMI and hypertension (i.e. no difference between the 'normal' and 'hypertensive' groups). Similarly, as the average z-scores of the group begin to separate past the age of 2 an increasing association is observed between BMI and hypertension. Therefore, the greatest association between BMI and hypertension is viewed at 19yrs (i.e. BMI_{19}). The increased gradient of the line indicates the rate of change in association. However, this is brought about by the original correlations between SBP and BMI. The covariates closer to the event have a greater association with the health outcomes and are positive, whilst the correlation at birth is weak and negative (see Table 7.1). Observing the change in association (i.e. a simple regression) is not adequate to determine which of BMI_0 , BMI_{19} or BMI_{0-19} is most important to predicting the onset of hypertension in later life. The plot amounts to observing a series of 'simple' regressions, whilst regression to the mean can only serve to bias the result.

The FOAD hypothesis was originally suggested by Professor David Barker at the University of Southampton in 1989. In application, it would imply that improving maternal health and foetal development could be hugely beneficial to the child adapting to its environment in later life - thus reducing the risk of adverse health outcomes. This would naturally suggest that the health intervention should occur prior to birth. In contrast, the DOHaD hypothesis suggests that the critical period occurs at the postnatal stage of development and is in fact the 'catch up' or change in BMI that is most significant to adverse later life health outcomes. If this were true, the intervention would occur in limiting weight gain in postnatal development. For epidemiological study, the interventions implied by the FOAD and DOHaD hypotheses are substantially different (Cole 2010). Plots such as Figure 7.4 could be interpreted to provide evidence to either theory (i.e. the low birth weight and the rapid 'catch up' growth of the hypertensive group). The z-scores indicate the changing association across the lifecourse, but are limited for identifying the critical phase of growth. Regression methods would then seem the obvious progression.

The lifecourse plot is a simple graphical approach that is generated by regressing SBP on the z-score body size measurements (i.e. not change) and plotting the partial regression coefficients (Cole 2004). The critical phase of growth is determined by a rapid change in the coefficients (i.e. an increasing risk of later life health outcomes). Whilst z-

score trajectories represent a series of simple regressions, the lifecourse plot ensures that the coefficients are conditional on other predictors entered into the model. This in turn introduces the guaranteed presence of collinearity amongst covariates. When the lifecourse BMI measures are split into a greater number of partitions (as will be used in section 7.8) the collinearity between neighbouring measures naturally increases. This typically results in the presence of many of the effects associated with collinearity that were discussed in chapter 3. This includes changes in sign and magnitude of the point estimates from the inclusion or exclusion of collinear variables. These can change rapidly upon the inclusion or exclusion of correlated covariates. This can make the identification of the critical phase of growth particularly difficult.

A greater attention may instead be placed on change in body size (i.e. BMI_{0-19} in the Metro Cebu study). This extension would seem intuitive, with aiming to identify a critical period of *growth* in the lifecourse. Consider regressing SBP on the change in z-score. This would appear a useful approach as the aims in the previous methodology have looked to identify steep increases in trajectory or regression coefficients. The researcher could enter BMI_0 and measures of BMI change; or BMI_{19} along with BMI change (although which is preferable is unclear). To include all seven covariates would bring about the identification problem. For p measurements of BMI, there would be a total of $p - 1$ BMI change covariates. In total, along with BMI_0 and BMI_{19} , there will be $p + 1$ predictors entered into the model with only p degrees of freedom - i.e. an infinite range of solutions for \mathbf{b} . Two attainable p -variable models with one of BMI_0 or BMI_{19} entered into the model are linked by a simple equation (similar to computing eqn(7.2) or eqn(7.3) from eqn(7.1) in the three predictor example). This is demonstrated in section 7.8 when the example is extended to seven predictors.

Advanced modelling techniques have been employed to tackle the lifecourse problem. Methods based on an SEM approach are conceptually appealing as relationships can be specified in the data. However, the model specifications must be carefully selected as different choices can lead to a widely varying interpretation. Tu et al. (2011) discuss the use of a growth mixture model, with the approach suffering from a similar problem regarding the model specification (along with issues of convergence and violation of assumptions). Without venturing into the complexities of the wider range of methods, this discussion will explore the notion of setting arbitrary constraints on the model parameters. It will also demonstrate many of the complexities surrounding this problem and what differentiates it from near perfect collinearity as discussed prior to this chapter.

7.6 Constraint Solutions of the Ill Defined Matrix

The problem that defines the lifecourse model is the need to invert a singular (or rank deficient) design matrix to compute estimates from the model. Whilst unconditional estimates can be found using OLS (i.e. via simple regressions), the need to compute individual estimates conditional on the remaining covariates appears key to identifying a critical phase of growth. An approach that directly computes an inverse of the $\mathbf{X}^T\mathbf{X}$ matrix is the generalized (or g-) inverse. When a matrix is non-singular, the g-inverse and the ordinary inverse will be equivalent. However, in cases in which an inverse is unattainable, a g-inverse (sharing some properties of the ordinary inverse) may still be found. The g-inverse of an $n \times k$ matrix \mathbf{A} is based on the SVD in eqn(7.9),

$$\mathbf{A}^- = \mathbf{V}\mathbf{D}^- \mathbf{U}^T \quad (7.9)$$

where \mathbf{U} is an $n \times k$ matrix with orthonormal columns, \mathbf{D} is a $k \times k$ diagonal matrix with non-negative elements, \mathbf{V} is a $k \times k$ orthogonal matrix and \mathbf{A}^- is the g-inverse of \mathbf{A} . A g-inverse is defined when the matrix \mathbf{A} satisfies the following properties,

$$\mathbf{A}\mathbf{A}^- \mathbf{A} = \mathbf{A} \quad (7.10)$$

$$\mathbf{A}^- \mathbf{A}\mathbf{A}^- = \mathbf{A}^- \quad (7.11)$$

When these properties hold, a g-inverse will exist, but is rarely unique unless the matrix \mathbf{A} is invertible. Through the g-inverse the least squares estimation of the regression model can be formed by solution of the normal equations (see section 2.2.2).

$$\begin{aligned} \mathbf{b}_{OLS} &= (\mathbf{X}^T\mathbf{X})^- \mathbf{X}^T\mathbf{y} \\ &= \mathbf{V}\mathbf{A}^- \mathbf{V}^T \mathbf{X}^T\mathbf{y} \end{aligned} \quad (7.12)$$

Justification of the choice of inverse is important to this work. This discussion will provide the basis to link a range of solutions from biased estimators that can be used to obtain parameter estimates from the perfectly collinear model. The following sections will look to generalize the solutions to these problems and discuss alternative estimates from shrinkage regression that fit a common least squares framework. The g- inverse lies at the heart of the proposed 'solutions' to the singular matrix and so the discussion will begin by considering various linear constraints.

7.6.1 Linear Constraints on the Estimation

Due to insufficient degrees of freedom in the lifecourse model, a linear restriction is required on the estimation to allow for the computation of a unique \mathbf{b} . There are in general an infinite set of solutions to the $(\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{y} = \mathbf{b}$ problem when $\mathbf{X}^T\mathbf{X}$ is not full rank. These differ in the constraints imposed on the estimation. Subsequently, the constraints define the usefulness of the estimation in the study. The following section will consider estimates from proposed solutions in the literature and look to identify the set of conditions imposed on the inverse.

The Metro Cebu example is used to consider some of the constraints that can be placed on the estimation. The normal equations for the unstandardized predictors are as follows,

$$(\mathbf{X}^T\mathbf{X})\mathbf{b} = \mathbf{X}^T\mathbf{y}$$

$$\begin{bmatrix} 1443.49 & -1081.12 & 362.38 \\ -1081.12 & 7886.57 & 6805.46 \\ 362.38 & 6805.46 & 7167.83 \end{bmatrix} \begin{bmatrix} b_{\text{BMI}_0} \\ b_{\text{BMI}_{0-19}} \\ b_{\text{BMI}_{19}} \end{bmatrix} = \begin{bmatrix} 117.03 \\ 9447.47 \\ 9564.49 \end{bmatrix} \quad (7.13)$$

As the first two rows equal that of the third there are only two equations for three unknowns (i.e. the null vector is $(1,1,-1)$). One of the normal equations can be replaced with a linear restriction that defines a desired relationship of the parameter estimates. Mazumdar et al. (1980) demonstrate that for each linear restriction there exists a corresponding g-inverse with the same solution,

1. Convert the rank-deficient matrix $(\mathbf{X}^T\mathbf{X})$ to a non-singular matrix by replacing one of the rows with a linear restriction.
e.g. The third row defined by the linear restriction $a_1b_1 + a_2b_2 + a_3b_3 = 0$ becomes $\rho_3 = (a_1, a_2, a_3)$.
2. The inverse of this matrix can then be calculated.
3. Finally, the third column of the inverted matrix is replaced by a column of zero's.
This produces the desired g-inverse.

Example 1 Set one of the coefficients equal to zero; e.g. $\text{BMI}_{19} = 0$ will correspond to the following restriction in step 1 - $(0, 0, 1)$.

$$\mathbf{b} = \begin{bmatrix} 1.09 \\ 1.35 \\ 0 \end{bmatrix}$$

Statistical software packages will often proceed with this solution. This restriction corresponds to removing one variable from the model to obtain a solution from OLS (see bivariable model I in Table 7.2). In other words, a variable is removed when a singularity is found. This restriction implies that in the population, β_3 has no effect. This is a strong assumption that should reflect exactly the population model. Similarly, setting a coefficient equal to a non-zero value would require similarly robust *a priori* evidence.

Example 2 Set a linear restriction such that two coefficients are equal. e.g. equating the coefficients for BMI_0 and BMI_{0-19} would correspond to the following restriction (0, 1, -1)

$$\mathbf{b} = \begin{bmatrix} 0.42 \\ 0.67 \\ 0.67 \end{bmatrix}$$

In the three predictor example this would appear limiting as the challenge is to separate the effects. However, when the example is later extended to consider a series of measurements we may assume two neighbouring estimates to have minimal change. This approach is discussed in Kupper (1985b) for the age-period-cohort model. The merits of this will be considered when the example is later extended (see section 7.8 and chapter 8).

Example 3 Set a linear restriction such that the coefficients sum to zero; e.g. $b_1 + b_2 - b_3 = 0$ would correspond to the following restriction on the coefficients (1, 1, -1)

$$\mathbf{b} = \begin{bmatrix} 0.28 \\ 0.53 \\ 0.81 \end{bmatrix} \quad (7.14)$$

This would appear a more 'natural' restriction than the previous two in that the solution follows the intrinsic relationship of the variables (i.e. $BMI_0 + BMI_{0-19} = BMI_{19}$). The degrees of freedom are still reduced by one, allowing a solution to be found, but not defining the value of any of the coefficients in the population. This constraint has advantages for interpretation and justification (see section 8.6.1), however it can still be argued that it is imposed on the solution in this example (for standardized variables the relationship is weighted, but the process remains equivalent – see section 7.7.1).

7.6.2 The Moore-Penrose Generalized Inverse

When \mathbf{A} is singular there will exist an infinite range of g-inverse solutions which only adhere to select assumptions of the ordinary inverse. The Moore-Penrose (MP) g-inverse (denoted by \mathbf{A}^-) is one that always exists and is unique for an $n \times k$ matrix \mathbf{A} of real or complex entries under eqn(7.10)-(7.11) and the following constraints,

$$(\mathbf{A}\mathbf{A}^-)^T = \mathbf{A}\mathbf{A}^- \quad (7.15)$$

$$(\mathbf{A}^-\mathbf{A})^T = \mathbf{A}^-\mathbf{A} \quad (7.16)$$

The MP g-inverse is defined by imposing the following condition on the diagonal matrix $\mathbf{\Lambda}$,

$$\lambda_{jj} = \begin{cases} 1/\lambda_{jj} & \lambda_{jj} > 0 \\ 0 & \lambda_{jj} = 0 \end{cases} \quad (7.17)$$

For the Metro Cebu example, this is as follows,

$$\begin{aligned} \text{SVD}(\mathbf{X}^T\mathbf{X}) &= \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \\ &= \begin{bmatrix} -0.04 & -0.82 & -0.58 \\ 0.73 & 0.37 & -0.58 \\ 0.69 & -0.44 & 0.58 \end{bmatrix} \begin{bmatrix} 1.44 \times 10^4 & 0 & 0 \\ 0 & 2.13 \times 10^3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -0.04 & -0.82 & -0.58 \\ 0.73 & 0.37 & -0.58 \\ 0.69 & -0.44 & 0.58 \end{bmatrix}^T \end{aligned} \quad (7.18)$$

$$\begin{aligned} &(\mathbf{X}^T\mathbf{X})_{\text{MP}}^- \\ &= \begin{bmatrix} -0.04 & -0.82 & -0.58 \\ 0.73 & 0.37 & -0.58 \\ 0.69 & -0.44 & 0.58 \end{bmatrix} \begin{bmatrix} 1/1.44 \times 10^4 & 0 & 0 \\ 0 & 1/2.13 \times 10^3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -0.04 & -0.82 & -0.58 \\ 0.73 & 0.37 & -0.58 \\ 0.69 & -0.44 & 0.58 \end{bmatrix}^T \\ &= \begin{bmatrix} 3.12 \times 10^{-4} & -1.44 \times 10^{-4} & 1.68 \times 10^{-4} \\ -1.44 \times 10^{-4} & 1.02 \times 10^{-4} & -4.26 \times 10^{-5} \\ 1.68 \times 10^{-4} & -4.26 \times 10^{-5} & 1.25 \times 10^{-4} \end{bmatrix} \end{aligned} \quad (7.19)$$

Notice that the sum of the first two columns of the MP g-inverse equals that of the third (before rounding) similar to example 3 in section 7.6.1. This naturally translates to the solution in eqn(7.20).

$$\mathbf{b} = \begin{bmatrix} b_{\text{BMI}_0} \\ b_{\text{BMI}_{0-19}} \\ b_{\text{BMI}_{19}} \end{bmatrix} = \begin{bmatrix} 0.28 \\ 0.53 \\ 0.81 \end{bmatrix} \quad (7.20)$$

Maintaining this relationship demonstrates an important result to be able to justify the use of the MP g-inverse in the lifecourse problem. PCA would consider the same SVD of the $\mathbf{X}^T\mathbf{X}$ matrix (i.e. the eigen-decomposition) as the MP g-inverse (prior to inversion). Therefore, the vector geometry of these matrices are equivalent for both procedures. Consider the structure of the eigenvectors in variable space geometry (i.e. an ordinary scatter plot). All of the data points are distributed on a common 2D plane \mathbf{p}_1 ,

$$\mathbf{p}_1: \text{BMI}_0 + \text{BMI}_{0-19} - \text{BMI}_{19} = 0 \quad (7.21)$$

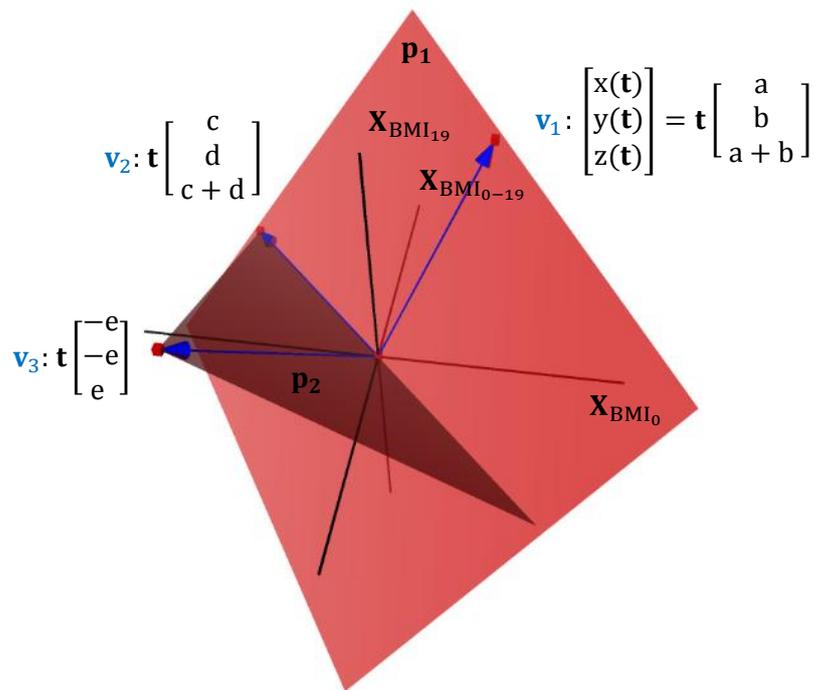


Figure 7.5: Illustration of the planes in which the three PC's must lie.

The first two eigenvectors must lie on the plane \mathbf{p}_1 with positional vectors that follow the relationship defined in eqn(7.21). The cosine of the angle (i.e. correlation) between the original axes and the eigenvectors can be found using the dot product (see Jackson (2003)),

$$\mathbf{v}_{11} = (1 \ 0 \ 0) \cdot (a \ b \ a+b) = a$$

$$\mathbf{v}_{21} = (0 \ 1 \ 0) \cdot (a \ b \ a+b) = b$$

$$\mathbf{v}_{31} = (0 \ 0 \ 1) \cdot (a \ b \ a+b) = a+b$$

The second eigenvector must lie on the plane \mathbf{p}_1 and be orthogonal to the first. Therefore, it corresponds to the same relationship.

$$v_{12} = (1 \ 0 \ 0) \cdot (c \ d \ c+d) = c$$

$$v_{22} = (0 \ 1 \ 0) \cdot (c \ d \ c+d) = d$$

$$v_{32} = (0 \ 0 \ 1) \cdot (c \ d \ c+d) = c+d$$

The plane \mathbf{p}_1 is fixed by the null vector. Therefore the redundant third dimension is represented by the normal vector to this plane (1,1,-1) – i.e. the null. The vector \mathbf{v}_3 must lie on the plane \mathbf{p}_2 orthogonal to \mathbf{p}_1 passing through the origin (i.e. $\mathbf{p}_2: \mathbf{X}_{\text{BMI}_0} + \mathbf{X}_{\text{BMI}_{0-19}} + 2\mathbf{X}_{\text{BMI}_{19}} = 0$). This determines a unit direction vector (fixed for *all* examples) for the PC weights $(-\sqrt{3}/3, -\sqrt{3}/3, \sqrt{3}/3)$ to maintain orthogonality.

The eigenvector for the Metro Cebu data maintains the intrinsic relationship of the variables in the first two components.

$$\mathbf{V} = \begin{bmatrix} -0.04 & -0.82 & -0.58 \\ 0.73 & 0.37 & -0.58 \\ 0.69 & -0.44 & 0.58 \end{bmatrix} \quad (7.22)$$

It is the scaling by the eigenvalues that determines the variance explained by each PC. This is demonstrated by the SVD of the MP g-inverse for the perfectly collinear covariates,

$$\begin{aligned} & (\mathbf{X}^T \mathbf{X})_{\text{MP}}^- \\ = & \begin{bmatrix} a & c & -\sqrt{3}/3 \\ b & d & -\sqrt{3}/3 \\ a+b & c+d & \sqrt{3}/3 \end{bmatrix} \begin{bmatrix} 1/\lambda_1 & 0 & 0 \\ 0 & 1/\lambda_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a & b & a+b \\ c & d & c+d \\ -\sqrt{3}/3 & -\sqrt{3}/3 & \sqrt{3}/3 \end{bmatrix} \\ = & \begin{bmatrix} a^2/\lambda_1 + c^2/\lambda_2 & ab/\lambda_1 + cd/\lambda_2 & a(a+b)/\lambda_1 + c(c+d)/\lambda_2 \\ ab/\lambda_1 + cd/\lambda_2 & b^2/\lambda_1 + d^2/\lambda_2 & b(a+b)/\lambda_1 + d(c+d)/\lambda_2 \\ a(a+b)/\lambda_1 + c(c+d)/\lambda_2 & b(a+b)/\lambda_1 + d(c+d)/\lambda_2 & (a+b)^2/\lambda_1 + (c+d)^2/\lambda_2 \end{bmatrix} \quad (7.23) \end{aligned}$$

The third column in the eigenvector (i.e. the null vector) has no effect on the computation of the g-inverse due to the corresponding zero eigenvalue (i.e. $\lambda_3 = 0$). The intrinsic relationship corresponds to each of the columns as well as rows of the MP g-inverse. This follows in turn to the estimated coefficients. The estimates gained from the (1, 1, -1) linear restriction and the MP g-inverse are identical, however the distinction between the approaches is that in the former the restriction is imposed and in the latter it is maintained from the intrinsic relationship in the data and the null vector. This is an important feature of the analysis that allows this restriction to be promoted as the most ‘natural’ given the structure inherent in the data.

7.7 The Application of Shrinkage Methods to the Lifecourse

In section 4.4, PCR and PLS were introduced as shrinkage regression methods. Although employing a biased estimator will impact on the accuracy of the estimation, it can substantially improve precision. The techniques are particularly useful for situations in which high levels of collinearity are likely to be present. As such, they are widely employed in areas such as chemometrics and bioinformatics – particularly when $k \gg n$. In comparison, these methods have experienced relatively little use and discussion in epidemiological study (Cole 2010). A particular benefit is that shrinkage techniques can impose (or rather maintain) constraints in the estimation process to provide estimates from the rank deficient matrix. In this section, PCR and PLS methodology are revisited to consider a potential application in the lifecourse problem. The interpretation of the results will be explored as well as the justification for employing the methodology.

7.7.1 Principal Component Regression

The use of OLS in the perfect collinearity problem dictates that one covariate must be removed to obtain unique estimates. For PCR, the same problem would exist if the full complement of components is retained (as the result is identical to OLS - see eqn(4.26)). However, by retaining the first two PC's in a PCR, the zero eigenvalue of $\mathbf{\Lambda}$ will be removed (and hence the null vector).

$$\mathbf{Z}_{m=2}^T \mathbf{Z}_{m=2} = \mathbf{\Lambda}_{m=2} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \quad (7.24)$$

$$\mathbf{\Lambda}_{m=2}^{-1} = \begin{bmatrix} 1/\lambda_1 & 0 \\ 0 & 1/\lambda_2 \end{bmatrix} \quad (7.25)$$

The PC's are entered into eqn(4.22) to obtain the regression coefficients for the PCR estimator. The components represent a combination of the original three covariates. This process (for $m = 2$) corresponds to performing an MP g-inverse on the $\mathbf{X}^T \mathbf{X}$ matrix. For the MP-inverse procedure, the zero row in $\mathbf{\Lambda}^{-1}$ cancels with the corresponding redundant eigenvector in \mathbf{V} (see eqn(7.23)). Therefore, retaining two components produces the same estimation of \mathbf{b} as the MP g-inverse and the $(1, 1, -1)$ linear restriction in section 7.6.1. This is an important step in being able to justify the result. In using a PCR the intrinsic structure of the variables is maintained.

As was also shown in section 7.6.2, each column of \mathbf{V} (apart from the third corresponding to the null) will follow the intrinsic relationship amongst the variables. This translates to the regression coefficients. For instance, the first eigenvector provides the ‘weight’ for the first component $(-0.04, 0.73, 0.69)^T$ (see eqn(7.22)). Using eqn(4.18) the weights can then be used to find the first PC (\mathbf{z}_1),

$$\mathbf{z}_1 = (-0.04)\mathbf{X}_{\text{BMI}_0} + 0.73\mathbf{X}_{\text{BMI}_{0-19}} + 0.69\mathbf{X}_{\text{BMI}_{19}} \quad (7.26)$$

Following eqn(4.20), \mathbf{y} is regressed onto \mathbf{Z} to find the single PCR regression coefficient $\mathbf{b}_{\text{PCR}}^{m=1} = 0.93$. Following eqn(4.23), $\mathbf{b}_{\text{PCR}}^{m=1}$ is multiplied by the weight matrix \mathbf{V} to rotate the estimates back to the original axes.

Variable	PCR Coefficients	
	1 Component Model (95% CI)	2 Component Model (95% CI)
Unstandardized		
BMI ₀	-0.04 (-0.06 to -0.01)	0.28 (-0.06 to 0.62)
BMI ₀₋₁₉	0.68 (0.56 to 0.8)	0.53 (0.34 to 0.73)
BMI ₁₉	0.64 (0.53 to 0.75)	0.81 (0.59 to 1.03)
Standardized		
BMI ₀	-0.42 (-0.73 to -0.11)	0.43 (-0.14 to 1)
BMI ₀₋₁₉	1.83 (1.5 to 2.16)	1.75 (1.41 to 2.08)
BMI ₁₉	1.73 (1.39 to 2.07)	2.03 (1.65 to 2.4)
Adj R², %	11.64	11.85

Table 7.3: PCR results for one and two component models on the Metro Cebu data.

Confidence intervals have been provided using jack-knife “leave one out” (Tukey 1958). Data is re-sampled with one observation removed and coefficient estimates calculated for each repetition. The cumulative R_y^2 in the two component model is the same as the least squares bivariate models (as no variance is discarded). The single component model explains 98% of the covariance relative to the full model. This appears preferable to reduce the collinearity impact. For the single component model BMI₀ had a small negative association with SBP. BMI₀₋₁₉ and BMI₁₉ both had significant positive associations. For the standardized model, BMI₀ had a negative association with the response, whilst BMI₀₋₁₉ and BMI₁₉ both had strong positive associations. In the 2 component models, BMI₀ reversed sign, but remained non-significant.

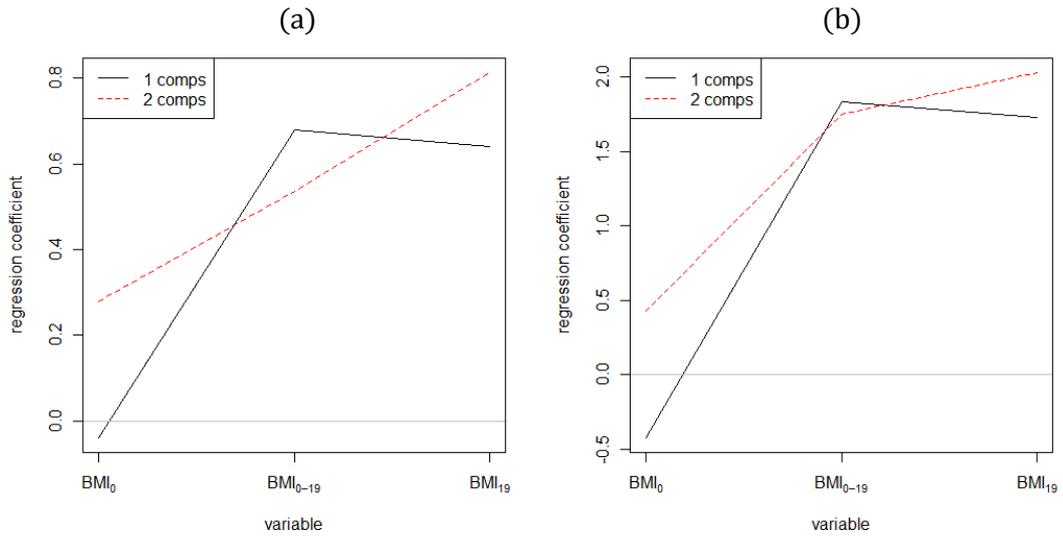


Figure 7.6: (a) Unstandardized and (b) standardized coefficients from PCR.

In the perfectly collinear example, reducing to two components still explains the same variance in the response (R_y^2) as OLS for any bivariable model because the data is two-dimensional. The vector geometry in Figure 7.2 explains this feature well in that the regression space spanned by any two predictors is the same as that of the full three predictor model. Therefore, the projected response \hat{y} is equivalent for both estimators. The intrinsic relationship amongst the covariates directly translates for the unstandardized predictors as follows,

$$\begin{aligned}
 \mathbf{y}_{\text{SBP}} &= 0.28 \mathbf{X}_{\text{BMI}_0} + 0.53 \mathbf{X}_{\text{BMI}_{0-19}} + 0.81 \mathbf{X}_{\text{BMI}_{19}} \\
 &= 0.28 \mathbf{X}_{\text{BMI}_0} + 0.53 (\mathbf{X}_{\text{BMI}_{19}} - \mathbf{X}_{\text{BMI}_0}) + 0.81 \mathbf{X}_{\text{BMI}_{19}} \\
 &= (0.28 - 0.53) \mathbf{X}_{\text{BMI}_0} + (0.81 + 0.53) \mathbf{X}_{\text{BMI}_{19}} \\
 &= -0.25 \mathbf{X}_{\text{BMI}_0} + 1.34 \mathbf{X}_{\text{BMI}_{19}}
 \end{aligned} \tag{7.27}$$

From eqn(7.27) a direct link can be observed to the regression coefficients and variance explained from the OLS result in Table 7.2 (although some rounding occurs). Therefore PCR can be viewed as a natural extension of OLS onto axes that partition variance for the ill-defined matrix. If maintaining the intrinsic relationship of the variables justifies the use of PCR in the 3 predictor lifecourse problem, then the next stage is to discuss questions regarding its application – i.e. how many components to retain and whether to use standardized or unstandardized covariates. For this purpose, the discussion returns to the use of vector geometry to illustrate PCR on perfectly collinear data.

When variances are equal for covariates in the bivariable model, the first PC bisects the two covariate vectors perfectly (e.g. see Figure 4.4). However, this is not the case in the Metro Cebu example. The first PC is weighted toward the covariate with the greater standard deviation (i.e. BMI_{0-19} in Figure 7.7),

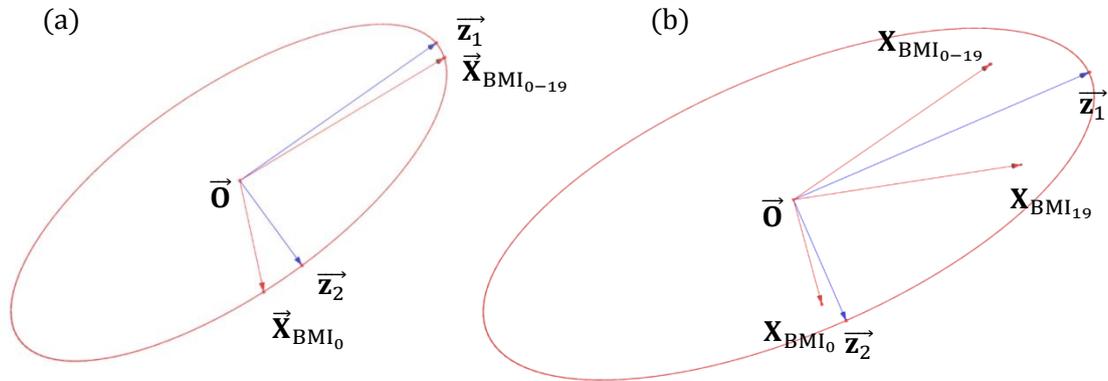


Figure 7.7: Illustration of PCA for (a) a bivariable and (b) full model.

Notice that in the bivariable example, the covariance ellipse passes through each of the covariate vectors (see Figure 7.7a). The covariance ellipse can only be maintained in the PCA process because it rests on the assumption that the sum of squared weights is equal to unity (see section 4.2.1). This ensures that the PC's are on the same scale as the original covariates. As demonstrated in Figure 7.7b, the ellipse is proportional to the three predictors entered and so will no longer pass through each vector. This is a result of considering the data for three variables on a 2-dimensional plane.

The process of standardization is to give a weighting to the constraint in the data, such that those variables with a smaller variance are not penalized in the analysis (see eqn(7.8)). All data are centered. However, the researcher may choose to enter standardized data, or to standardize the coefficients after calculating the components. If the data are entered unstandardized, the PCR procedure based on SVD would produce unstandardized predictors. These would directly follow the intrinsic relationship of the data in eqn(7.21) (as demonstrated in section 7.6.2). If the coefficients are then standardized directly using eqn(2.30), the following weighting would transform back to the intrinsic relationship,

$$\frac{1}{sd_{BMI_0}} b_{BMI_0} + \frac{1}{sd_{BMI_{0-19}}} b_{BMI_{0-19}} = \frac{1}{sd_{BMI_{19}}} b_{BMI_{19}} \quad (7.28)$$

If instead the covariates are standardized prior to PCR, the covariates would be weighted by their standard deviation as follows,

$$sd_{\text{BMI}_0} \underline{\mathbf{X}}_{\text{BMI}_0} + sd_{\text{BMI}_{0-19}} \underline{\mathbf{X}}_{\text{BMI}_{0-19}} - sd_{\text{BMI}_{19}} \underline{\mathbf{X}}_{\text{BMI}_{19}} = 0 \quad (7.29)$$

where $\underline{\mathbf{X}}$ are scaled covariates with unit length. The ordinary SVD PCR procedure would naturally translate the same relationship to the coefficient estimates,

$$sd_{\text{BMI}_0} \underline{\mathbf{b}}_{\text{BMI}_0} + sd_{\text{BMI}_{0-19}} \underline{\mathbf{b}}_{\text{BMI}_{0-19}} - sd_{\text{BMI}_{19}} \underline{\mathbf{b}}_{\text{BMI}_{19}} = 0 \quad (7.30)$$

The coefficients are weighted by standard deviation. Similarly, unstandardized estimates can then be generated based on standardized data by using the relationship in eqn(2.30),

$$sd_{\text{BMI}_0}^2 \underline{\mathbf{b}}_{\text{BMI}_0} + sd_{\text{BMI}_{0-19}}^2 \underline{\mathbf{b}}_{\text{BMI}_{0-19}} - sd_{\text{BMI}_{19}}^2 \underline{\mathbf{b}}_{\text{BMI}_{19}} = 0 \quad (7.31)$$

The coefficients would now be weighted by variance (this result can be verified using the coefficients in Table 7.3). The important result is that in each of these examples, whilst the intrinsic relationship may not hold directly, these weights prove that the variables still adhere to the same condition. Therefore, the decision over whether to employ standardized or unstandardized weights is very much conceptual and must be factored into the interpretation of any analysis.

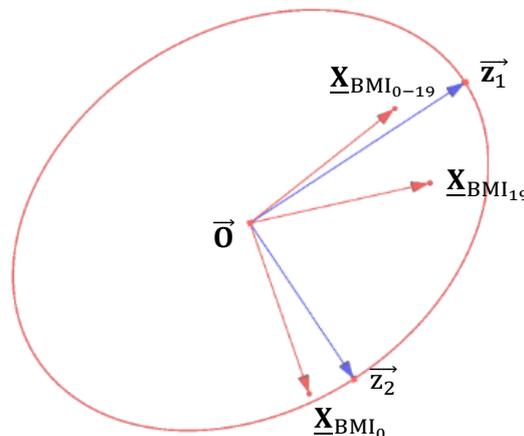


Figure 7.8: Illustration of PCA for the standardized predictors in the full model.

The first PC in Figure 7.8 is not orthogonal to BMI_0 and so standardization has generated a greater weight on this variable. This result was expected as BMI_0 is heavily penalised in the unstandardized geometry. The greater covariance between BMI_{19} and BMI_{0-19} still dictates that these covariates load highly on the first PC.

7.7.2 Partial Least Squares Regression

For PLS the focus returns to the $\mathbf{X}^T\mathbf{y}$ matrix. The intrinsic relationship of the lifecourse covariates directly translates to the covariances in this matrix as follows,

$$\begin{aligned}
 \text{Cov}(\mathbf{X}_{\text{BMI}_{19}}, \mathbf{y}) &= \text{Cov}(\mathbf{X}_{\text{BMI}_0} + \mathbf{X}_{\text{BMI}_{0-19}}, \mathbf{y}) \\
 &= \frac{1}{2} [\text{Var}(\mathbf{X}_{\text{BMI}_0} + \mathbf{X}_{\text{BMI}_{0-19}} + \mathbf{y}) + \text{Var}(\mathbf{X}_{\text{BMI}_0} + \mathbf{X}_{\text{BMI}_{0-19}}) - \text{Var}(\mathbf{y})] \\
 &= \frac{1}{2} [\text{Var}(\mathbf{X}_{\text{BMI}_0}) + \text{Var}(\mathbf{X}_{\text{BMI}_{0-19}}) + \text{Var}(\mathbf{y}) + 2\text{Cov}(\mathbf{X}_{\text{BMI}_0}, \mathbf{X}_{\text{BMI}_{0-19}}) + \\
 &\quad 2\text{Cov}(\mathbf{X}_{\text{BMI}_0}, \mathbf{y}) + 2\text{Cov}(\mathbf{X}_{\text{BMI}_{0-19}}, \mathbf{y}) - \text{Var}(\mathbf{X}_{\text{BMI}_0}) - \text{Var}(\mathbf{X}_{\text{BMI}_{0-19}}) - \\
 &\quad 2\text{Cov}(\mathbf{X}_{\text{BMI}_0}, \mathbf{X}_{\text{BMI}_{0-19}}) - \text{Var}(\mathbf{y})] \\
 &= \text{Cov}(\mathbf{X}_{\text{BMI}_0}, \mathbf{y}) + \text{Cov}(\mathbf{X}_{\text{BMI}_{0-19}}, \mathbf{y})
 \end{aligned} \tag{7.32}$$

Using this relationship, the SVD of $\mathbf{X}^T\mathbf{y}$ generates eigenvectors that follow the same intrinsic relationship (see section 5.3 for SVD of $\mathbf{X}^T\mathbf{y}$).

$$\mathbf{X}^T\mathbf{y} = \mathbf{L}\mathbf{M}\mathbf{N}^T = \begin{bmatrix} \text{Cov}(\mathbf{X}_1, \mathbf{y})/\|\mathbf{L}\| \\ \text{Cov}(\mathbf{X}_2, \mathbf{y})/\|\mathbf{L}\| \\ \text{Cov}(\mathbf{X}_3, \mathbf{y})/\|\mathbf{L}\| \end{bmatrix} \left[(n-1) \sum_{j=1}^3 \sqrt{\text{Cov}(\mathbf{x}_j, \mathbf{y})^2} \right] [\mathbf{1}]$$

The SVD for the Metro Cebu data is as follows,

$$\begin{aligned}
 \mathbf{X}^T\mathbf{y} &= \begin{bmatrix} 11.06 \\ 892.89 \\ 903.94 \end{bmatrix} = \begin{bmatrix} 0.01/1.33 \\ 0.94/1.33 \\ 0.95/1.33 \end{bmatrix} \left[(955-1) \sqrt{0.01^2 + 0.94^2 + 0.95^2} \right] [\mathbf{1}] \\
 &= \begin{bmatrix} 0.01 \\ 0.7 \\ 0.71 \end{bmatrix} [1270.61] [\mathbf{1}]
 \end{aligned} \tag{7.33}$$

This is an important result as each factor extracted in PLS will maintain this relationship within the data and the coefficients are given a weighting (\mathbf{p}) that corresponds to this relationship. For instance, the left singular vector in eqn(7.33) provides the following weights for the first PLS component to obtain the scores (\mathbf{t}_1) (see eqn(4.27)),

$$\mathbf{t}_1 = [\mathbf{X}_{\text{BMI}_0}, \mathbf{X}_{\text{BMI}_{19}}, \mathbf{X}_{\text{BMI}_{0-19}}] \begin{bmatrix} 0.01 \\ 0.7 \\ 0.71 \end{bmatrix}$$

Following eqn(4.28), \mathbf{y} is regressed onto the scores matrix \mathbf{T} (which contains only \mathbf{t}_1 in this case) to attain the single PLS regression coefficient $b_{\text{PLS}}^{m=1} = 0.939$. Following eqn(4.29), the PLS coefficients are transformed back onto the original axes,

$$\begin{bmatrix} 0.01 \\ 0.7 \\ 0.71 \end{bmatrix} 0.93 = \begin{bmatrix} 0.01 \\ 0.66 \\ 0.67 \end{bmatrix}$$

The residuals of both \mathbf{X} and \mathbf{y} matrices are then re-entered into the algorithm and the SVD is once again implemented to obtain the second PLS component. As the left singular eigenvector of $\mathbf{X}^T \mathbf{y}$ maintains the intrinsic relationship in the data, the same relationship translates to the regression coefficients for both $m = 1$ and $m = 2$ components.

	PLS Coefficients 1 Component Model (95% CI)	PLS Coefficients 2 Component Model (95% CI)
Unstandardized		
BMI ₀	0.01 (-0.05 to 0.06)	0.28 (-0.06 to 0.62)
BMI ₀₋₁₉	0.66 (0.54 to 0.78)	0.53 (0.34 to 0.73)
BMI ₁₉	0.67 (0.55 to 0.79)	0.81 (0.59 to 1.03)
Standardized		
BMI ₀	0.05 (-0.3 to 0.41)	0.43 (-0.14 to 1)
BMI ₀₋₁₉	1.81 (1.49 to 2.14)	1.75 (1.41 to 2.08)
BMI ₁₉	1.93 (1.58 to 2.28)	2.03 (1.65 to 2.4)
Adj R², %	11.73	11.85

Table 7.4: PLS results for one and two component models on the Metro Cebu data.

The single component model explains more than 98% of the covariance between SBP and the three BMI measures. Based on the R_y^2 it would seem that the single component model would be preferable. The PLS analysis with a single component retained produced coefficients that were similar to those from the simple OLS regressions (The reasons for this were discussed for the D-index - see Figure 5.2). BMI₀ had no association with the response, whilst BMI₁₉ and BMI₀₋₁₉ both had strong significant positive associations with SBP. In the standardized single component model, the coefficients followed a similar pattern with positive associations strengthening further. For the two component model, BMI₀ demonstrated no association with SBP. BMI₁₉ had a slightly stronger positive association, whilst BMI₀₋₁₉ had a slightly weaker positive association.

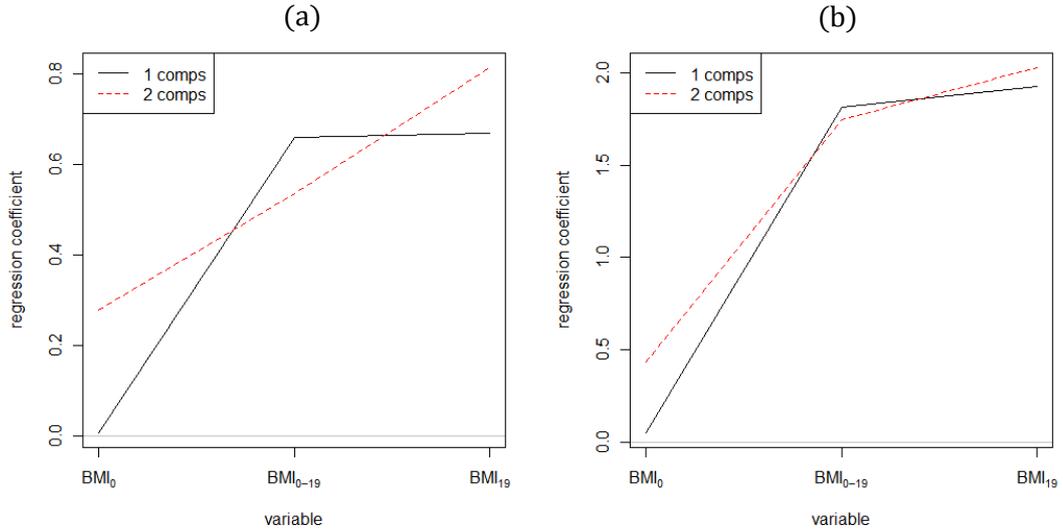


Figure 7.9: Plot of the (a) unstandardized and (b) standardized coefficients from PLS.

Similar to PCR, the two component model has a direct relationship to the OLS result from Table 7.2,

$$\begin{aligned}
 y_{\text{SBP}} &= 0.28 \mathbf{X}_{\text{BMI}_0} + 0.53 \mathbf{X}_{\text{BMI}_{0-19}} + 0.81 \mathbf{X}_{\text{BMI}_{19}} \\
 &= 0.28 \mathbf{X}_{\text{BMI}_0} + 0.53 (\mathbf{X}_{\text{BMI}_{19}} - \mathbf{X}_{\text{BMI}_0}) + 0.81 \mathbf{X}_{\text{BMI}_{19}} \\
 &= (0.28 - 0.53) \mathbf{X}_{\text{BMI}_0} + (0.53 + 0.81) \mathbf{X}_{\text{BMI}_{19}} \\
 &= -0.25 \mathbf{X}_{\text{BMI}_0} + 1.34 \mathbf{X}_{\text{BMI}_{19}}
 \end{aligned} \tag{7.34}$$

It is clear that PLS (like PCR) is partitioning the variation in OLS as to the intrinsic relationship that defines the lifecourse covariates. PLS and PCR coefficients for two component models are identical. No variance is discarded by either method and so the results would always be the same when $k - 1$ components are retained.

The discussion regarding the weighting of the unstandardized and standardized models is identical to that presented for the PCR model. Although the coefficients do not directly follow the intrinsic relationship, it can be seen from the weightings that this result still holds. Once again, the discussion rests on whether conceptually it seems appropriate to penalize BMI_0 due to the smaller variance in relation to BMI_{19} and BMI_{0-19} . In a PLS analysis, the variables with a greater correlation with the outcome are naturally given a greater weight due to the covariance maximizing aim of the method. If two components are retained the result is identical to PCR, however the axes are built on the covariance maximizing philosophy.

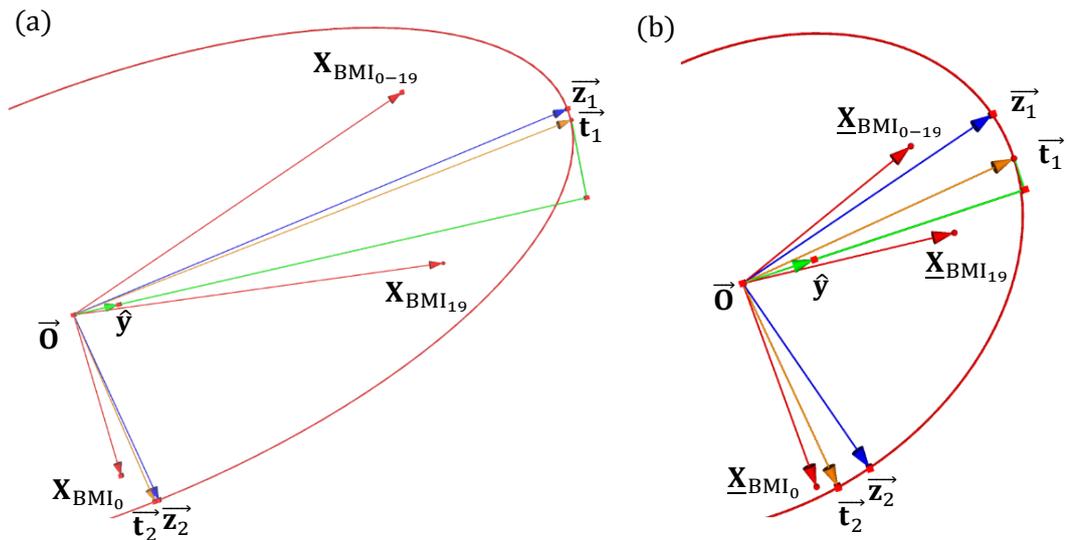


Figure 7.10: PLS components from (a) standardized and (b) unstandardized data.

The vector geometry in Figure 7.10 illustrates this idea well. Removing t_2 in the unstandardized model removes almost all of the variance of BMI_0 (this is demonstrated in the one component models in Table 7.4). However, in the standardized model, the change between PCR and PLS components is much greater. This is because of the greater weighting given to BMI_0 through standardization is countered by the greater weighting given by PLS to those variables closer to the outcome (i.e. BMI_{0-19} and BMI_{19}). Although the two component model produces the same estimates after reverse transformation to the original axes, the single component model places a lesser weight on BMI_0 . Therefore, choosing unstandardized or standardized single component models (in this example) has a minimal effect due to the small association between BMI_0 and SBP.

This discussion has been useful to demonstrate that any application of PCR or PLS is equally justifiable based on maintaining the intrinsic relationship of the variables. These coefficients are weighted differently, whether the standard deviations of the variables are believed to be important to the interpretation, however the same relationships exist (if only indirectly). The decision of standardization becomes a conceptual one, with either seemingly plausible if accompanied by justification of the choice and incorporated into the interpretation. With retaining only a single component in PLS, a greater variation on BMI_{0-19} and BMI_{19} is retained as these have a greater association with the response. Standardizing the variables has no effect on the correlations, but it does on variance distribution. One issue that has been largely ignored in this example is the decision of how many components to retain. In both PCR and PLS examples the single component model has been adequate with a high explained variance in the response.

7.8 Identifying the Critical Phase of Growth

The Metro Cebu data is extended to incorporate BMI measurements on each of the participants at birth (BMI_0), 1 year (BMI_1), 2 years (BMI_2), 8 years (BMI_8), 15 years (BMI_{15}) and 19 years (BMI_{19}) to identify the critical phase of growth for SBP in later life. A total of five BMI change measurements (defined by the change in neighbouring BMI raw scores) are entered into the analysis, along with BMI_0 and BMI_{19} . Whilst variance explained may increase with the addition of covariates, the perfect collinearity remains.

	BMI_0	BMI_{0-1}	BMI_{1-2}	BMI_{2-8}	BMI_{8-15}	BMI_{15-19}	BMI_{19}
BMI_0							
BMI_{0-1}	-0.57						
BMI_{1-2}	-0.06	-0.41					
BMI_{2-8}	-0.01	-0.06	-0.33				
BMI_{8-15}	0.02	0.10	-0.02	0.14			
BMI_{15-19}	0.00	-0.04	-0.04	0.08	-0.21		
BMI_{19}	0.11	0.14	-0.03	0.45	0.66	0.43	
SBP	0.01	0.04	-0.03	0.16	0.23	0.19	0.35

Table 7.5: Correlations of the Cebu data including BMI change measures.

The example here differs from our original paper (Tu et al. 2010) in that BMI is considered rather than weight to provide a different interpretation on the results gained from the methods. Also, the results from unstandardized covariates are considered in addition to the standardized measures (i.e. centered and scaled to unit length). This differs from the z-scores that were used in the original paper. The use of z-score change in place of BMI_{0-19} (as defined here) is perhaps confusing. It amounts to standardizing BMI_0 and BMI_{19} and then observing change (which is then not unit variance). This will maintain direct links between the estimates that were observed for the unstandardized covariates, however it does become difficult to interpret on the original weightings.

As Table 7.5 demonstrates, the inclusion of additional BMI measurements will predictably increase the correlation amongst the covariates. This will subsequently increase any potential impact of collinearity on parameter estimates obtained from regression procedures (without remedial action). Similar to the three covariate example in section 7.7, all seven covariates cannot be entered simultaneously into an OLS regression. This is due to the covariates spanning only six dimensions (i.e. insufficient degrees of freedom). Instead, five BMI change covariates can be entered along with BMI_0 , or five BMI change covariates along with BMI_{19} .

Variable	Univariable	Multivariable Regression	Multivariable Regression
	Regression	Model I Coefficient	Model II Coefficient
	Coefficient (95% CI)	(95% CI)	(95% CI)
Unstandardized			
BBMI	0.08 (-0.47 to 0.63)	0.58 (-0.12 to 1.29)	
BMI ₀₋₁	0.29 (-0.15 to 0.73)	0.70 (0.06 to 1.35)	0.12 (-0.42 to 0.66)
BMI ₁₋₂	-0.31 (-0.89 to 0.27)	0.62 (-0.11 to 1.36)	0.04 (-0.71 to 0.79)
BMI ₂₋₈	1.24 (0.74 to 1.73)	1.03 (0.48 to 1.57)	0.44 (-0.32 to 1.21)
BMI ₈₋₁₅	1.38 (1.01 to 1.75)	1.50 (1.12 to 1.88)	0.92 (0.06 to 1.77)
BMI ₁₅₋₁₉	1.32 (0.90 to 1.75)	1.65 (1.23 to 2.07)	1.07 (0.25 to 1.89)
CBMI	1.33 (1.10 to 1.56)		0.58 (-0.12 to 1.29)
Standardized			
BBMI	0.10 (-0.57 to 0.77)	0.72 (-0.15 to 1.59)	
BMI ₀₋₁	0.45 (-0.23 to 1.12)	1.07 (0.09 to 2.06)	0.18 (-0.64 to 1.00)
BMI ₁₋₂	-0.36 (-1.03 to 0.31)	0.72 (-0.13 to 1.57)	0.04 (-0.82 to 0.91)
BMI ₂₋₈	1.65 (0.98 to 2.31)	1.37 (0.64 to 2.10)	0.59 (-0.43 to 1.61)
BMI ₈₋₁₅	2.43 (1.78 to 3.09)	2.65 (1.98 to 3.32)	1.62 (0.11 to 3.13)
BMI ₁₅₋₁₉	2.04 (1.38 to 2.70)	2.54 (1.89 to 3.19)	1.64 (0.38 to 2.90)
CBMI	3.66 (3.03 to 4.29)		1.60 (-0.34 to 3.54)
Adj R ²		12.12	12.12

Table 7.6: Coefficient estimates from two attainable models using OLS regression.

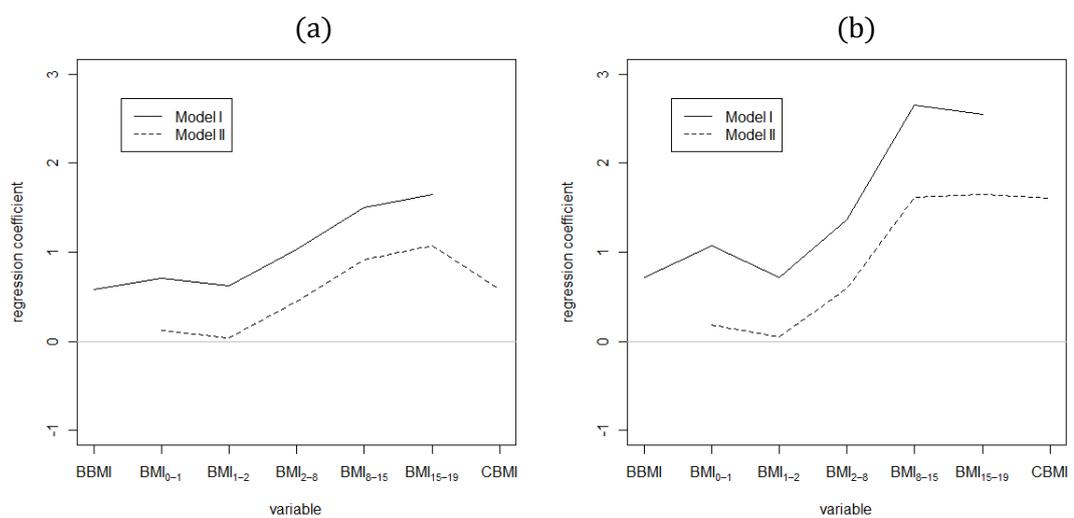


Figure 7.11: Plots of (a) unstandardized and (b) standardized coefficient estimates.

When SBP was regressed on BMI_0 and the 5 BMI change covariates, a positive association was observed for all covariates with the outcome. BMI_0 had the lowest unstandardized coefficient (0.58 (-0.12 to 1.29)) and was non-significant. BMI_{1-2} was also non-significant (0.72 (-0.13 to 1.57)). It is noticeable from Figure 7.11a that the coefficients follow parallel trends. BMI_{19} is the summation of the incremental BMI change and BMI_0 , therefore the coefficients in model II can be calculated from model I.

$$y_{SBP} = b_{BMI_0} X_{BMI_0} + b_{BMI_{0-19}} X_{BMI_{0-19}} \quad (7.35)$$

$$= b_{BMI_0} (X_{BMI_{19}} - X_{BMI_{0-19}}) + b_{BMI_{0-19}} X_{BMI_{0-19}}$$

$$= (b_{BMI_{0-19}} - b_{BMI_0}) X_{BMI_{0-19}} + b_{BMI_0} X_{BMI_{19}} \quad (7.36)$$

Notice that the coefficient estimate for BMI_{19} in model II is equal to that of BMI_0 in model I (for the non-standardized model). Also when the incremental BMI change covariates are entered, the coefficients in model II would always be reduced by the same coefficient of b_{BMI_0} in model I (i.e. reduced on the condition that b_{BMI_0} is positive). Therefore, these models are intrinsically linked and it again appears difficult to justify one as superior to the other or any more useful in answering the original research hypothesis.

The fundamental problem of tracing the z -scores is that it will suffer from regression to the mean (see section 7.5), whilst examples such as the lifecourse plot will incorporate high levels of collinearity using an increased number of measures. It was discussed in section 7.6.1 that dropping a variable from the model is equivalent to setting a linear restriction equal to zero. This does not seem an ideal solution to the collinearity problem and in particular to answer the initial research question (as shown by the plots in Figure 7.11). Generating any of the five remaining attainable 6 covariate OLS models by dropping incremental BMI change covariates rests on similar *prior* assumptions.

A possible 'solution' that would (on the surface) appear more feasible when a greater number of covariates are entered is setting any neighbouring change in coefficients equal. This is discussed in Kupper (1985b) for the age-period-cohort model (see chapter 8). With increased measures, the difference between two covariates is seemingly reduced and so setting two effects equal would appear less dangerous. However, as Kupper discusses, this restriction must be representative of the (unattainable) population model, otherwise the results will be biased. Without robust *a priori* information to justify the restriction, this 'solution' would seem best avoided due to the increasing collinearity in the model with increased measurements.

	PCR			
	1 Component	2 Component	3 Component	4 Component
Unstandardized				
BMI₀	0.04 (0.00 to 0.08)	0.11 (-0.05 to 0.28)	0.03 (-0.15 to 0.22)	0.00 (-0.21 to 0.21)
BMI₀₋₁	0.09 (0.03 to 0.15)	-0.04 (-0.32 to 0.24)	0.08 (-0.21 to 0.37)	0.09 (-0.20 to 0.39)
BMI₁₋₂	-0.04 (-0.07 to 0.00)	0.00 (-0.08 to 0.09)	-0.06 (-0.17 to 0.05)	0.02 (-0.26 to 0.30)
BMI₂₋₈	0.23 (0.15 to 0.31)	0.24 (0.14 to 0.34)	0.26 (0.16 to 0.37)	0.17 (-0.18 to 0.51)
BMI₈₋₁₅	0.48 (0.37 to 0.59)	0.41 (0.13 to 0.68)	0.27 (0.03 to 0.51)	0.29 (0.03 to 0.54)
BMI₁₅₋₁₉	0.22 (0.12 to 0.31)	0.31 (0.00 to 0.62)	0.46 (0.21 to 0.72)	0.50 (0.21 to 0.78)
BMI₁₉	1.02 (0.83 to 1.20)	1.04 (0.85 to 1.23)	1.06 (0.87 to 1.24)	1.07 (0.88 to 1.26)
Standardized				
BMI₀	-0.10 (-0.63 to 0.42)	0.34 (-0.02 to 0.70)	0.34 (-0.09 to 0.77)	-0.11 (-0.52 to 0.30)
BMI₀₋₁	0.58 (0.00 to 1.15)	0.08 (-0.30 to 0.47)	0.08 (-0.32 to 0.48)	0.17 (-0.20 to 0.54)
BMI₁₋₂	-0.67 (-1.04 to -0.30)	-0.45 (-0.79 to -0.11)	-0.45 (-1.12 to 0.21)	0.23 (-0.26 to 0.72)
BMI₂₋₈	1.08 (0.64 to 1.52)	1.15 (0.80 to 1.50)	1.15 (0.66 to 1.64)	0.79 (0.36 to 1.21)
BMI₈₋₁₅	1.12 (0.76 to 1.49)	1.21 (0.80 to 1.62)	1.21 (0.32 to 2.09)	1.17 (0.71 to 1.63)
BMI₁₅₋₁₉	0.52 (0.03 to 1.01)	0.63 (0.17 to 1.08)	0.63 (-0.14 to 1.40)	1.28 (0.78 to 1.79)
BMI₁₉	1.53 (1.02 to 2.05)	1.70 (1.35 to 2.04)	1.70 (1.30 to 2.10)	2.00 (1.63 to 2.36)
Adj R², %	11.99	12.01	12.11	12.05
	5 Component	6 Component		
Unstandardized				
BMI₀	-0.02 (-0.43 to 0.39)	-0.29 (-0.80 to 0.23)		
BMI₀₋₁	0.09 (-0.24 to 0.41)	-0.17 (-0.62 to 0.28)		
BMI₁₋₂	0.04 (-0.38 to 0.46)	-0.25 (-0.80 to 0.30)		
BMI₂₋₈	0.18 (-0.26 to 0.63)	0.16 (-0.29 to 0.61)		
BMI₈₋₁₅	0.29 (0.03 to 0.54)	0.63 (0.14 to 1.12)		
BMI₁₅₋₁₉	0.49 (0.20 to 0.79)	0.78 (0.34 to 1.23)		
BMI₁₉	1.07 (0.88 to 1.26)	0.87 (0.58 to 1.16)		
Standardized				
BMI₀	0.06 (-0.45 to 0.58)	-0.13 (-0.85 to 0.59)		
BMI₀₋₁	0.27 (-0.16 to 0.69)	0.02 (-0.77 to 0.81)		
BMI₁₋₂	0.11 (-0.44 to 0.65)	-0.08 (-0.81 to 0.66)		
BMI₂₋₈	0.53 (-0.08 to 1.13)	0.45 (-0.18 to 1.07)		
BMI₈₋₁₅	1.22 (0.76 to 1.68)	1.43 (0.70 to 2.16)		
BMI₁₅₋₁₉	1.36 (0.84 to 1.87)	1.48 (0.87 to 2.09)		
BMI₁₉	2.03 (1.66 to 2.40)	1.89 (1.39 to 2.39)		
Adj R², %	11.96	12.12		

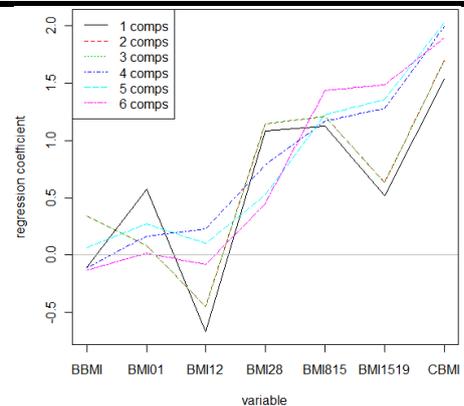
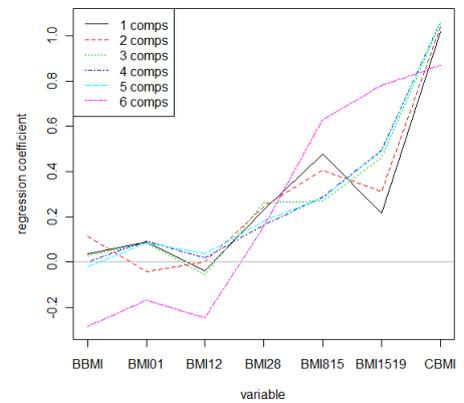


Table 7.7: Results of the PCR analysis with 1-6 components retained.

	PLS			
	1 Component	2 Component	3 Component	4 Component
Unstandardized				
BMI₀	0.01 (-0.07 to 0.10)	-0.08 (-0.41 to 0.25)	-0.30 (-0.92 to 0.32)	-0.28 (-0.85 to 0.28)
BMI₀₋₁	0.07 (-0.04 to 0.18)	0.00 (-0.37 to 0.38)	-0.11 (-0.60 to 0.38)	-0.18 (-0.63 to 0.28)
BMI₁₋₂	-0.04 (-0.12 to 0.04)	-0.07 (-0.39 to 0.25)	-0.18 (-0.93 to 0.58)	-0.23 (-0.83 to 0.37)
BMI₂₋₈	0.22 (0.12 to 0.33)	0.19 (-0.13 to 0.51)	0.11 (-0.51 to 0.74)	0.15 (-0.36 to 0.66)
BMI₈₋₁₅	0.44 (0.31 to 0.57)	0.35 (0.08 to 0.62)	0.60 (0.07 to 1.14)	0.63 (0.13 to 1.13)
BMI₁₅₋₁₉	0.32 (0.21 to 0.44)	0.62 (0.30 to 0.93)	0.76 (0.30 to 1.23)	0.78 (0.33 to 1.23)
BMI₁₉	1.03 (0.85 to 1.21)	1.01 (0.82 to 1.21)	0.89 (0.55 to 1.23)	0.87 (0.58 to 1.16)
Standardized				
BMI₀	0.05 (-0.30 to 0.40)	0.00 (-0.51 to 0.51)	-0.07 (-0.69 to 0.55)	-0.08 (-0.83 to 0.67)
BMI₀₋₁	0.23 (-0.13 to 0.59)	0.12 (-0.36 to 0.60)	0.16 (-0.37 to 0.69)	0.06 (-0.77 to 0.89)
BMI₁₋₂	-0.19 (-0.54 to 0.17)	0.09 (-0.44 to 0.62)	0.02 (-0.63 to 0.66)	-0.08 (-0.84 to 0.69)
BMI₂₋₈	0.86 (0.49 to 1.23)	0.56 (0.01 to 1.10)	0.48 (-0.14 to 1.10)	0.43 (-0.20 to 1.06)
BMI₈₋₁₅	1.27 (0.87 to 1.66)	1.28 (0.78 to 1.79)	1.33 (0.74 to 1.91)	1.41 (0.67 to 2.16)
BMI₁₅₋₁₉	1.06 (0.68 to 1.45)	1.39 (0.85 to 1.93)	1.43 (0.86 to 2.01)	1.47 (0.86 to 2.08)
BMI₁₉	1.91 (1.56 to 2.25)	1.98 (1.61 to 2.36)	1.96 (1.54 to 2.37)	1.91 (1.40 to 2.43)
Adj R², %	12.19	12.35	12.38	12.30
	5 Component	6 Component		
Unstandardized				
BMI₀	-0.29 (-0.83 to 0.25)	-0.29 (-0.80 to 0.23)		
BMI₀₋₁	-0.17 (-0.62 to 0.28)	-0.17 (-0.62 to 0.28)		
BMI₁₋₂	-0.25 (-0.81 to 0.32)	-0.25 (-0.80 to 0.30)		
BMI₂₋₈	0.16 (-0.32 to 0.64)	0.16 (-0.29 to 0.61)		
BMI₈₋₁₅	0.63 (0.14 to 1.12)	0.63 (0.14 to 1.12)		
BMI₁₅₋₁₉	0.78 (0.33 to 1.23)	0.78 (0.34 to 1.23)		
BMI₁₉	0.87 (0.58 to 1.16)	0.87 (0.58 to 1.16)		
Standardized				
BMI₀	-0.13 (-0.85 to 0.59)	-0.13 (-0.85 to 0.59)		
BMI₀₋₁	0.02 (-0.77 to 0.81)	0.02 (-0.77 to 0.81)		
BMI₁₋₂	-0.08 (-0.82 to 0.66)	-0.08 (-0.81 to 0.66)		
BMI₂₋₈	0.45 (-0.18 to 1.07)	0.45 (-0.18 to 1.07)		
BMI₈₋₁₅	1.43 (0.70 to 2.16)	1.43 (0.70 to 2.16)		
BMI₁₅₋₁₉	1.48 (0.87 to 2.09)	1.48 (0.87 to 2.09)		
BMI₁₉	1.89 (1.39 to 2.39)	1.89 (1.39 to 2.39)		
Adj R², %	12.21	12.12		

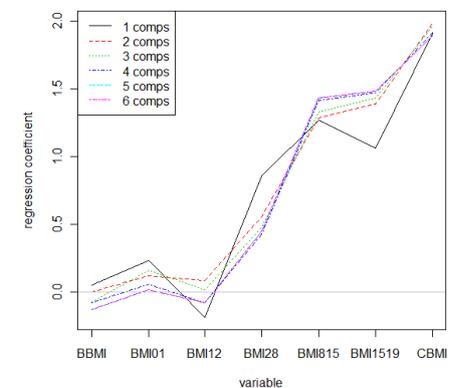
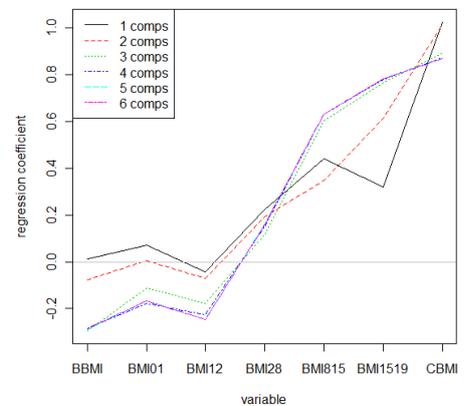


Table 7.8: Results of the PLS analysis with 1-6 components retained.

Summary of Results

The univariate OLS coefficient for BMI_0 was positive and non-significant (0.08 (-0.47 to 0.63)), but became significant upon adjustment for BMI_{0-19} (1.09 (0.55 to 1.63)). The coefficient reversed sign but remained non-significant upon adjustment for BMI_{19} (-0.26 (-0.77 to 0.26)). When six of the seven BMI measures were entered simultaneously (i.e. BMI change along with BMI_0 or BMI_{19}) all of the coefficients were positive. The covariate BMI_{15-19} had the strongest association with SBP in the unstandardized case for both models, whilst BMI_{9-15} became the strongest for BMI_0 and the five BMI change variables for the standardized model. Using eqn(7.2) and eqn(7.36) a unique relationship amongst the OLS coefficients was defined by the dependency inherent within the lifecourse data. This was illustrated by the parallel curves in Figure 7.9 showing that the coefficients are describing a trend in the data.

Whilst the OLS example was forced to consider one fewer than the full complement of predictors, this was not the case for PCR and PLS methods. Components replace the predictors in the estimation and each component is a combination of the original predictors. Therefore, one or more components can be removed from the estimation to reduce the dimensionality of the data and obtain unique estimates on the original axes. When a two component PCR (or PLS) was employed on the three covariates, positive coefficients were obtained for all predictors (although BMI_0 was not significant). Due to the dimensionality of the data, removing one component in either approach would produce the same result as no information is lost when the maximal attainable components are used. However, whilst the coefficients are estimable, the design matrix is still likely to be collinear and experience common features attributed to collinearity (as with any other problem). Therefore, it may be decided to retain only the first component. BMI_0 became negative and significant for PCR (-0.04 (-0.06 to -0.01)), whilst in PLS the same coefficient remained positive and non-significant (0.01 (-0.05 to 0.06)). The change in coefficients between the one and two component models were small for PLS and explained 99% of the original variance.

The analysis was extended to include five BMI change variables in place of the single predictor. The results from PCR are displayed in Table 7.7. For the single component model the coefficients early in the lifecourse are small (with BMI_{1-2} negative) with a fluctuation in magnitude at BMI_{8-15} . The change in R_y^2 is greatest up to the 3 component model. Beyond this, the incremental change is small. For the 3 component model, there was no significant association between BMI_0 , BMI_{1-2} , BMI_{2-8} and SBP. Following this, the

predictors BMI_{8-15} , BMI_{15-19} and BMI_{19} had an increasing significant positive association with SBP. The addition of more components only produced small changes on the coefficient estimates, but the width of the confidence intervals increased with some changes of sign on the coefficients. There is a great variability in the estimates from the PCR plots, indicating that discarding components carries a greater risk of discarding important information. The results of the PLS analysis are displayed in Table 7.8. In the single component model the results follow a similar pattern to PCR with 3 components retained. The coefficients in the early lifecourse (i.e. BMI_0 , BMI_{0-1} and BMI_{1-2}) have no significant association, whilst those later in the lifecourse (i.e. BMI_{2-8} , BMI_{8-15} , BMI_{15-19} and BMI_{19}) demonstrate a positive association. Moving to the 2 and 3 component models BMI_{2-8} becomes non-significant. However, it is noticeable that increasing the number of components does little to effect the pattern of the coefficients. The variation is explained predominantly in the first two components and there is no association between the early measurements (i.e. BMI_0 , BMI_{0-1} and BMI_{1-2}) and SBP, whilst the effects of growth generally increase with age up to BMI_{19} . Including multiple measurements of the lifecourse will increase the information, however it also increases collinearity and so PLS becomes a very useful tool to reduce the dimensionality of the data further than the necessary degrees of freedom. Employing PLS can extract signal whilst discarding noise in the data, although this is directly related to the 'quality' of the data.

A final word of caution must be made with regard to the original discussion of the problem. In many ways the problem has been simplified in that each of the predictors are assumed to have a unique and independent contribution and are interpreted as such. However, this contradicts our conceptual understanding of the problem and thus promotes a simple application of methods such as PCR and PLS for such application. The path model in Figure 7.1 would define current weight as a mediator of BMI_0 to CD. Therefore, if such a causal model is adopted the results must be carefully analysed. Following the discussion in section 3.2, the three coefficients from the causal model would not be interpreted. The total effect of BMI_0 on SBP (i.e. unadjusted) may be interpreted or BMI_0 considered as a confounder of BMI_{19} . This analysis has broken away from the causal model that was used to define the problem, however as always the conceptual understanding of the problem must underpin the statistical analysis. The analysis could be extended to methods such as partial least squares path modelling (PLS-PM) to incorporate these causal ideas, however this is beyond the scope of the current work.

7.9 Discussion

The constraint in PCR and PLS is not being imposed on the solution and so would seem to provide the most simple interpretation for a solution. The relationship is directly maintained for all of the coefficients in the unstandardized model. It is also true if the differential weighting is applied as standardized predictors (as defined in eqn(7.30)). However, it remains that any constraint that has been imposed on the estimation must reflect that of the population model exactly to obtain unbiased estimates (Kupper et al. 1983). Even with knowledge of the population this is an unachievable aim. Methods such as the lifecourse plot will typically constrain BMI_0 or BMI_{19} to zero. Such a solution could only be fully justified if it was reflective of the population constraint. Any deviation from this “truth” will generate a bias on the estimation.

Shrinkage methods will input some bias into the estimation when $m < k$ components are retained (unless the constraint employed is correct), however this seems less of a problem when unbiased estimates are seemingly unachievable. A PCR or PLS with maximal components retained (i.e. $m = p - 1$ for the lifecourse example), will produce the same estimation, as no information is discarded in the process. However, ‘solving’ the identification problem in this manner does not avoid the remaining collinearity. The design matrix is still likely to be near singular. When an increasing number of predictors are entered into the model, such as in the extension task, the degree of collinearity will increase. This is only natural as neighbouring BMI measures are broken down into finer and finer measurements. We are then likely to experience common ‘symptoms’ of collinearity such as a changes of sign, inflated variance or magnitude of the parameter estimates. Therefore, shrinkage methods such as PCR and PLS provide a useful extension to the MP-inverse and related linear restrictions in that fewer components can be retained to further reduce the dimensionality of the data, but the ‘natural’ constraint maintained.

It must be decided how many components to retain. In the 3 predictor example this decision was justified using the high variance explained in the single component model. With the goal of maximizing variation in the components, the decision in PCR is generally defined based on measures such as eigenvalues. This may involve scree plots, parallel analysis or some arbitrary rule of thumb such as Kaiser’s criterion (as discussed for PCA in section 6.4.1). PLS looks to maximize covariance. Using measures based on explained variance (such as R_y^2) would seem a natural extension. A range of methods have been developed for this purpose. One such measure is labelled the “predictive residual

error sums of squares" (or PRESS) (Allen 1974). This is found by splitting the data into groups. A model is formulated on one group and tested on the other. PRESS seeks the differences between the predicted and observed values. When the PRESS statistic is low, the model has a 'high internal consistency'. The PRESS statistic calculates residuals for each data sample and so effectively represents a cross-validated RSS. This value is used in place of RSS to calculate a cross-validated R_y^2 . Whilst this statistic is based on the predictive capability of the model, this does not necessarily demonstrate the usefulness of the coefficients for interpretation purposes.

It would be beneficial if an optimal number of components to retain could be decided based on something other than a predictive guide. For instance, changes of sign on the estimates are a common feature of collinear data. It is well known that the full component model gives the same result as the OLS estimate. As the number of components retained is reduced, a single component solution in the direction of a simple regression would be reached (see Figure 5.2). This may be used as a guide based on the number of sign changes away from simple regressions. Whilst this would seem a rather rough guide (with the potential for much criticism), it would depart from relying on a predictive index. Comparing a vector in the direction of simple regression to an OLS result was also the basis for our D-index developed in chapter 5. With some adaptation the index could be used as an aid for PLS. The vectors could represent PLS solutions from differing numbers of components and the angles of projection adjusted for each additional component. This index could allow a judgement to be formed on the balance between explained variance and the impact of collinearity between components.

This chapter has focussed on the statistical aspects of PCR and PLS to illustrate the constraints that are made by the methods and the computation of the coefficients. However, variations of these approaches could be employed such as partial least squares path modelling (PLS-PM) to incorporate the path modelling ideas (see section 3.2.1). Alternative approaches can also be identified that would be analogous to current methods in the lifecourse literature. Employing PLS on the original lifecourse variables (i.e. not BMI change) would be similar to the lifecourse plot that utilized OLS estimates. Similar to the original approach the user would then look for steep changes in the estimates (considering different numbers of components) to demonstrate a critical period in the lifecourse. Raw predictors could also be included in addition to change. The implementation of this method would be dictated by the available degrees of freedom deciding how many components could be retained.

7.10 Conclusions

The identification problem of perfectly collinear covariates is not one that is just limited to lifecourse epidemiology. There are numerous examples in the social sciences, economics, bioinformatics etc. The theme to this problem is that there is no single solution to the ill-defined matrix, instead there are infinitely many (as demonstrated by the g -inverses) which demonstrate an equal model fit. In studies such as the Cebu example, the researcher is effectively seeking a distortion of an unattainable 'truth'. As such, numerous proposals for computation of the coefficients have been suggested, but many appear to have little interpretational benefit. Whilst it appears unrealistic to search for a single 'solution' to the perfect collinearity problem (which would be impossible to prove) the researcher should look to find one that is interpretable and justifiable for the problem.

This chapter has considered why an additional constraint is required for a 'direct' solution in the lifecourse model and the options available to the researcher in implementing a linear restriction. The restriction that would appear most suitable (without incorporating external information) is that maintained by the MP g -inverse. This is because the corresponding constraint (i.e. 1, 1, -1) is an intrinsic condition of the lifecourse data (i.e. $BMI_0 + BMI_{0-19} - BMI_{19} = 0$). This particular inverse is equivalent to the PCR shrinkage regression method. General theory on the PCR estimator can also identify a fundamental limitation of applying this methodology. The components attempt to maximize only the variance amongst the predictors and thus ignore the association of the response with the predictors.

An extension to the PCR approach can be found in the use of PLS. PLS is focussed on maximizing the covariance between \mathbf{X} and \mathbf{y} and as such estimates are obtained on rotated axes to achieve this feature. Whilst a discussion such as this highlights justifiable criteria for selecting a methodology, one researcher may argue that simply dropping the covariate with the smallest regression coefficient is as good as any approach. Due to the nature of the problem PLS can be promoted as the most justifiable approach, but it would be incorrect to argue it as the 'best' based on providing the most accurate estimation. Chapter 8 considers a small selection of methods in the age-period-cohort literature to bypass the singularity problem. There are also additional difficulties when studying the critical phase of growth that have been ignored in this chapter to simplify the mathematical problem (e.g. measurement error, sampling variation, non-linear terms). The effects of these factors are also considered.

8. Perfect Collinearity in the Age-Period-Cohort Model

A second example of perfectly collinear covariates occurs in the analysis of age-period-cohort (APC) data. The APC model studies the effects of three temporal covariates – age, period and cohort – on disease incidence and mortality rates of various health outcomes. These rates are usually presented in the form of a two way table of age versus period. Birth cohorts would be identified on the diagonals of the table. These variables once again present a perfect collinearity, such that $\text{Age} + \text{Period} = \text{Cohort}$. In the APC analysis the problem is changed to consider the data as interval variables. In this case there is more than one example of perfect collinearity. The categories of age, period and cohort will present an exact dependency within each group in the dummy coding (as regularly seen in categorical regression), whilst there will be an overall perfect dependency across the categories (analogous to the continuous lifecourse problem).

Early approaches to the APC problem focussed on graphical techniques. Whilst these methods provide a direct description of the data, they are limited in studying individual effects. The work followed to regression analysis to provide quantitative estimates of the effects. The methods proposed in the literature generally focus on bypassing the identification problem (see section 7.4) by imposing constraints on the coefficients or by considering ‘estimable’ effects. In the former, the interpretation of the results must reflect the constraints imposed. In contrast, whilst linear effects are perfectly collinear, the deviations from linearity are not reliant on the choice of constraint. Chapter 8 provides a brief review of the development of influential methods in the APC literature and considers some of the more recent proposals to alleviating the identification problem. The novel application of latent variable methods such as PCR, PLS and the intrinsic estimator (IE) are considered using a recent APC dataset. The work looks to identify the constraints imposed on the APC solution and the rationale for choosing one over another. The discussion is intended to identify a place for these methods in the APC literature.

8.1 Definition of the Variables and APC Data

APC analysis in epidemiology is used to model and forecast trends of disease incidence and mortality rates. The model attributes risk to three influences; (1) an age effect, (2) a period effect and (3) a generational (or cohort) effect. Each of these variables have historically been considered important by epidemiologists (Holford 1992; Mortimer and Shanahan 2006). In particular, the influence of birth cohorts has long been cited as an important determinant of later life disease (Fienberg and Mason 1982; Frost 1939; Kermack et al. 1934). APC analyses have been used in a wide range of fields such as demography, sociology and economics. Epidemiologists will look to improve understanding of the aetiology of a health outcome and to identify important cohort effects that lead to a long lasting impact on disease risk. The definition of a cohort effect rests on the definitions of period and age effects. An age effect reflects a particular age (or stage in the lifecourse) in which physiological change or accumulation of exposure to some risk factor or social influence are directly associated with the process of aging (Keyes et al. 2010). A period effect represents a change in the incidence rate that may be the result of an environmental, medical or technological change impacting on subjects of all ages. A cohort effect is present when the incidence rate of a health outcome changes due to subjects exposed to new or changing risk factors in different time periods affecting various age groups differently. In epidemiology a cohort effect is often viewed as a type of restricted age by period interaction (Kupper et al. 1983).

APC data is usually presented in the form of a two-way table of rates ($R_{ij} = O_{ij}/N_{ij}$, where O_{ij} is the observed response at age i and period j (i.e. cases) and N_{ij} is the total number of subjects at risk within this time (i.e. person years)). This table is labelled a 'lexis diagram', with age groups forming the rows and period groups the columns of the table,

		Period Interval				
		$j = 1$	$j = 2$	$j = 3$	$j = 4$	
Age Interval	$i = 1$	R_{11}	R_{12}	R_{13}	R_{14}	Cohort 8
	$i = 2$	R_{21}	R_{22}	R_{23}	R_{24}	
	$i = 3$	R_{31}	R_{32}	R_{33}	R_{34}	Cohort 6
	$i = 4$	R_{41}	R_{42}	R_{43}	R_{44}	
	$i = 5$	R_{51}	R_{52}	R_{53}	R_{54}	Cohort 5
		Cohort 1				

Table 8.1: An example lexis diagram.

The rates of the chosen response (R_{ij}) are displayed in the cells of the table. There are a total of $i = 1, \dots, I$ age groups, $j = 1, \dots, J$ periods of time and therefore $k = 1, \dots, I + J - 1$ birth cohort groups. The cohort groups (shown in red) are found diagonally in the table, with the oldest cohort in the bottom left and the youngest in the top right. The data obtained for APC studies are often vital statistics which are readily available and easily digested, but limited in detail. The variables listed may be assumed to represent latent constructs. For instance, the definition of age is a generalisation of what is actually observed. Individuals will typically age both physiologically and socially at different rates (Hobcraft et al. 1982). Also, a variable such as socio-economic status (SES) would appear to be a strong candidate for adjustment in such a model. A technological advancement or change in health care may be developed in a particular period of observation; however such an improvement may not be readily available to the less wealthy or those living in poorer areas. This could not be attributed directly to a period or cohort effect. The observable measure that is utilized for period and cohort are similarly likely to be poor representations of the latent constructs that they are assumed to represent.

By the nature of the intervals selected, the cohort groups often overlap. For instance, if $i = 1$ represents the interval 25-29 and $j = 1$ the period 1985-1989, then the birth cohort group attributed to this cell will contain subjects born 1956-1964. Assuming equal interval widths, the following cohort with age 25-29 during the period 1990-1984, will contain subjects born 1961-1969. Therefore, a subject can move between neighbouring cohorts as they age (i.e. they are not uniquely assigned to one cohort). When date of births are known, it would be possible to incorporate a finer grid to produce a uniquely defined cohort group such that this ambiguity is lost (this is described further in section 8.5.3). It is also common for the cohorts in the extremes of the table to contain very few observations in relation to the intermediate cohorts (Kupper et al. 1985b).

It is noticeable from the lexis diagram that the groups are not balanced. The number of age and period groups may be specified to be equal, however cohort would naturally be unbalanced (i.e. $I + J - 1$ cohorts for I age and J period). This feature of the data can be of importance to the analysis and is one that may easily be understated or hidden in the results of a statistical analysis (see section 8.6.2). Also, the predictors are of interval categorical form, whilst they represent continuous time effects. Therefore, it may be argued that "cross-time" effects are hidden when considering a coarser grid of data. This will play an important role when discussing various solutions and tabulation on the potential impact of collinearity on estimates computed.

8.2 Example Application

The dataset considered in this chapter is a recent example published by Dahlquist et al. (2011). The study considers onset of type-1 diabetes in Sweden between 1983 and 2007.

	1985	1990	1995	2000	2005
2	17.6 (217) 17.2 (201)	16.7 (241) 14.9 (204)	22.8 (335) 22.4 (313)	27.4 (328) 24.3 (276)	28.1 (360) 25.0 (303)
7	28.1 (352) 28.8 (343)	28.6 (361) 27.6 (331)	31.2 (462) 35.9 (504)	39.8 (590) 40.0 (564)	46.4 (570) 53.3 (621)
12	32.0 (448) 33.1 (442)	37.1 (473) 31.0 (376)	38.7 (502) 33.9 (417)	42.3 (635) 38.5 (547)	59.2 (894) 46.6 (670)
17	20.2 (301) 12.4 (176)	20.0 (285) 12.3 (166)	17.1 (223) 11.9 (148)	22.2 (293) 11.9 (149)	20.4 (312) 12.4 (180)
22	18.2 (273) 9.2 (132)	15.1 (232) 11.1 (163)	15.7 (229) 9.6 (135)	18.3 (243) 11.3 (144)	17.4 (238) 9.8 (128)
27	16.2 (232) 9.8 (135)	18.0 (283) 7.9 (118)	14.4 (229) 8.2 (125)	15.8 (235) 8.6 (124)	14.6 (204) 6.2 (83)
32	15.0 (224) 6.9 (99)	13.9 (206) 6.7 (94)	12.9 (208) 6.1 (93)	11.5 (185) 6.3 (97)	10.6 (164) 5.4 (81)

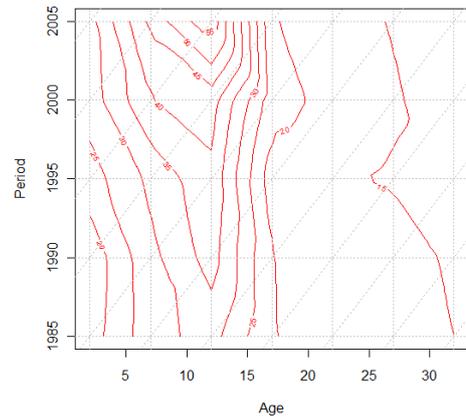
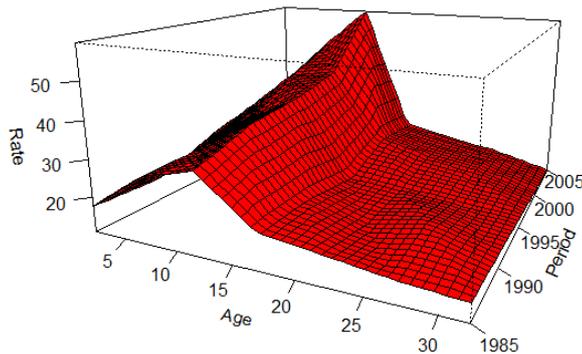
Table 8.2: Cumulative incidence at 35 years in Sweden during 1983-2007, with absolute numbers of cases in brackets for males (standard font) and females (bold font).

The incidence rate of childhood onset type-1 diabetes has been reported to be increasing worldwide, although the reasons for this trend are not well known (Pitkaniemi et al. 2004). Scandinavian countries, such as Sweden and Finland, have reported the highest increases over the past 20 years (Dahlquist and Mustonen 2000). Type-1 diabetes is a disease that destroys insulin producing beta-cells of the pancreas. This produces a shortage of insulin which generates an increased blood and urinary glucose. A decreasing trend has been identified for young adults (>15 years), which suggests a shift in risk towards younger age groups. The nationwide study by Dahlquist considered 20,249 subjects followed from birth to 34 years. There are 7 categories for age at diagnosis and 5 categories for period at diagnosis, generating a total of 11 birth cohort groups (i.e. $7 + 5 - 1 = 11$). Each age and period category represents a mid-point measure from an interval width of 5 years. Cohort groups (i.e. diagonals of the table) follow a particular generation of subjects sharing similar birth years. The data spans a time period of 25 years and a wide demographic area and as such could reflect a number of potentially important period and cohort influences.

8.3 Graphical Analysis

Analysis of APC data in the early 20th century began with the use of descriptive graphical tools to understand the trends of the data. A 3D surface and contour plot of the Dahlquist data in Table 8.2 is constructed in Figure 8.1.

(a) Male Sample



(b) Female Sample

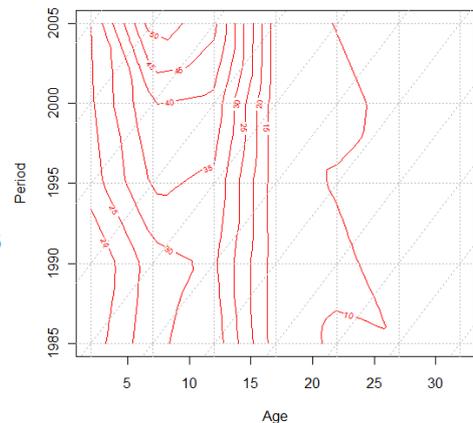
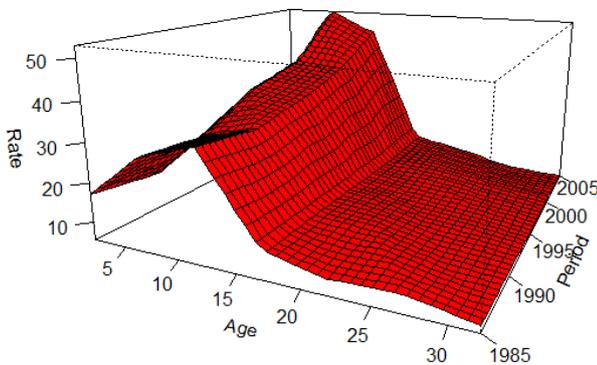
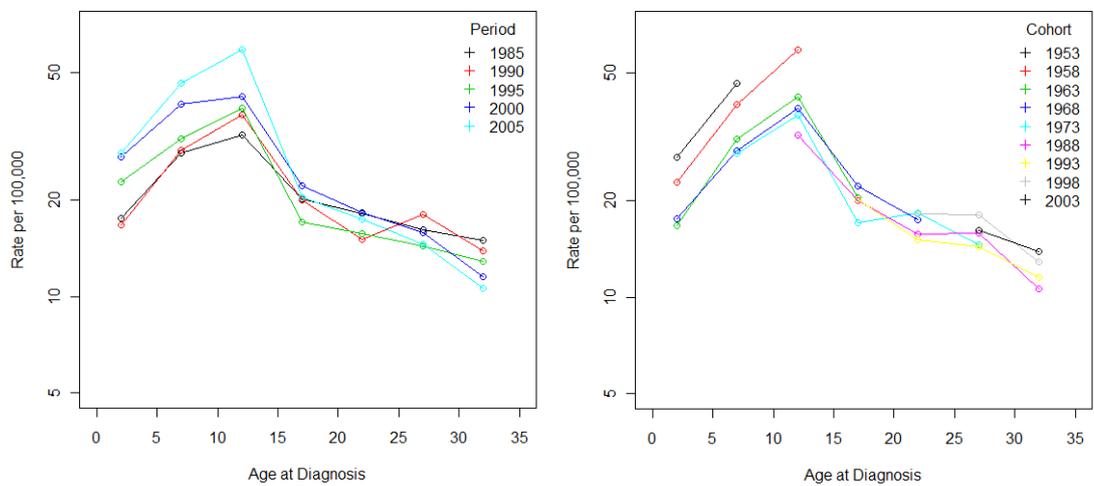


Figure 8.1: Diagnosis rates of males as (a) a response surface and (b) contour plot.

Due to the nature of the 3D plot (and APC data in general), the user is limited to controlling only two of the three variables. Generally, these plots have considered age and one of cohort or period in clinical examples. One of the reasons is that the physiological change is based in biology, which presents a more solid base for making forecasts than variables that are generally sociological concepts such as cohort and period (Fienberg and Mason 1982). The plots in Figure 8.1 consider age and period; this means that cohort effects are allowed to vary across the surface. Whilst 3D plots give an overview of the overall trends in the data, it is difficult to identify subtle changes (Holford 1992).

The limitations of the 3D surface plot can be bypassed to some extent by projecting the rates onto a 2D surface, such as a contour (or level) plot (see Figure 8.1b). The cohort groups are illustrated by the grey dashed diagonal lines. For both groups the contour lines are generally parallel with the y-axis and so it can be observed that age-specific risk is not changing rapidly over time. For the male sample, the incidence rate has peaked at around age 14 (which appears consistent across periods). The female incidence rate is initially similar to the males, however in later periods the rates peak at an earlier age (approximately 9 years). A general decrease is observed for both groups as the subjects enter early adulthood.

(a) Male Sample



(b) Female Sample

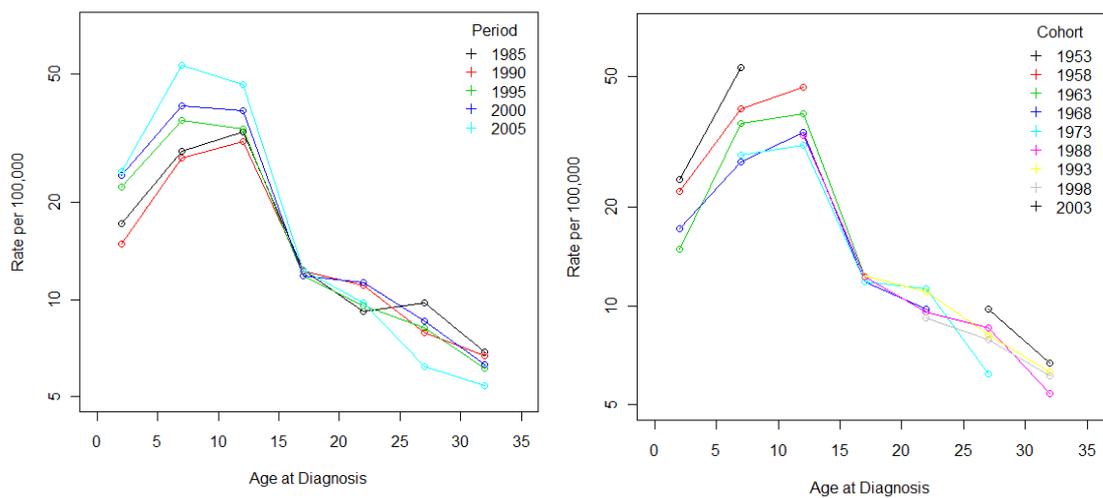


Figure 8.2: Age specific plots stratified by period and cohort.

From the contour plots there does not appear to be a strong cohort influence with the contours generally cutting through the cohort groups. However, the scaling of the groups in these plots makes direct observation difficult. As it is a disease rate that is being described in this example, age is likely to play a key factor and so most clinical studies will report age-specific plots (see Figure 8.2). Splitting by period gives 'cross-sectional' age-specific rates, whilst by cohort gives 'longitudinal' age-specific rates (Carstensen 2007). A general pattern observed in early graphical APC analysis was that mortality rate at a younger age would demonstrate a fall at an earlier cohort than those at a later age. This could suggest a potential cohort effect early in the lifecourse such that the impact is only realised when the birth cohort reaches its later years (e.g. FOAD, DOHaD). Kermack (1934) suggested that in some cases mortality rate of the child is dependent on the health of the mother, again suggesting a potential generational effect (Smith and Kuh 2001). If stratified cohort groups follow a parallel trend then it would suggest that neither hypothesis is likely. This has led some to suggest that non-parallelism of the lines in contour and age-specific plots demonstrates the potential for an interaction effect (Kupper et al. 1985b).

The plots illustrated in Figure 8.2 demonstrate a greater variation when stratified by period although it is not clear which variable is producing the greater effect. If the plots were overlapped, the cohort curves would cut through period curves. For both sexes, a general trend is observed as an increase in incidence up until early teenage years before a rapid decline to age 16 years followed by a very gradual reduction beyond this point. For males it appears that the trajectories for the early cohorts follow a similar pattern, although for the more recent 1993 and 1996 birth cohorts the rates look to be increasing greatly. This is also illustrated in stratification by period, with the more recent periods 2000 and 2005 demonstrating an inflated maximum at around age 12. In comparison, this inflation is occurring at an earlier stage in the female subjects. There is a change to an earlier maxima of age 7 (or the interval this mid-point represents) for periods 1995 and later. Dahlquist (2011) reported "a cohort effect dominating over a period effect" in this data and "a shift to a younger age (in incidence rate)". From the graphical information it would seem difficult to justify the first claim, although it certainly does not disprove it either. It would seem that trajectories are relatively constant until more recent cohorts which demonstrate an inflated incidence rate. For females, there appears to be a shift in maxima to an earlier age group when stratified by period. From the descriptive analysis it appears that caution must still be placed on the claims made in the original paper. An analytical assessment of the data is needed to strengthen these observations.

8.4 Regression Analysis

Early examples of APC analysis would typically look at a two predictor model (usually age and birth cohort), with age-specific plots sufficing from a descriptive perspective. The graphical analysis could then be extended to a quantitative assessment of the effects using full rank regression methods to obtain statistical estimates free from any model identification problems. In the early 1980's researchers began to place an increased emphasis on quantifying the additive effects of age, period and cohort predictors (Holford 2005). From a statistical perspective it may appear to be a nonsensical task to include the predictors as separate additive effects. However, from a conceptual viewpoint it may be argued that individuals born at the same time and experiencing some common aging effect, not only respond to period effects, but also to an additional impact of birth cohort.

Keyes (2010) presents one definition sourced from a “predominantly sociological” viewpoint regarding the role and potential importance of the birth cohort. A focus is placed on birth cohort as the main exposure for explaining patterns in mortality rate and disease incidence. They would hypothesize that it is a reflection of the cohort to which you are born into, such as the “conditions, barriers and resources”, that determines trend and incidence of health outcomes in later life. The research question becomes focussed on determining the risks associated with belonging to a particular birth cohort. What are seen as short term effects brought about by physiological processes of aging and variation of environmental factors are considered to be masking a generalized effect of birth cohort. Therefore, age and period are seen as confounding effects of this main effect.

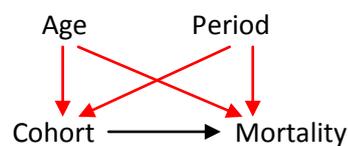


Figure 8.3: Path diagram of the sociological interpretation.

This definition would seem to suggest only interpreting the cohort effect when each of the predictors are entered as additive effects (see section 3.2.1) – similar to the lifecourse example. Definitions such as the epidemiological and sociological presented here are vital to the analysis of APC data to understand what the research question is asking and the potential to achieve generalizable results (Fienberg and Mason 1982).

Another interpretation of the APC data provided by Fienberg and Mason (1982) is that birth cohort carries a general group effect that impacts on the individuals in the study (i.e. the cohort effect being multi-level in nature). However, due to the aggregate forms of the data, verification of such concepts is rarely achievable. Fienberg provides the example of a hypothesis suggesting that cohort size is an important factor at the macro level impacting on factors such as crime, employment and divorce rates at an individual level. However, to test such a theory would require these rates at a micro level and not simply aggregated data. The research question must depend on what is achievable with the data provided. Any statements made from the analysis (suggesting macro and micro effects) rest on the definitions of the (potentially latent) variables entered. APC is by definition, a broad, wide-ranging concept so the ultimate goal of the analysis, with the grouped data provided, should perhaps in turn remain fairly broad.

A regression model can be considered with the effects of age and period included along with an interaction between the variables,

$$y_{ij} = \beta_0 + \beta_{Ai}X_{Ai} + \beta_{Pj}X_{Pj} + \beta_{ij}X_{Ai}X_{Pj} + \varepsilon_{ij} \quad (7.37)$$

In eqn(7.37), β_0 denotes the intercept, β_{Ai} the I fixed effects for age, β_{Pj} the J fixed effects for period and β_{ij} the IJ fixed effects specific to age i and period j . There are a total of IJ available degrees of freedom (df) for the model. There is 1 df required to estimate β_0 , $(I - 1)$ df to estimate β_{Ai} , $(J - 1)$ df to estimate β_{Pj} and $(I - 1)(J - 1)$ to estimate β_{ij} . This equates to a total of $(I + 1)(J + 1)$ model parameters required with only IJ df available, leaving the model non-identifiable. The “zero-sum” constraints often used for ANOVA may be applied such that,

$$\sum_{i=1}^I \beta_{Ai} = \sum_{j=1}^J \beta_{Pj} = 0 \quad (7.38)$$

$$\sum_{i=1}^I \beta_{ij} = 0 \text{ (for each } j\text{)}$$

$$\sum_{j=1}^J \beta_{ij} = 0 \text{ (for each } i\text{)}$$
(7.39)

(Kupper et al. 1983)

Applying these constraints means that only $(I - 1)$ age and $(J - 1)$ period effects need estimating, reducing the necessary df both by 1, whilst the number of interaction effects requiring estimation are reduced by $(I + J - 1)$ leaving $(I - 1)(J - 1)$ to be estimated. For any constraint applied, any one of age, period or an interaction effect is naturally defined as the negative sum of the rest in that category (e.g. $\beta_{AI} = -\sum_{i=1}^{I-1} \beta_{Ai}$) – i.e. effects coding. For dummy coding, the mean may be subtracted from each of the categories and the estimated effects interpreted as a deviation from the category mean to which it belongs (i.e. analogous to estimates of continuous predictors after centering). This maintains the natural structure of the data, creating no distortion (Kupper et al. 1985b). However, whilst this negates perfect collinearity within the categories for main and interaction effects, all of the df available to the model are used estimating main and interaction effects, resulting in a saturated model (i.e. no df available to estimate random error ϵ_{ij} , giving an automatic $R_y^2 = 1$). A further constraint would be necessary to release another degree of freedom.

APC models are often formulated to assume that cohort can stand alone as an additive effect separate of age and period (Keyes et al. 2010). The commonly used APC multiple classification model (Mason and WINSBORO.HH 1973) is a three factor ANOVA-type model used to identify unique effects of the three predictors,

$$y_{ij} = f(R_{ij}/N_{ij}) = \beta_0 + \beta_{ai}X_{Ai} + \beta_{pj}X_{Pj} + \beta_{c(I-i+j)}X_{Ck} + \epsilon_{ij} \quad (7.40)$$

Whilst still assuming constant effects for age (for each i) and period (for each j) as before, sets of interaction effects are now assumed to be constant relating to specific cohort groups (i.e. the cohort k). The zero sum constraints can once again be applied as demonstrated in eqn(7.38), whilst eqn(7.39) would now be replaced by the following,

$$\sum_{k=1}^{I+J-1} \beta_{ck} = 0 \quad (7.41)$$

This is only one constraint for the interactions, which means that the number of cohort groups to be estimated is reduced by 1 to leave a total of $I + J - 2$. Similar to the previous definition for the constraints on the additive effects of age and period, this has allowed for one of the effects to be defined as the negative sum of the first $I + J - 2$ cohort effects in the effects coding (i.e. $\beta_{c I+J-1} = -\sum_{k=1}^{I+J-2} \beta_{ck}$).

A further constraint is still required to obtain unique estimates. This is shown in Kupper (1983;1985b) by collapsing the categories by component using orthogonal polynomials to assess the linear effects,

$$\sum_{i=1}^{I-1} \left[i - \frac{I+1}{2} \right] \mathbf{X}_{Ai} - \sum_{j=1}^{J-1} \left[j - \frac{J+1}{2} \right] \mathbf{X}_{Pj} + \sum_{k=1}^{I+J-2} \left[I - i + j - \frac{I+J}{2} \right] \mathbf{X}_{Ck} = 0 \quad (7.42)$$

Similarities can be observed with the lifecourse model examined in chapter 7. It is clear that the design matrix is not full rank, even after zero sum constraints have been applied. Kupper et al. (1985b) also highlights that these polynomials represent the linear component of the categories in the model, hence it is only the “slope” estimates that are perfectly collinear. This provides the motivation for some to study “estimable functions”. Some effects or combinations of effects can be labelled “estimable”, which means that they are invariant to the constraint applied (Holford 1992;Holford 2005;Rodgers 1982) – see section 8.5.1.

For the multiple classification model in eqn(7.40), \mathbf{y}_{ij} represents some linear function of the incidence rate. Count data from a contingency table are often modelled using a generalized linear model (GLM) with poisson error and a log link. This type of model has been widely used in fields such as demography and epidemiology (Yang et al. 2008). They could similarly be modelled by other count models such as the negative binomial or the zero inflated poisson if the application suggests these to be preferable (i.e. the response is over dispersed or contains a large number of zeros respectively – e.g. modelling cases for a rare disease). By taking the anti-log of the coefficients, relative risk type effects are produced for age/period/cohort. It is necessary to enter the logged ‘person years’ (i.e. the quantity in parentheses in Table 8.2) as an ‘offset’ term. The poisson likelihood is based on the cases O_{ij} , but it is the rates R_{ij} that the researcher wishes to model, therefore these rates are adjusted as follows in the poisson model.

$$\log(R_{ij}) = \log\left(\frac{O_{ij}}{N_{ij}}\right) = \log(O_{ij}) - \log(N_{ij}) \quad (7.43)$$

The logged person years N_{ij} is entered as an offset term on the right hand side of the model with the coefficient set to equal unity. The probability model assumed (poisson or otherwise) takes on a particular importance as it will impact on any interaction effect observed both in graphical and analytical work (Holford 1985).

8.4.1 Geometry of the Solutions

A discussion of the vector space geometry for perfect collinearity was presented in section 7.6.2. A null vector $(1, 1, -1)$ was highlighted which defined the plane P_1 in which the first two eigenvectors must lie for PCA (i.e. that capture all the variance in \mathbf{X}) (see Figure 7.5). The plane is fixed along with the null vector by the APC (or lifecourse) intrinsic relationship, whilst the first two eigenvectors can vary within this plane. The null also defines that the normalized third eigenvector is $(\sqrt{3}, \sqrt{3}, -\sqrt{3})$. O'Brien (2011) produces a similar illustration of the perfect collinearity problem and the solutions produced by various g-inverses. In this paper he uses a 3-dimensional parameter space spanned by the coefficient solutions to demonstrate the ideas.

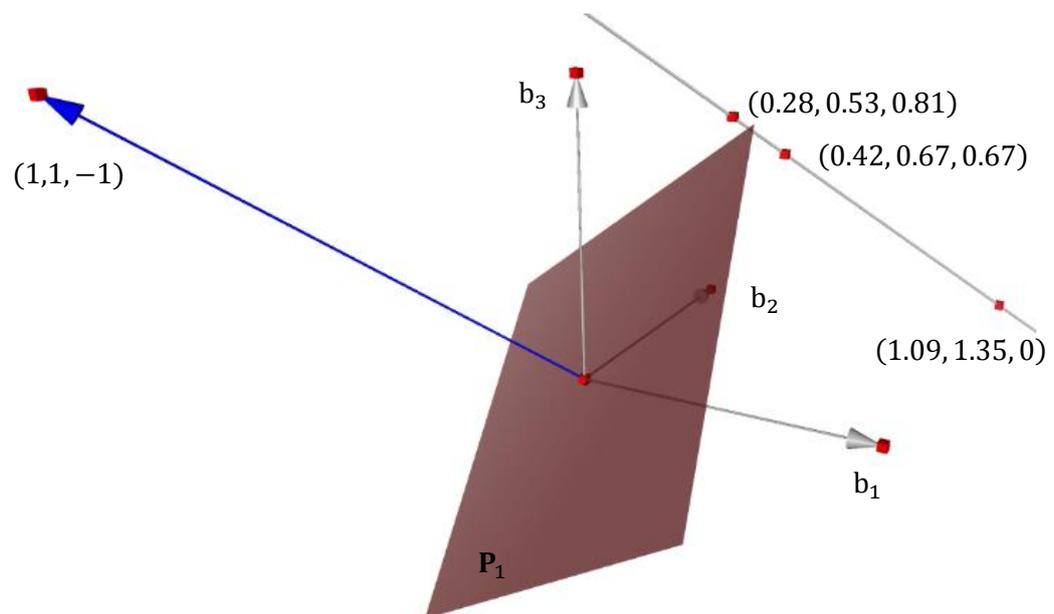


Figure 8.4: Solution space geometry for the Cebu lifecourse example.

The vector geometry in Figure 8.4 shows the three solutions that were attained for the Metro Cebu data using the three linear constraints from section 7.6.1. The MP g-inverse constraint is in the direction orthogonal to the null vector (shown by the orthogonal solution plane). This dictates that the dot product of the solution and the null (i.e. the blue vector) will always equal zero (as noted for the PCA solutions). The two remaining solutions, representing the constraints $b_3 = 0$ and $b_2 = b_3$, form part of a common “line of solutions” with direction parallel to the null vector. This line of solutions represents least squares estimates from various g-inverses.

The null vector is a combination of values such that when multiplied by the columns of \mathbf{X} results in a column vector of zeros (O'Brien 2011). In the lifecycle example, this combination is indicated by the third eigenvector which captures zero variance in \mathbf{X} (or more generally the final eigenvector under a single perfect dependency). Consider the normal equations in eqn(7.13), it is clear that the null vector in this case is $(1,1,-1)$ (or some scalar multiplication of this vector – e.g. $\sqrt{3} \cdot (\sqrt{3}, \sqrt{3}, -\sqrt{3})$). The existence of this single null vector demonstrates that the design matrix is only one fewer than full rank (i.e. the kernel space). Each of the solutions formed from generalized inverses are least squares estimates in that all minimize SSR under the constraint assumed. Thus the estimated coefficients produce the same model fit and are unbiased (under the constraint applied). From linear algebra the difference between any of these estimators lies in the null space (hence the estimators all falling on the same line of solutions). Therefore, model fit will not determine the appropriateness of the constraint.

The population (or generating) set of parameters must lie at some point along this line of solutions, but assumptions must be made on the estimation. Following the definition of Mazumdar (1980), the constraint \mathbf{c} must be orthogonal to the vector of coefficients to produce unbiased estimates ($\mathbf{c}^T \mathbf{b} = 0$). This property is naturally translated to the geometry.

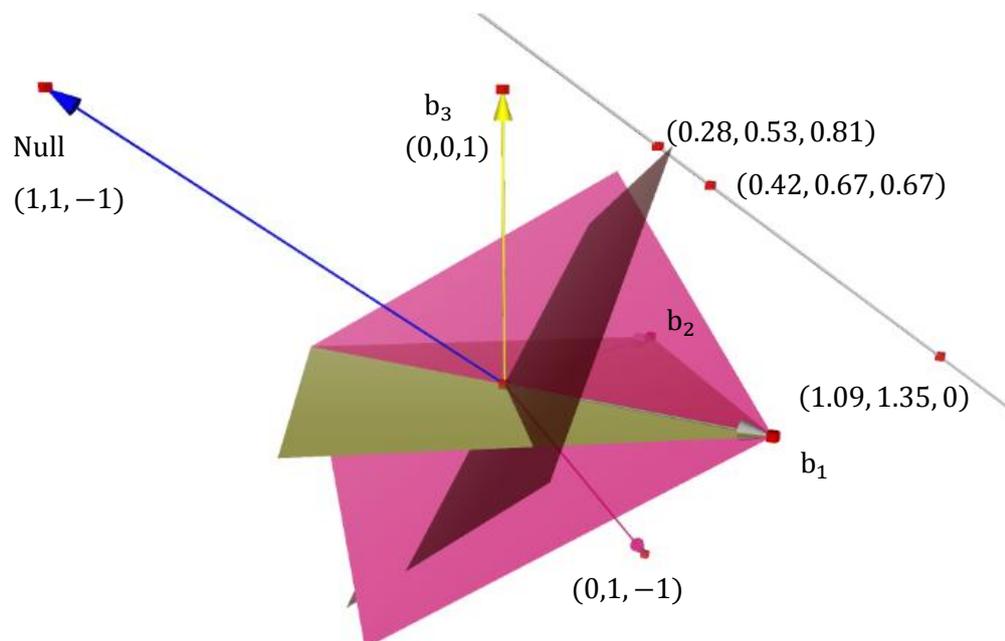


Figure 8.5: Solutions planes for the linear constraints used in the Cebu example.

In Figure 8.5 the equal constraint vector (0, 1, -1) (shown in pink) and the zero constraint vector (0, 0, 1) (shown in yellow) have been projected to demonstrate the orthogonality of the solution plane with the null vector. The planes that are shown orthogonal to the constraints are labelled “solution planes”. In this sense, the PCR solution is no more guaranteed to be reflective of the population parameter than any other g-inverse. If the constraint does not perfectly reflect the population parameters, the resulting estimates will be biased. Importantly, it is noticeable from the geometry that the bias introduced to each parameter follows the equation of the null vector (by parallelism of the lines). Moving from one constrained estimate to the next is simply a movement along this line of solutions. For instance, consider that b_3 is zero in the generating model, then the g-inverse that relates to the linear constraint (0, 0, 1) is an unbiased estimator. In contrast, the estimate from the MP g-inverse is then biased. However, the bias is only in a single dimension (along this line of solutions) and so there is only one indeterminate parameter (labelled w) that produces the bias in the estimates,

$$\beta_A = \hat{\beta}_A + w \qquad \beta_P = \hat{\beta}_P - w \qquad \beta_C = \hat{\beta}_C + w$$

For example, the true population parameters would be (1.09, 1.35, 0). A transformation by -0.81 (1, 1, -1) would provide the MP-inverse solution. The quantity 0.81 represents a type of bias in this sense, although the scaling is dependent on the length of the null vector (which can be any scalar multiple). These relationships show that if the population parameter of any one of age, period or cohort is known, then the analyst would automatically know the remaining two. It is also possible to consider linear combinations of the estimates. For instance summing β_A and β_P would remove the indeterminate parameter from the estimation and the resultant estimates $\hat{\beta}_A + \hat{\beta}_P$ would be unbiased of the generating parameters (that are similarly summed).

This simple geometrical discussion provides the basis for many of the ‘solutions’ proposed in the APC literature and helps to guide the understanding of the APC problem. It is a conceptually appealing idea (if only for understanding the problem rather than solving it) that any solution from a generalized inverse can be found on a single line of solutions that is orthogonal to the plane that contains all of the data points. The movement from one constraint estimation to the other is given by a single scalar value and the null vector. To determine the bias from generating parameters in application is impossible as the population values are unattainable, therefore the parameter w that would demonstrate the bias in the estimation remains undetermined.

8.4.2 The Impact of a Linear Constraint

Constraints must be made on the estimation to obtain a unique set of linear estimates from the APC model. This is an unavoidable feature of the analysis. These constraints will influence the magnitude of the parameter estimates and potentially the direction of the trends (Holford 2005). However, infinite least squares solutions (i.e. minimizing SSR) still exist that produce the same fitted values (Weinkam and Sterling 1991). These are found along the line of solutions as illustrated in Figure 8.4. In the discussion of least squares estimation (see section 2.2.3) the ‘quality’ of an estimator was based on the ‘accuracy’ and ‘precision’ of the estimation with an unbiased estimator optimizing the accuracy (i.e. $\boldsymbol{\beta} = \mathbf{b}$). However, labelling an estimator unbiased defines that the constraint imposed perfectly reflects that of the population model. Due to the nature of the APC data, such a constraint could never be proven. This has led many researchers to propose a “most appropriate solution”. However, whether this is based on some arbitrary constraint, defined by the data or based on *prior* assumptions, we simply cannot determine optimality in this sense (Kupper et al. 1985b).

Most statistical software are programmed to ‘make up’ degrees of freedom (i.e. being 4 df less than full rank) by automatically employing some generalized inverse. For example, R (i.e. using the *lm* function in the *stats* library) automatically removes variables determined by the order in which they are entered (i.e. the variable removed when a singularity is identified). In terms of the linear constraint imposed on the model, those predictors ‘dropped’ from the model are said to have zero effect in the population (e.g. $\boldsymbol{\beta}_{A1} = \boldsymbol{\beta}_{P1} = \boldsymbol{\beta}_{C1} = \boldsymbol{\beta}_{C2} = 0$) (i.e. a one step solution). In contrast, authors such as Barrett (1973, 1978), Fienberg (1978) and Mason (1973, 1979) have set two effects from each category to be equal (e.g. $\boldsymbol{\beta}_{A1} = \boldsymbol{\beta}_{A2}$ etc.), which can then be followed by a full rank regression to obtain unique estimates based on these constraints. Although not impossible, it would seem unlikely that these assumptions would be perfectly reflected in the population model. Whilst some researchers have encouraged external information to guide assumptions, the lack of any way to definitively check them can be uncomfortable. If automatic assumptions are made that deviate from those naturally defined in the data, the constraints appear more dangerous and cast doubt on the interpretation of the results. If there is an effect in the generating population, but it has been constrained to zero, then an indeterminate parameter (w) will produce bias as part of the estimation.

Following a study by Rodgers (1982), population parameters are set for an APC model and simulated estimates attained from various constraints for the following model.

$$\hat{y} = \beta_0 + 2\beta_{A2} + 4\beta_{A3} + \beta_{P2} + 4\beta_{P3} - 2\beta_{C1} - 3\beta_{C2} - 4\beta_{C3} - \beta_{C4} \quad (7.44)$$

It is not an entirely comfortable procedure to set such generating parameters arbitrarily as demonstrated in eqn(7.44). As the underlying structure is never actually known for these processes, we simply do not know whether a specific combination of values lies at the heart of the problem. However, the purpose of these simulations is to explore the statistical impact of different constraints, therefore this approach would seem acceptable for this purpose. Three sets of constraints are applied. The first constrains β_{A1} , β_{P1} and β_{C5} to equal zero (i.e. reflective of the population model and the constraints explored in the original study). The second implements the zero sum constraints (eqn(7.38) and eqn(7.41)) and the final constrains β_{A1} , β_{P2} and β_{C5} to zero (a structure not reflected by the generating parameters).

For the first simulation (identical to Rodgers example) β_{A1} , β_{P1} and β_{C5} are set to zero. This is reflected in the population model and thus adds zero bias to the estimation.

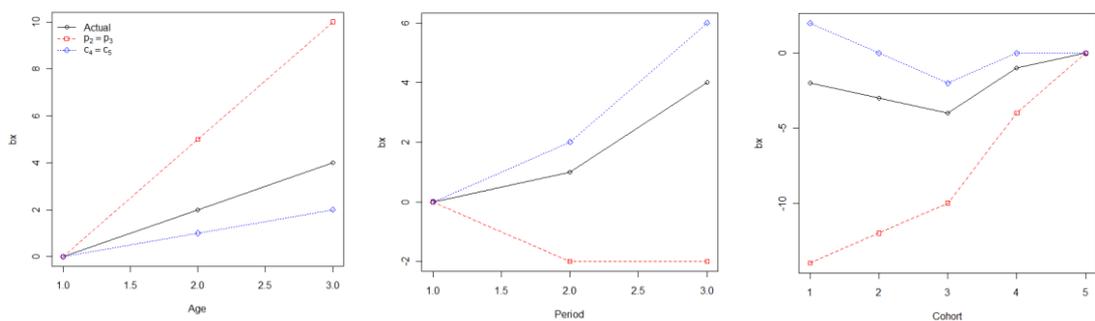


Figure 8.6: Point estimates from simulation with constraint $\beta_{A1} = \beta_{P1} = \beta_{C5} = 0$.

A further constraint is applied (as required to produce a non-singular design matrix) which sets $\beta_{C4} = 0$ (shown in blue) and $\beta_{P2} = \beta_{P3}$ (shown in red). Neither of these constraints are reflective of the population model, therefore a bias is introduced into the estimation. The predictors constrained to equal zero form the pivot point of the estimation and bias is introduced in the form of a rotation about these zero coefficients. The further the constraint from the population model parameters, the greater the bias. Notice that no distortion is introduced into the estimation (i.e. the shape of the curve is consistent in each category). Also, there is an equal bias (i.e. rotation) added to age, period and cohort. This is a clockwise rotation for age/cohort and anti-clockwise for period.

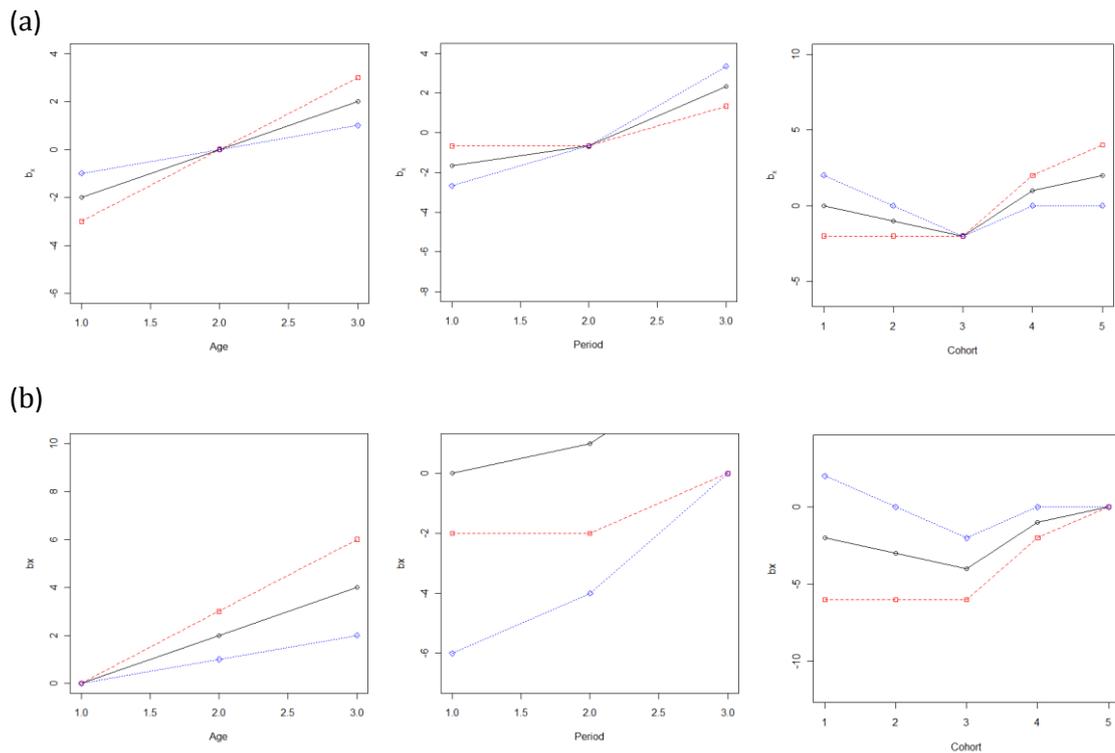


Figure 8.7: Point estimates with constraints - (a) $\sum_{i=1}^I \beta_{Ai} = \sum_{j=1}^J \beta_{Pj} = \sum_{k=1}^{I+J-1} \beta_{Ck} = 0$ and (b) $\beta_{A1} = \beta_{P3} = \beta_{C5} = 0$.

In the second simulation (see Figure 8.7a) zero sum constraints are implemented on the design matrix. No information regarding the population model is assumed. The estimates represent a deviation from the category mean (i.e. rotate about a common origin). The additional constraints applied are consistent with the first simulation, and therefore a bias is introduced, once again demonstrated as a rotation about the pivot. No distortion is introduced and the additional bias is consistent across age, period and cohort curves in the same clockwise/anti-clockwise pattern, as in the first. A final simulation is illustrated with β_{P3} now constrained to zero in place of β_{P1} (see Figure 8.7b). The original constraints are not reflective of the population parameters. Again a bias is introduced by the additional constraints (i.e. the rotation). This demonstrates the importance of each of the constraints to reflect the unknown population model, which will limit the bias in the estimation.

These simulations illustrate that no distortion to the model curves is introduced after applying any of the single identifying constraints. The degree of bias illustrated by the rotation from the true slope is equal for each category and conforms to the same relationship of a clockwise = anti-clockwise rotation for age and cohort with an anti-clockwise rotation for period. This is a result of maintaining the intrinsic relationship amongst the slope

coefficients as illustrated by Kupper (1983) (see eqn(7.42)) and shown by the indeterminate parameter w . These final two results are labelled ‘estimable’ as they are consistent when any g -inverse is used to obtain estimates. Whilst it is theoretically true that solutions cannot be eliminated with any certainty, a range of slopes can be identified for which it would be biologically implausible for the estimates to lie (e.g. a positive or negative slope). It would then be possible to define a range of values for w that fit our *a priori* assumptions of the relationships (Holford 2005). However, the correct rotation is not estimable and is akin to moving along the line of solutions.

The article by Rodgers progresses to imposing “multiple constraints” on the covariates. Constraining any two coefficients equates to setting anchoring points for the estimation, but not introducing a distortion to the curves. Therefore, the model fit remains consistent. The motivation to impose more constraints is borne from the fact that it would force some distortion into the curves and vary the model fit. For instance, removing a variable entirely from the analysis would produce a constraint such that all of the categories for that factor are zero. However, the intrinsic relationship defining the data must still hold and so effects for the remaining two coefficients are adjusted accordingly. This is in contrast to the lifecourse example and the geometry in section 8.4.1 which only considered the three predictors and a single identifying constraint. When the lifecourse was extended to seven predictors, BMI_0 and BMI_{19} (in separate models) were constrained to zero with parallel plots observed as only the anchor point is changed. Setting these single constraints changed the magnitude of the age-specific curves (i.e. the intercept), but not the slope. It was shown mathematically why this relationship exists in eqn(7.36). The problem with the multiple constraint strategy is that forcing a distortion into the curves and considering model fit will not validate the curve with the best fit to the data.

In the final part of this section data is simulated following the approach of Rodgers (1982) to illustrate the potential impact of sampling variation. Data are generated from the model in eqn(7.44) but now with a zero cohort effect. An error from a normal distribution with mean=0 and variance=100 is added to the response to generate 100 cases for each age and period combination and averaged.

	P_1	P_2	P_3
A_1	0.90	2.60	7.91
A_2	2.32	4.93	5.88
A_3	5.11	9.07	9.20

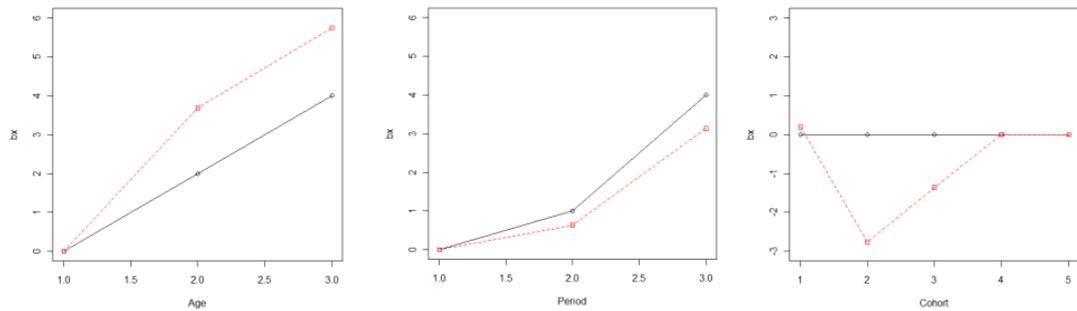


Figure 8.8: The impact of sampling error when the correct constraint is employed.

From a conceptual view, equating two single year cohorts would seem more biologically justifiable than equating wider cohorts (i.e. a lesser change would be expected). Similar to the discussion of the lifecourse example in chapter 7, when the continuous time variable were split into finer measurements, the collinearity amongst these categories naturally increased. Therefore, any deviation from the population generating model, either from a constraint not reflected perfectly in the generating model or from sampling variation, will become much inflated in the parameter estimates. Rodgers (1982) states that this difference is multiplied by a factor inversely proportional to the width of the cohort groups. Sampling error becomes a further important consideration in the APC analysis.

Rodgers (1982) suggests that it is not only the researcher without strong *a priori* assumptions that faces problems, but also those who do. No matter how strong the assumption (from a theoretical basis), there is no certainty of its appropriateness as there is no definitive way of checking. It is also very easy to underestimate the significance of the constraint applied as they may appear to have clear and understandable definitions. For instance, setting two constraints equal in each category may appear justified after looking at descriptive plots or the resultant fit of the data, however a seemingly simple constraint such as this may have substantial effects on the resultant analysis when it deviates from the population generating model. Even when the constraint reflects closely the population generating structure, sampling and measurement error can introduce a much inflated distortion into the estimates that has a potentially high impact on the interpretation of the coefficients. It must be noted that whilst the simulation approach employed by Rodgers is certainly of interest as a statistical exercise, there is a more fundamental issue to address with this approach. From an APC perspective, we must consider whether the population generating model and in particular the interpretation of the coefficients actually makes sense conceptually. This interesting question will be discussed in section 8.6.

8.5 Selected Solutions from the Literature

8.5.1 Curvature & Drift

Holford (1992) describes dependence on arbitrary constraints as a limitation of APC solutions and instead sought inferences from estimable functions. Approaches were suggested such as restricting the range of the slopes through the input of external knowledge and considering functions of estimates that are unbiased of the generating parameters. These methods rest on the indeterminate parameter w . Holford (1992) suggests that our knowledge of chronic diseases (speaking in a clinical context) is in general too basic to generate an accurate model of the population generating parameters. Whilst estimable functions cannot provide the coefficient estimates that we would ordinarily seek from the APC multiple classification model, the inferences gained from those functions can be made with confidence, regardless of the constraint chosen.

First order linear effects are not interpreted as they rely heavily on the constraints imposed on the model (i.e. an unknown rotation of the slope from an underlying model), however they are still sought from the data as “slopes”. Unlike the bias added to the estimation of linear effects, the estimability of non-linear functions is the reason that the curve is not distorted when a single identifying constraint is applied. Holford proposes a least squares solution partitioned into a simple linear regression to provide an “overall slope” or trend of the effects. The deviation away from this slope is then labelled the curvature (i.e. the residuals from the full constraint model and a simple regression model).

$$\beta_{Ai} = \beta_{A0} + i \beta_A + \gamma_{Ai} \quad (7.45)$$

Eqn(7.45) demonstrates how each regression coefficient is built from a category specific intercept term β_{A0} , an overall slope estimate β_A and a curvature γ_{Ai} . The curvature can be defined by using linear contrasts as demonstrated in eqn(7.42),

$$\gamma_{Ai} = \beta_{Ai} - \left[i - \frac{I+1}{2} \right] \beta_A \quad (7.46)$$

Eqn(7.46) illustrates that the curvature is defined as the regression coefficient with the linear trend removed. The curvature is dependent on the method by which the slope is estimated, however it is invariant to the constraints applied. Thus, it is possible to identify whether there is an increasing/decreasing or concave/convex nature to the estimates.

To find unbiased estimates of population parameters would appear to be a desirable property in any analysis. However, there are critics of Holford's curvature approach and how much it can actually provide as part of an APC analysis. Kupper (1985a) discusses that it is the linear effects (i.e. the slope) which often contribute most to the variation and that it is these estimates that are hidden when only considering curvature in isolation. If a factor demonstrates a linear effect, the curvature would indicate no change, but it could nevertheless still be an important effect. It is instead the curvature relative to the linear trend that would be of most interpretational benefit. In this sense, whilst statistically the estimates are unbiased, the interpretation is still very much dependent on the (arbitrarily) chosen method for eliminating the linear trends.

Clayton and Schiffrers (1987a, 1987b) propose a 'drift' coefficient δ that represents a quantity of variation that is common to both period and cohort (separable from age). The authors note that "no analysis of cancer epidemiology can ignore age". The age variable is often favoured in a clinical context as it has a solid biological basis and is usually assumed to play an important role in the biological system or mechanism under study. This is very much the basis of the proposed method, with the age, age-period and age-cohort models of most interest. Consider that moving from an age only to an age-period model indicates a significant period effect. Similarly moving from an age only to an age-cohort model also demonstrates a significant cohort effect. The similarly high model fits would suggest that there is some variation that cannot be attributed uniquely to either period or cohort factors. This quantity is labelled the 'drift' component by the authors and is considered as a separate parameter entered alongside age.

Curvature is unchanging under statistical constraints on the estimation, whilst the linear component is not. The indeterminate parameter w cannot be attributed to either period or cohort and so it is this that represents the 'drift', or instead the rotation in overall linear trend. For example, assume that the 'true' generating model is the age-period model.

$$y_{ij} = \beta_{Ai}X_{Ai} + \beta_{Pj}X_{Pj} \quad (7.47)$$

Next re-parameterize the model to be a function of age and cohort such that (as previously illustrated in eqn(7.2) for the lifecourse model) the true age slope is incorrect by cohort.

$$\begin{aligned} y_{ij} &= \beta_{Ai}X_{Ai} + \beta_{Pj}(X_{Ai} + X_{Ck}) \\ &= (\beta_{Ai} + \delta_P) X_{Ai} + \delta_P X_{Ck} \end{aligned} \quad (7.48)$$

This enhancement by β_{Pj} can be replaced by a period ‘drift’ term δ_{Pj} , the overall linear slope of β_{Pj} . It would represent an average change in the rates over time, which is identifiable, but cannot be partitioned to one of period or cohort. This could be a simple regression, as proposed by Holford, or a mean of successive differences, as suggested by Clayton & Schiffers (or indeed by any other appropriate estimate of trend). The latter approach cancels to the mean difference between the first and last groups.

$$\begin{aligned}\delta_P &= [(\beta_{PJ} - \beta_{P(J-1)}) + (\beta_{P(J-1)} - \beta_{P(J-2)}) + \cdots + (\beta_{P2} - \beta_{P1})]/(J - 1) \\ &= (\beta_{PJ} - \beta_{P1})/(J - 1)\end{aligned}\quad (7.49)$$

This process could similarly start with the age-cohort model as the true generating model and re-parameterise to an age-drift model with period included. The indeterminate parameter w would have an equal and opposite effect on the age coefficients in this case.

A ‘net drift’ is defined by a re-parameterisation of the full APC model in a similar way. Once again, drift is defined by the non-estimable overall trend,

$$y_{ij} = \beta_{Ai}X_{Ai} + \beta_{Pj}X_{Pj} + \beta_{Ck}X_{Ck} \quad (7.50)$$

$$= \beta_{Ai}X_{Ai} + \delta_P(X_{Ai} + X_{Ck}) + \delta_C X_{Ck}$$

$$= (\beta_{Ai} + \delta_P)X_{Ai} + (\delta_P + \delta_C)X_{Ck} \quad (7.51)$$

Net drift is defined by $\delta_P + \delta_C$ (the same result is also achieved by replacing X_{Ck} with $X_{Pj} - X_{Ai}$ in eqn(7.50)). The three variable model can be used to derive estimates from any two variable model, but not the reverse. These are sub-models of the full model (Osmond 1996). The approach is to test the models using deviance, or some variant of this measure such as the AIC/BIC, to assess the improvement in fit moving from the age only model to the full APC multiple classification model. Age-period and age-cohort are not directly comparable by deviance alone (although AIC/BIC may be appropriate). The age-cohort model has more parameters than the age-period model and so is likely to produce a better fit to the data. Also, model fit is invariant to the choice of constraint (as discussed for the simulations in section 8.4.2), therefore if one of the factors were to follow a linear pattern in the population, it may be seen to have zero effect (i.e. two categories within the factor are constrained to be equal) or a non-zero relationship dependent on the constraint employed. In either case model fit will be equal and moving from a two-factor model (assuming zero effect for the excluded variable) to the full model will produce a non-

significant improvement in fit (Kupper et al. 1985b). Hence the two factor model would be chosen as optimal, whilst the factor constrained to have zero effect (perhaps falsely) could produce a potentially ‘serious’ bias on the resulting estimates.

The motivation for the Clayton & Schiffers approach is to identify whether the effects are either attributable to period or cohort to simplify the analysis towards a two predictor model. Whilst this would seem a useful aim to avoid the identification problem, the use of model fit criteria to assess the validity of a model is not adequate. A strong model fit only indicates that the model produces fitted values close to the observed data. However, as previously noted, an infinite number of models can produce these same fitted values and so model fit is no indication of the appropriateness of the individual coefficients (i.e. any along the line of solutions). The notion of Clayton & Schiffer’s paper regarding the focus on age is interesting in comparison to the sociological example described earlier that saw birth cohort as a dominant factor. This is a reflection of the interests between clinical and sociological fields. A similar analysis could be performed with the focus on cohort from a sociological perspective, with a similar drift term defined.

8.5.2 Minimal Differences

Osmond & Gardner (1982) considered the bivariable sub-models of the full three factor APC model to generate a solution. The motivation of the minimal differences approach is to partition net drift variance into either period or cohort using a penalty function based on the magnitude of the non-drift effects – i.e. the curvatures (Doll, 2004). The authors considered the estimation of the bias parameter w , such that a specified penalty function is minimized. This method may be considered geometrically within the framework previously used. The method looks at the Euclidean distance between the two and three predictor model estimates along the line of solutions. The three two factor solutions are computed by constraining the excluded variable to equal zero. To calculate the value of w used for the three variable solution the authors propose an “intermediate value”. The distance between the full solution and the three two variable solutions is then inversely weighted by the mean residual sums of squares. The value of w is found by minimizing the following function, although alternative functions could also be specified (see also Decarli & La Vecchia (1987)),

$$\mathbf{g}(w) = \|\boldsymbol{\beta}(w) - \boldsymbol{\beta}_{(A)}\|/\rho_A + \|\boldsymbol{\beta}(w) - \boldsymbol{\beta}_{(P)}\|/\rho_P + \|\boldsymbol{\beta}(w) - \boldsymbol{\beta}_{(C)}\|/\rho_C \quad (7.52)$$

where ρ_A , ρ_P and ρ_C are the residual mean squares from the two variable models and $\beta_{(A)}$, $\beta_{(P)}$ and $\beta_{(C)}$ are the corresponding sets of estimates. This is in effect an ‘averaging’ of the solutions to produce a three variable solution inversely weighted by model fit. This remains one selection out of the infinitely many available. It would seem sensible to give models with a poor fit to the data a lesser weighting in the full model. However, as with Clayton & Schiffer’s approach, basing criteria on model fit is dangerous to identify parameters that we wish to interpret. Holford (2005) criticises the method in that curvature of the full model can only be estimated using a full three factor model. Hence, the two factor models do not provide the information required to produce the desired parameter. Minimum differences is an approach that would make the necessary leap into constraint based solutions and as expected leaves itself open to criticism (as any in this category would under perfect collinearity). The method provides a useful extension for the current discussion as it attempts a partition of the drift component.

8.5.3 Individual Records & Natural Weights

Robertson and Boyle (1986) suggested that the identification problem could be overcome by forming a three-way lexis diagram for the three factors.

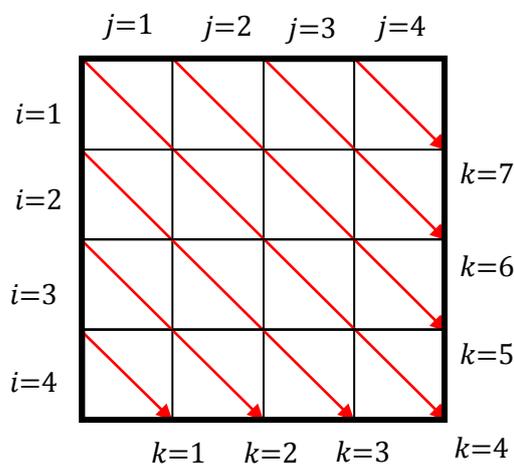


Figure 8.9: The Standard Structure of APC data.

Figure 8.9 demonstrates that in the ordinary structure of APC data each rate belongs to only one cohort. This guarantees the overlap of cohort groups and can make seemingly simple constraints produce very complex interpretations. The viability of using overlapping birth cohorts to solve the identification problem has been discussed extensively in the literature (Clayton and Schiffers 1987; Osmond and Gardner 1989).

When each subject's date of birth is known, the user can define subjects to lie within 'non-overlapping' cohort groups (i.e. as a subject ages, they will only ever belong to one cohort). This is achieved by splitting each cell into an upper and lower triangle. The first potential problem with this is that the detailed individual records must be available to the user (which is not the case in the Dahlquist APC data). However, this theory would still warrant discussion if this would solve the problem. The APC multiple classification model (see eqn(7.40)) can be re-written as follows.

$$\log(y_{ij}^t) = \beta_0 + \beta_{Ai} + \beta_{Pj} + \beta_{C1-i+j+t} + \epsilon_{ij} \quad (7.53)$$

When $t = 0$, the rate is part of a lower triangle in each cell and when $t = 1$ it is a rate in an upper triangle (see Figure 8.10 for an example cell). On the surface, this model would seem to relieve the identification problem, however it masks important underlying assumptions.

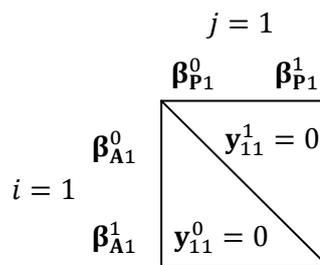


Figure 8.10: A single cell of the Lexis diagram.

Through the process of splitting each cell to be assigned to two cohorts, two separate age effects (β_{Ai}^0 and β_{Ai}^1) and two separate period effects (β_{Pj}^0 and β_{Pj}^1) have been created for each cell. Therefore, a finer grid has been employed, however the problem still remains. It is clear from Figure 8.10 that y_{11}^0 occurs at an earlier period (i.e. β_{Pj}^0) and at a later age (i.e. β_{Ai}^1). This is not accounted for in the model in eqn(7.53). If it were, there would be a return to the singularity problem presented in the original data. Equivalently neighbouring age and period groups could be merged to generate a broader grid and still produce non-overlapping cohorts, but the underlying problem would remain. This is defined by the intrinsic relationship present in the data.

Employing a finer grid would not overcome the underlying structure in the data that leads to the identification problem. However, the notion of splitting each cell, such that cohort groups can be interpreted on the same scale as age and period is a feature that should not be ignored. Weinkam and Sterling (1991) agreed in principle with this graphical representation of the APC problem by producing a finer grid on a triangular mesh to give

an equal spacing to each factor. This representation can provide a useful insight into the problem from an interpretive view. An important feature of APC data is that the time dependent observations have been tabulated to produce categorical factors. One of the limitations of tabulation in time data is that it restricts “cross-time linkages” (Fienberg and Mason 1982). This is perhaps the conceptual limitation of the arbitrary constraints employed by Robertson, in that the neighbouring categories are given the same weighting as belonging to the same cell. Instead, the tabulated data in the APC model can be considered continuous, with each cell representing a “knot” on the diagram. A contour plot could be produced on the following grid with a potentially simpler interpretation of effects.

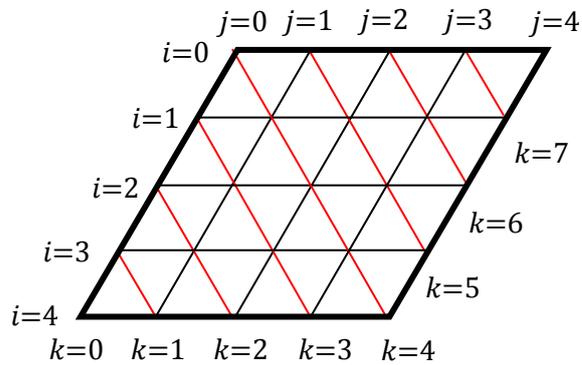


Figure 8.11: Lexis table on a triangular mesh, with linear effect directions A, P, C.

Lee and Lin (1996b) used this structure to identify “natural” weights (i.e. no *prior* information assumed) regarding the structure of the data. The technicalities of the original paper will not be repeated here as they add little to the current methodological discussion. However, the general aims of the exercise are worth discussing as it provides a link between the methods that were discussed in the literature review and a justification for use of the modern latent variable approaches.

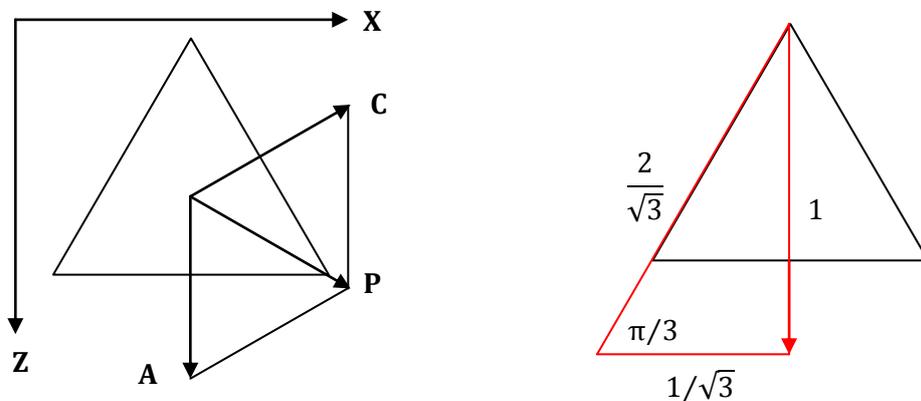


Figure 8.12: Cell weightings obtained directly from the lexis diagram.

In traditional trend surface analysis effects are separated into ‘regional’ and ‘local’ trends. Any effects captured by a 2D trend surface are considered regional, whilst any departure (or residuals) from this surface become local effects. Each observation in the data is a summation of a regional (or systematic) trend and a residual (i.e. akin to a contour plot defining a 2D surface for 3 variables). The uniqueness of the perfect collinearity problem is that there is no residual - all of the variation must be captured in the systematic component as one variable is defined by a relationship inherent in the data. It is this relationship that defines the 2D surface and allows us to explore properties of the estimates. Lee & Lin use reference axes that are illustrated in Figure 8.12 to produce the trend surface function, however any arbitrary orthogonal axes of \mathbf{X} and \mathbf{Z} could be defined that span the 2D regression space. A principal axis transformation of i, j, k onto \mathbf{X}, \mathbf{Z} in this example (see section 3.3.2) is used to define the new coordinate system for the Lexis diagram in Figure 8.11 as follows ($\mathbf{x} = j - i/2, \mathbf{z} = i \cdot \sqrt{3}/2$).

To analyze linear trends, a compromise is made in the direction that allows the target variable to change, whilst giving an equal weighting to the remaining predictors (e.g. for age this direction is marked \mathbf{A} in Figure 8.12, period marked \mathbf{P} and cohort \mathbf{C}). These are directions orthogonal to the contour levels of the associated predictor in Figure 8.11 to achieve maximal squared increment in the target variable relative to the remaining two covariates (Lee and Lin 1996b). As the categories are represented equally on the grid, the justification for the equal angles between \mathbf{A}, \mathbf{P} and \mathbf{C} can be directly observed. Any other weighting (representing other constraint solutions) would require justification to move away from that naturally defined by the design matrix. For a unit change in the direction \mathbf{A} , age will increase by $2/\sqrt{3}$, period by $1/\sqrt{3}$ and cohort decrease by $1/\sqrt{3}$ (see Figure 8.12b). This result is the same for \mathbf{P} and \mathbf{C} with the target variable (i.e. period in the direction of \mathbf{P} and cohort in the direction of \mathbf{C}) changing by $2/\sqrt{3}$, and the remaining covariates equal and half of this value. The intrinsic relationship that defines the variables is clearly followed by these weights (although a $\sqrt{3}$ constant is used in the original paper to scale the weights to “adjusted variables”). It should be clear that to choose the surface as in the above figure is to select the MP/PCR/PLS solution (although different arbitrary axes would be selected to differentiate these methods when fewer components are retained). It was not a feature noted by the authors in the original paper that the trend surface approach defines these methods, but it is nevertheless based on a similar motivation. The next section will look to provide a novel motivation for why these methods may be considered optimal in the absence of robust external evidence.

8.6 Latent Variable Methods in APC Analysis

In section 8.4.1 it was demonstrated that the PCR solution is on the solution hyper plane orthogonal to the null space. This is the same constraint that is employed with the MP g -inverse (demonstrated in section 7.6.2), PLS and also a more recent proposal, the intrinsic estimator (IE). There is no external clinical or biological reasoning for choosing this constraint out of the infinitely many least squares solutions available, but it has been preferred for a few reasons. Zero variance is assigned to the null dimension and so the constraint is entirely data driven. In contrast, moving elsewhere along the line of solutions is making some *prior* assumption about the null and hence the nature of the relationships in the data. Such an assumption is not directly supported by the data. The solution is at the centre of the line of solutions with the shortest distance from the origin. This may represent an average of the g -inverse solutions (O'Brien 2011). Therefore, it could be suggested that in the absence of better reasoning from *a priori* knowledge, this would represent the generalized solution based objectively on the data provided.

From a purely statistical perspective it seems sensible that the constraint is based on the structure of the design matrix and not the response. There are an infinite number of models that can produce the same fitted values of \mathbf{y} . Therefore, to use \mathbf{y} to determine \mathbf{c} would seem misguided. Kupper (1983) suggests that if “reasonably reliable” information can be utilized based on information external to the observed response then this should present a valid approach to determining the constraint employed. Also, if several constraints can be suggested based on external knowledge that do not differ greatly in the estimates produced, then this should raise confidence in the accuracy of the estimated effects. Caution should perhaps be placed on the rationale behind this approach. The discussion of perfect collinearity in chapters 7 and 8 is primarily motivated by application and interpretation of the coefficients to the definitions of the variables. Whilst an approach such as Rodgers’ simulations is certainly a useful statistical exercise, we must seek to understand what these simulations would signify in application.

8.6.1 Justification for the Application of Latent Variable Methods

It would appear to be the natural progression of the APC discussion to reduce the number of components entered into the model and see if this approach brings us nearer to the generating model used to produce the simulations in section 8.4.2. However, the notion of

specifying an arbitrary set of generating parameters for the underlying model is an uncomfortable one. The models specified in the original Rodgers (1982) paper were repeated for the simulations to facilitate a statistical discussion regarding the role of error and the choice of constraints. As expected, the choice of a constraint that is not reflected in the population model provides a solution that suffers from bias of the coefficients. However, do these underlying models make sense as part of an APC setting? Should it be possible to have a zero cohort effect (as defined in the second simulation) within the structure defined by the age and period variables (which are all intrinsically related)? This is an important point that is rarely discussed in the simulation literature and for this reason it may be considered nonsensical by some that in an APC setting we are simulating from such a model. It is this discussion that brings us to the root of the perfect collinearity problem that has been discussed in this and the previous chapter.

It is important to consider what various constraint solutions mean with regard to an underlying model (which may or may not be assumed to exist). Through the use of latent variable methods, a solution is provided that is guaranteed to follow the intrinsic structure of the variables. This makes a direct interpretation of the coefficients simpler. To justify this stance a simple example is provided. Consider that two variables are entered into a regression that are perfectly collinear and have equal variance. It would seem that the two variables are statistically equivalent and must receive the same weighting. However, if the variables are assumed in practice to be measuring something conceptually different than the data provided based on *a priori* knowledge, then perhaps it could be argued that one should receive a greater weight. From a statistical perspective this would not be reflected in the measured variables and so the research question is actually based on data that is different to what has been provided. The problem is how to interpret the coefficients from this weighted approach. One could argue that the concepts of age, period and cohort are not actually perfectly collinear and instead it is only the coarse structure of the lexis table and the definitions of the latent variables that defines them as such (i.e. the notion of micro level variables defining the actual research question not being perfectly collinear). Hence the Rodgers model could still be acceptable as a generating model, but it must be based on this specific conceptual understanding of the variables entered and adjusted for. This is a difficult concept to grasp as effectively the variables must be interpreted as something that statistically they appear not to be.

It is the structure that was presented in the trend surface approach of Lee & Lin that would be 'naturally' defined by the data. Any other solution along the line of solutions

would be moving away from that directly supported by the data. Therefore, for the Rodgers paper the final model used to illustrate the impact of error requires more information about what the author assumes and what each of the solutions compared actually represents in application. It is this distortion in the estimation that highlights the questionable approach of the original paper for anything other than a statistical discussion of bias. If one of the variables is weighted (which are perfectly collinear and equal variance), what is the interpretation of the remaining variables that are adjusted for (which are still held by the intrinsic relationship)? If the measured variables are not reflecting the research question then perhaps we require variables that will. This data may not be viable due to practicality and monetary constraints of obtaining it, but would allow for a direct interpretation. Therefore, whilst the use of alternative constraints would appear supported by statistical exercises and theoretical discussion, the suitability of these methods in application would seem limited. For this reason, it would seem sensible that the population generating model assumes the intrinsic structure of the variables (as proposed by Lee and Lin, 1996) in the perfectly collinear relationship. The simulation exercise is not repeated using such a model in the current discussion as a comparison of the properties of latent methods has been demonstrated earlier in this work. However, if a simulation exercise were to be repeated, it would seem justified that the coefficients employed follow the intrinsic relationship defined by the APC problem.

8.6.2 Application of Latent Variable Methods

It is not essential to once again compare properties of the PLS, PCR and the MP-inverse in this section as these have been previously discussed in chapter 7. The difference in this chapter is that the data are in aggregate form. It is clear from the dummy coding that the sum of each of the categories for each variable equals unity. After the constraints in eqn(7.38) and eqn(7.41) are applied, these categories sum to zero in each approach. An interesting development in the APC literature is a method recently proposed by Yang et al. (2004). The authors recognise that the IE method has very close links to the PCR procedure that has been discussed throughout this work. Yang et al. (2008) states that “the IE uses the extra step of inverse orthonormal transformation of the coefficient estimates of the principal components regression back to the original space of age, period and cohort”. As this step has already been assumed as part of the PCR process it is not a surprise that this work considers the methods fundamentally the same (see section 4.4.1).

For the Yang et al. paper, $t\mathbf{B}_0$ is representing the uni-dimensional vector spanning the null space. \mathbf{B}_0 is a unit vector (i.e. the eigenvector associated with the zero eigenvalue that makes the design matrix computationally singular) and s a scalar multiple determining the influence of the null vector on the estimates (i.e. w in our discussion). They highlight that in the IE case (as in PCR, PLS etc.) $t = 0$ (i.e. the eigenvalue). The authors state that the IE estimates are in fact 'estimable functions' (as defined by Holford). This is to say that the estimates in the 2-dimensional space are simply functions of the unattainable estimates from the full three variable model (this was demonstrated similarly in the lifecourse study – see eqn(7.27) and eqn(7.34)). The eigenvectors are independent of the response and so the null eigenvector is determined by the number of age and period groups in the data only.

For the estimator to be unbiased the constraint employed must be orthogonal to $\vec{\beta}$. In the Yang et al. paper they state that the estimator is unbiased, however this is conditional on the estimator lying in the space orthogonal to the null vector. Therefore, the unbiased property is still dependent on the constraint imposed. The null vector is not dependent on the response and therefore becomes an estimable function of the generating parameters. A further important property is the variance (or efficiency) of the estimator compared with other constrained estimators. The MP-inverse (which is equivalent to these latent variable solutions with maximal components retained) has been shown to have the minimum variance as only non-zero eigenvalues and the associated eigenvectors form part of the regression (Kupper et al. 1983).

An important feature of the IE script is that effect coding is adopted. One fewer than the maximal number of covariates is entered for each category with the final group (i.e. the reference) coded as minus one. The coefficient for this covariate is defined as the negative summation of the other effects. When a singularity is identified, the latent variable is removed, which in the standard case results in the final cohort category. However, if the procedure were repeated, but a different cohort category removed, a different set of coefficients would be obtained (due to unbalanced groups). In comparison, PLS uses standard dummy coding to directly extract all of the estimates. A category mean is subtracted from each group and therefore removing any category would produce the same coefficient estimates. This is a minor discrepancy between the methods. It is not clear which approach is correct and the result is unlikely to drastically change the interpretation, however for the interests of consistency in application PCR/PLS coding should be encouraged in favour of the more recent IE proposal.

8.7 Regression Study

The intention of chapter 7 was to consider the mathematics behind a PCR/PLS analysis applied to perfectly collinear data. Chapter 8 has focussed more on the interpretation of applying these methods and where they may be placed as part of the vast literature of statistical techniques in the APC problem. This section focuses on the data from Dahlquist et al. (2011) and considers the results attained from the application of selected methods discussed in this chapter. Holford (1985;1992;2005) and Clayton & Schiffrers (1987) both proposed methods based on estimable functions that provide complete confidence in the measures but do not attempt to identify linear effects of the factors. These methods are applied in this section. It is also considered what additional insight can be gained from the application of PCR/PLS to the same data.

The study considers incidence of childhood onset type-1 diabetes and so this analysis (in agreement with the authors) assumes age to be a dominant factor in the analysis (Robertson et al. 1999). Age-specific descriptive plots were illustrated in section 8.3 to continue this idea. A descriptive simplification of age specific rates is to produce age-standardized rates by collapsing the period columns and choosing a reference period. However, these statistics are intended only as a summary as they assume a constant age trend consistent in each period group (Osmond and Hardy 2004). This study begins with the two factor models of age entered along with period or cohort (see Table 8.3-Table 8.6). Following the approach of Osmond & Hardy (2004), the age coefficients are presented analogous to the age-specific incidence rates presented in section 8.3.

Age	β_{Ai} (95% CI)	Period	β_{Pj} (95% CI)
2	22.32 (21.29 to 23.40)	1985	0.91 (0.87 to 0.94)
7	34.64 (33.30 to 36.03)	1990	0.92 (0.88 to 0.95)
12	41.74 (40.26 to 43.28)	1995	0.93 (0.90 to 0.97)
17	19.84 (18.91 to 20.82)	2000	1.08 (1.04 to 1.11)
22	16.88 (16.04 to 17.77)	2005	1.20 (1.16 to 1.24)
27	15.78 (14.98 to 16.62)		
32	12.65 (11.96 to 13.38)		
Deviance=196.61			

Table 8.3: Coefficients from an Age + Period regression of the male subjects.

Age	β_{Ai} (95% CI)	Period	β_{Pj} (95% CI)
2	20.58 (19.53 to 21.69)	1985	0.91 (0.87 to 0.95)
7	36.80 (35.27 to 37.98)	1990	0.86 (0.82 to 0.90)
12	36.42 (34.92 to 37.98)	1995	0.98 (0.94 to 1.03)
17	12.09 (11.35 to 12.88)	2000	1.08 (1.03 to 1.12)
22	10.17 (9.51 to 10.88)	2005	1.21 (1.16 to 1.26)
27	8.12 (7.55 to 8.73)		
32	6.20 (5.72 to 6.73)		
Deviance=27.37			

Table 8.4: Coefficients from an Age + Period regression of the female subjects.

As part of the coding, the zero sum constraints have been applied only to the period and cohort categories. The anti-log of these coefficients can then be viewed similar to relative risk for the associated period/cohort. These rates were obtained using the constrained regression function in STATA (Version 12). The intercept is collapsed into the age effects to obtain coefficients that can be interpreted similar to age-specific rates. This would appear beneficial as the primary interest is in interpreting the importance of period and cohort in the clinical example.

Age	β_{Ai} (95% CI)	Cohort	β_{Ck} (95% CI)
2	18.63 (17.46 to 19.87)	1953	1.04 (0.91 to 1.20)
7	32.57 (30.96 to 34.26)	1958	0.91 (0.82 to 1.01)
12	44.83 (42.94 to 46.81)	1963	0.92 (0.85 to 1.00)
17	23.80 (22.61 to 25.06)	1968	0.79 (0.74 to 0.85)
22	20.12 (19.03 to 21.28)	1973	0.77 (0.73 to 0.82)
27	18.74 (17.62 to 19.93)	1978	0.82 (0.78 to 0.87)
32	14.37 (13.32 to 15.51)	1983	0.89 (0.84 to 0.94)
		1993	0.92 (0.87 to 0.98)
		1998	1.27 (1.20 to 1.34)
		2003	1.44 (1.34 to 1.56)
Deviance=130.69			

Table 8.5: Coefficients from an Age + Cohort regression of the male subjects.

Age	β_{Ai} (95% CI)	Cohort	β_{Ck} (95% CI)
2	17.84 (16.59 to 19.18)	1953	1.01 (0.82 to 1.25)
7	35.54 (33.55 to 37.65)	1958	1.03 (0.89 to 1.18)
12	39.73 (37.73 to 41.83)	1963	0.83 (0.74 to 0.92)
17	13.99 (13.09 to 14.96)	1968	0.9 (0.83 to 0.99)
22	12.06 (11.21 to 12.96)	1973	0.81 (0.75 to 0.86)
27	9.26 (8.51 to 10.08)	1978	0.84 (0.79 to 0.9)
32	6.82 (6.14 to 7.58)	1983	0.95 (0.89 to 1.01)
		1993	1.17 (1.09 to 1.25)
		1998	1.45 (1.34 to 1.58)
		2003	1.40 (1.23 to 1.59)
Deviance=27.99			

Table 8.6: Coefficients from an Age + Cohort regression of the female subjects.

The drift δ in the model is 0.015 ($\log(\delta) = 1.02$), which represents the linear component that can be equally attributed to either period or cohort. Both models demonstrate an increasing incidence with each period/cohort until age 12 before a decline later in the lifecourse. Model fit statistics of the nested models can be studied as suggested in Clayton & Schiffrers (1987). Dahlquist et al. (2011) utilized the drift approach of Clayton & Schiffrers (1987) to assess whether there was a period or cohort effect in operation. The authors use AIC to adjust for the difference in parameters between period and cohort models. They conclude that there is a predominant cohort effect due to the minimal AIC for the model and a significant improvement in model fit moving from the A+P to the full APC model.

Model	df	Male		Female	
		Deviance	AIC	Deviance	AIC
Age	28	338.68	617.08	258.29	522.41
Age + Drift	27	214.98	495.38	145.84	411.96
Age + Period	24	196.61	483.01	130.69	402.81
Age + Cohort	18	27.37	325.76	27.99	312.11
Age + Period + Cohort	15	24.25	328.64	22.85	313.97

Table 8.7: Deviance and AIC statistics from nested APC models.

The curvature estimates are next considered based on the approaches of Holford and Clayton & Schiffers to consider estimates from the full APC model. The 'default constraints' that are listed in Table 8.8 are those applied by R using *lm* in the *stats* library in which the last category of age and period are constrained to zero whilst the final two categories of cohort are similarly constrained to zero (i.e. applied whenever the program identifies a singularity amongst the columns). The curvature components have been obtained by fitting least squares and mean difference trends (as described in section 8.5.1). The least squares curvatures were found by fitting a simple linear slope to the data (similar to the drift) and extracting the residuals from this model. The mean differences approach is applied by including linear age and cohort terms in the model and constraining the first and last categories to equal zero. The anti-logs of the curvature trends are demonstrated in Figure 8.13 for males and Figure 8.14 for female groups. Despite many papers discussing this approach, few seem to have made meaningful insight into the data using curvature estimates only. Observations can be made regarding the shape of the trends (e.g. concave upwards or downwards). For instance, birth cohort for both groups demonstrates a 'U' shape. From this, a decelerating trend can be described from 1953 to 1973 cohorts for males, before remaining relatively stationary to around 1988, followed by an accelerating trend to more recent birth cohorts. The female curve is similar, but starts from a lesser risk and with a downward trend in the most recent cohort.

The estimates gained from the full APC model are considered next. To attain these estimates, the constrained linear regression function is once again used with constraint (1, 1, -1) to follow the intrinsic structure of the covariates. This is equivalent to a PCR/PLS analysis with maximal components retained (19 for this example). In this analysis the drift component is partitioned to be attributed to either period or cohort effects. The age curves for males and females both demonstrate a peak at age 12 (the 10-14 years interval), although the maxima is better differentiated for males. The age-specific plots (see Figure 8.2) indicated a shift to an earlier age in peak incidence for later periods in the female group. It is clear for both groups that period at diagnosis contributes to a lesser degree than birth cohort. Period values for males are positive and maximised at year 2000 (1.04 (1.00 to 1.08)), whilst two peaks are seen for females at the earlier years of 1985 (1.04 (0.99 to 1.09)) and 1995 (1.04 (1.00 to 1.09)) following a negative trend. Cohort values for both groups peak at more recent years, with males at 2003 (1.40 (1.28 to 1.54)) and females 1998 (1.57 (1.47 to 1.68)). The female cohort trend then demonstrates a small decrease in the 2003 birth cohort (1.55 (1.39 to 1.72)).

MALES		Default Constraints	Curvatures	
			Least Squares	Mean Change
Intercept		-8.78	-39.28	-22.63
Period	1985	-0.22	0.00	0
	1990	-0.15	0.01	0.02
	1995	-0.13	-0.02	-0.02
	2000	-0.03	0.02	0.03
	2005	0	-0.01	0
Cohort	1953	0.20	0.28	0
	1958	0.00	0.09	-0.18
	1963	-0.04	0.06	-0.20
	1968	-0.25	-0.14	-0.39
	1973	-0.33	-0.21	-0.45
	1978	-0.33	-0.20	-0.43
	1983	-0.31	-0.16	-0.39
	1988	-0.33	-0.18	-0.39
	1993	-0.07	0.10	-0.11
	1998	0	0.18	-0.02
	2003	0	0.19	0
	Deviance		24.25	24.25

Table 8.8: Default constraints along with least squares and mean difference curvatures.

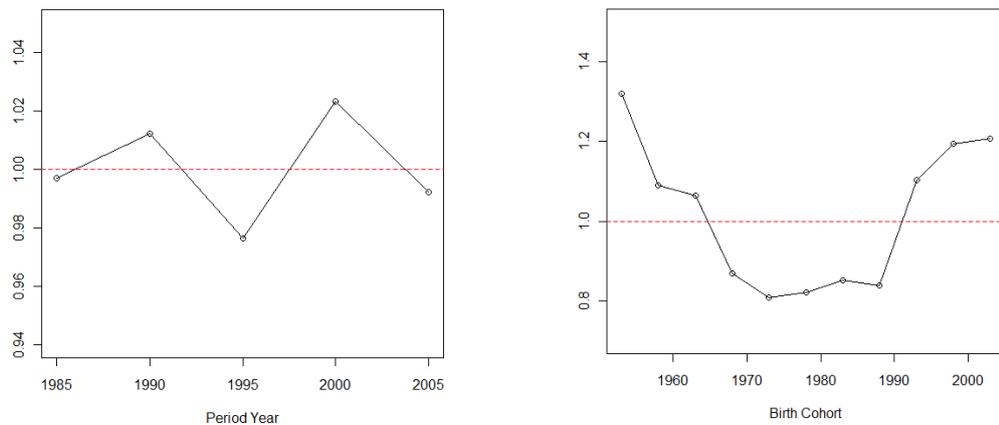


Figure 8.13: Anti-log least squares curvature estimates for males in the study.

FEMALES		Default Constraints	Curvatures	
			Least Squares	Mean Change
Intercept		-9.08	-42.95	-20.26
Period	1985	0.14	0.01	0
	1990	0.06	-0.04	-0.04
	1995	0.11	0.04	0.04
	2000	0.05	0.01	0.01
	2005	0	-0.01	0
Cohort	1953	-0.64	0.20	0
	1958	-0.57	0.19	0.00
	1963	-0.78	-0.08	-0.27
	1968	-0.66	-0.03	-0.21
	1973	-0.70	-0.14	-0.32
	1978	-0.71	-0.21	-0.39
	1983	-0.65	-0.22	-0.39
	1988	-0.50	-0.14	-0.31
	1993	-0.26	0.04	-0.13
	1998	0	0.23	0.06
	2003	0	0.16	0
	Deviance		22.85	22.85

Table 8.9: Default constraints along with least squares and mean difference curvatures.

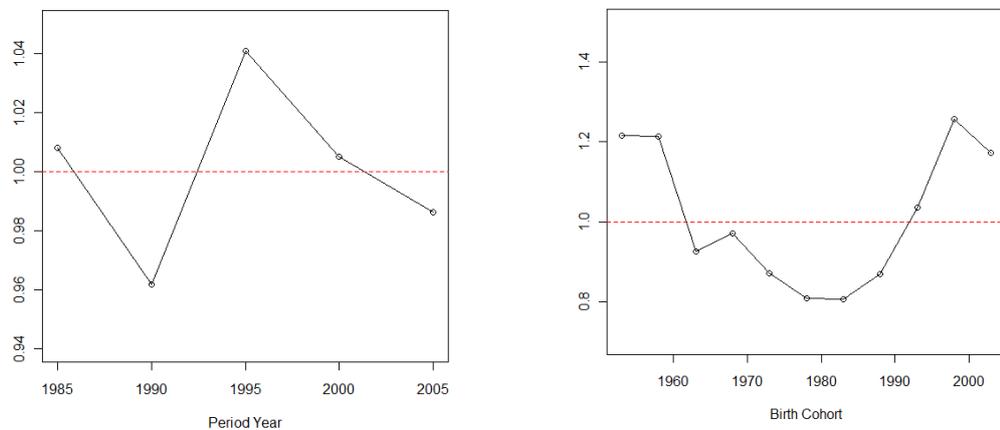


Figure 8.14: Anti-log least squares curvature estimates for females in the study.

Age + Period + Cohort					
Age	β_{Ai} (95% CI)	Period	β_{Pj} (95% CI)	Cohort	β_{Ck} (95% CI)
2	19.56 (18.62 to 20.55)	1985	0.96 (0.93 to 1.00)	1953	1.14 (1.01 to 1.27)
7	33.66 (32.26 to 35.12)	1990	1.00 (0.96 to 1.04)	1958	0.97 (0.89 to 1.05)
12	45.62 (43.81 to 47.50)	1995	0.98 (0.94 to 1.01)	1963	0.97 (0.90 to 1.04)
17	23.83 (22.62 to 25.10)	2000	1.04 (1.00 to 1.08)	1968	0.82 (0.77 to 0.88)
22	19.79 (18.73 to 20.91)	2005	1.02 (0.99 to 1.06)	1973	0.78 (0.74 to 0.83)
27	18.15 (17.18 to 19.17)			1978	0.82 (0.78 to 0.87)
32	13.69 (12.93 to 14.50)			1983	0.88 (0.83 to 0.92)
				1988	0.89 (0.84 to 0.94)
				1993	1.21 (1.15 to 1.27)
				1998	1.35 (1.26 to 1.43)
				2003	1.40 (1.28 to 1.54)

Table 8.10: PLS Coefficients from a maximum components model for Males.

Age + Period + Cohort					
Age	β_{Ai} (95% CI)	Period	β_{Pj} (95% CI)	Cohort	β_{Ck} (95% CI)
2	16.90 (15.99 to 17.87)	1985	1.04 (0.99 to 1.09)	1953	0.92 (0.78 to 1.09)
7	34.38 (32.83 to 36.00)	1990	0.98 (0.93 to 1.03)	1958	0.97 (0.86 to 1.10)
12	39.29 (37.48 to 41.18)	1995	1.04 (1.00 to 1.09)	1963	0.78 (0.71 to 0.87)
17	14.07 (13.15 to 15.06)	2000	0.99 (0.95 to 1.03)	1968	0.87 (0.79 to 0.95)
22	12.31 (11.45 to 13.23)	2005	0.95 (0.91 to 1.00)	1973	0.82 (0.77 to 0.89)
27	9.55 (8.84 to 10.32)			1978	0.81 (0.76 to 0.87)
32	7.19 (6.63 to 7.79)			1983	0.85 (0.80 to 0.91)
				1988	0.97 (0.92 to 1.03)
				1993	1.22 (1.16 to 1.29)
				1998	1.57 (1.47 to 1.68)
				2003	1.55 (1.39 to 1.72)

Table 8.11: PLS Coefficients from a maximum components model for Females.

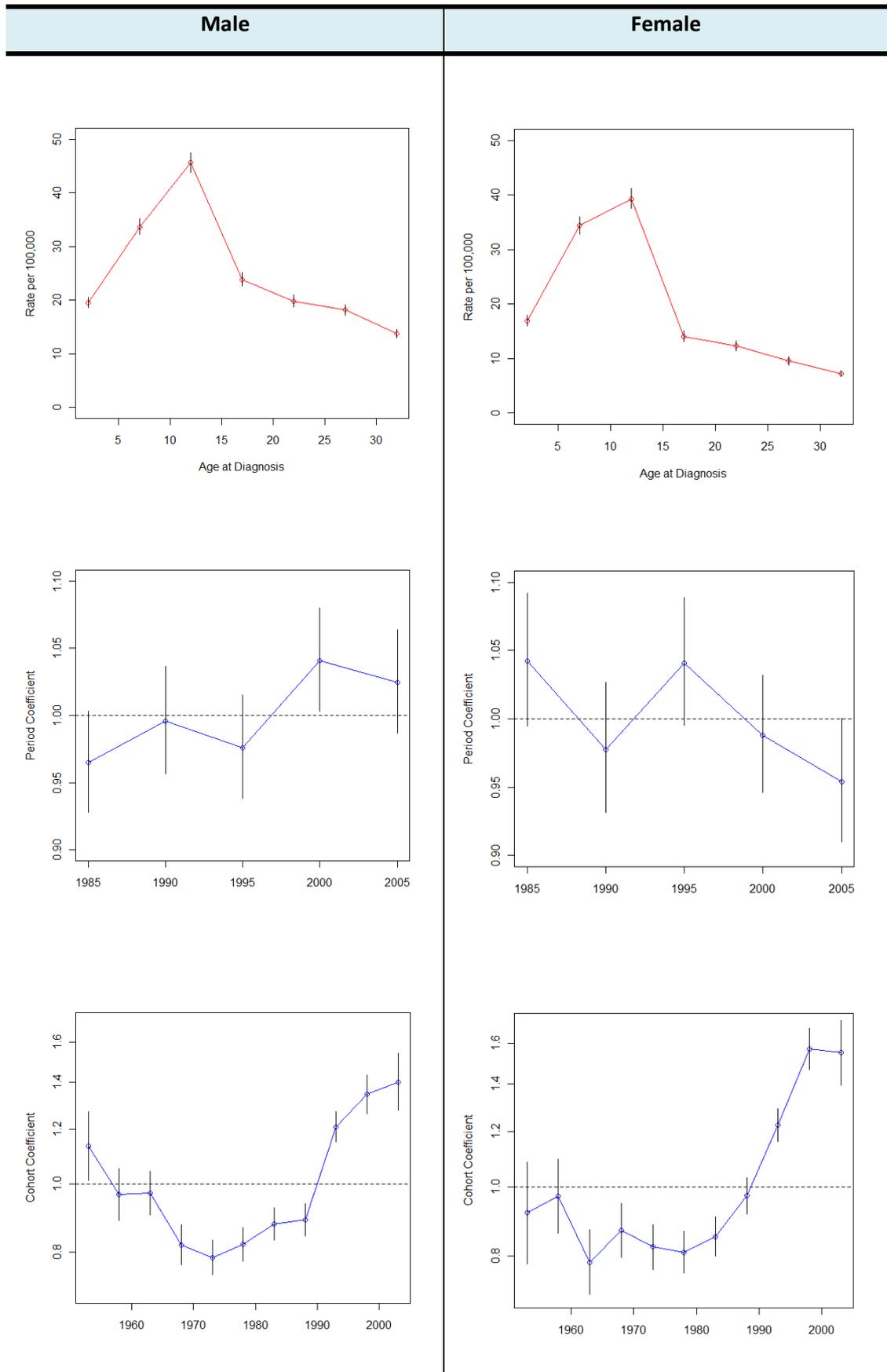


Figure 8.15: Coefficients from a full component PLS regression analysis.

Dahlquist et al. suggest birth cohort as the dominant factor (in comparison to period) based on a drift analysis and this is supported in the PCR/PLS results. The coefficients from the PCR/PLS regression are a rotation of the curvature components to incorporate a general trend that partitions the drift component. It would be possible to reduce the number of components retained in either estimation method. A demonstration of the benefits of this process has been previously illustrated in the lifecourse example. It is also important to note that the same intrinsic relationship would be retained regardless of the number of components entered into the regression. The results of this study would appear to suggest that birth cohort is playing an important role in dictating the onset of type-1 diabetes. This would seem to highlight the importance of exposures early in the lifecourse (whether at a prenatal or postnatal stage) to later life effects. The cohort risk begins high in the early years of the study, before a relatively low stable risk in the middle years followed by an increased risk in the later most recent cohorts. The authors put forward an explanation for the moderately high risk in the early years. They suggest that rather than biological significance, the result is most likely due to classification error, with different criteria of diagnosis and validation methods introduced in later years.

A final comment should be made with regard to those researchers who are interested in forecasting from APC models. For instance, actuaries will typically be interested in forecasting health care needs and the potential impact of health care proposals (Holford 1992). It appears dangerous to assume that what has happened in the past will be reflective of the future. This caution is perhaps only heightened by the sheer generality of the measurements made from the APC model. Early papers by Kermack et al. (1934) and Derrick (1927) attempted predictions for mortality rates in later years in England and Wales that it is now possible to test. These predictions were not close to what was actually observed. One of the main reasons is that diseases prominent at the time of the observed data were very much influenced by period effects specific to that time. Davey-Smith & Kuh (2001) reports that TB accounted for more than 30% of deaths for subjects aged 15-44 years and 15% for 45-64 years. Considering this, it is not surprising that the estimates were not accurate for data observed at a later date when TB was far less prominent. Prediction becomes yet another difficult part of APC work in trying to analyze macro effects based on environmentally changing variables such as period. This opens a new range of methodology such as the Lee-Carter model (Lee and Carter 1992; Li et al. 2009), the mixed model (O'Brien et al. 2008; Yang and Land 2008) and the autoregressive approach (Lee and Lin 1996a) which are beyond the scope of this work.

8.8 Conclusions

This chapter has looked to provide a framework of prominent approaches in the APC literature (although many more exist than it would be possible to cover). The running theme in the work on perfect collinearity, and perhaps the primary source of confusion regarding the problem, is that the population ‘truth’ is unavailable. Even if complete data were available and free from error it would still not be possible to identify three unique estimates and so it is necessary to understand exactly what solution it is that we are striving for. Whilst it may seem statistically nonsensical to attain estimates of the linearly dependent parameters, the underlying concepts that these macro level variables are said to represent are not necessarily confounded. Therefore, it may be considered that the aggregated data that are utilized is at the heart of the statistical problem.

“We must either limit our analysis to effects that are estimable, or else devise means whereby external information about the disease can be brought to bear”

(Holford 1992)

Data may be generated based on some population model and this model will represent the parameters that we are striving for. However, a statistically equivalent set of data can be produced by an infinite number of alternative models. These solutions each lie on a common ‘line of solutions’ in the geometry (when there is only one perfect dependency). Therefore, without *a priori* knowledge we are choosing a constraint that reflects one of an infinite number of statistically equivalent models (based on R_y^2).

Any deviation from the population coefficient will introduce a bias and we can have no certainty regarding the degree of bias added with each constraint. Estimable functions remain the only method that can provide complete confidence. It may be argued that when robust external information is not available, a PLS regression analysis would appear the most beneficial of the constraint based approaches. Perhaps the most informative feature is that the coefficients reflect the intrinsic structure of the data leading to a simpler interpretation. To assume any other structure would require a more complex interpretation of the underlying model. Also, a near identifiable model is likely to still suffer from the effects of collinearity and so the flexibility of these methods to retain fewer components can be of great benefit to the interpretation and variance shrinkage. However, the fact remains that with the study of APC data our hands remain somewhat tied by the perfect collinearity inherent in the data.

9. Conclusions and Future Developments

This thesis has considered three diverse problems linked by collinearity; (1) the ‘diagnosis’ of collinearity; (2) the formation of a dependency structure; and (3) the analysis of perfectly collinear predictors. Links between methods have developed through chapters studying what appear to be very different problems. A new technique is formed in the D-index and the matroid approach has been developed from first principles to be applied to a real epidemiological problem. The use of certain statistical methods has been critically assessed for particular applications and alternative techniques considered in their place. This section reviews the development process of the work and outlines some of the findings from the chapters. Possible extensions are suggested that were beyond the current scope and time limitations of this work.

The motivation for a new collinearity index was to move away from a ‘traditional’ measure of collinearity and instead focus on the impact of collinearity on the modelling process. The use of vector geometry to assess the role of model environment and the motivation of PLS to maximize the covariance on rotated axes provided a starting point for the development of an alternative index. From this idea, a ‘global’ measure was developed to identify whether a near dependency is present and geometrical angles (converted to correlations) used to determine the involvement of predictors in the dependencies. This method was developed from the bivariable model and extended to a general case independent of the visible geometry.

One of the interesting challenges with developing this index was finding existing methods to compare the results to. The VIF itself is not a ‘global’ measure, although it is easy to mistake it as one when only analyzing a bivariable model. Belsley (1991) highlights two major limitations of the VIF. Firstly, it will not identify the number of near dependencies (or the variables involved) and secondly it is based on some arbitrary cut-point to determine what should be considered “severe” collinearity. An arbitrary cut-point should be employed with extreme caution. Chapter 3 demonstrated that this can be

particularly dangerous as the VIF ignores important features of the data, such as the role of the response. Therefore, an extension was proposed for the VIF by multiplying it by the VDF. An arbitrary cut-point would then appear to have more foundation, although important issues associated with this practice cannot be avoided. The chosen threshold would preferably be application specific and make use of external knowledge of biological or clinical relationships. A threshold that may be considered suitable for one field is not likely to be transferable to all. As an alternative to the VIF, eigenvalues present a type of 'global' index to identify the number of near dependencies in the data. Due to limitations associated with eigenvectors, Belsley suggests the use of variance decomposition proportions (VDP) to determine covariate involvement in those dependencies.

All-subsets regression provides a different perspective with all rearrangements of predictors analyzed to identify how many dependencies exist and which covariates are involved. However, typically t-values and p-values are used to identify dependencies, which are themselves affected by collinearity. Utilizing the VIF was instead considered within an all-subsets framework. However, some global measure would be required to determine the strength of the dependency and an additional measure to indicate which covariates are involved (similar to what could be achieved by the CI and VDP). This is where a method such as the matroids would fit into the collinearity diagnosis section. In this method, all relationships are analyzed (for the example provided using a VIF criteria) and the 'essential flats' extracted to identify near dependencies present. The threshold value would then represent a type of 'global' collinearity measure. It would seem that this method provides the best comparison to achieving the benefits of the D-index.

To use the D-index as part of the matroid structure would provide an interesting challenge for the future. The power of the D-index is incorporating the information of the response into the analysis of the impact of collinearity. This did not fit with the MetS example in chapter 7. Although the methods have very close links it would still appear a challenge to see how this might work. The D-index would naturally replace the minimum VIF criteria in the process, however it would seem difficult to incorporate the information from the angles/correlations. The matroid approach could identify the number of dependencies using the 'global' index. The angles/correlations of the D-index can then be used to demonstrate covariate involvement in the relationships utilized as an additional fit measure (replacing the R_x^2 attached to each variable in our development of the matroid approach). Further work into the confidence intervals of the index is also required to extend to the general case and incorporate into a matroid structure.

The use of matroids as a collinearity diagnosis tool for the D-index was not the original motivation for developing the method. It was intended as a technique for identifying near dependencies in data (making use of existing collinearity indices). The matroid methodology remains at an early stage of development in comparison to EFA and VARCLUS. The Shen (2003) and Maison (1973) studies have demonstrated the potential of such a technique; however some key areas must be improved before it can be considered a viable option for MetS study. Greene (1990) suggests that the portion of subsets that change from 'independent' to 'dependent' to adhere to the matroid axioms would indicate the "consistency" of the analysis. Consistency appears to be important measure in this approach, which wouldn't be required for other methods in the field. It is one indicator of the suitability of the diagnostic measure entered into the framework. The use of Greene's consistency in the project seemed unreliable and so was left out of the analysis presented in this thesis. Whilst at some thresholds a greater number of flats changed status from independent to dependent, it didn't necessarily change the resulting structure in the LHD. Therefore, this measure would often appear overly pessimistic regarding the consistency of the structure. A bootstrap approach may be a viable option to measure overall consistency of the thresholds, however due to the computational costs attached in generating the matroid flats, a faster algorithm would need developing.

An R_x^2 statistic was introduced into the matroid diagram to demonstrate the fit of the variable to the flat in which it is assigned. This feature was motivated by the study of the VARCLUS methodology. The VARCLUS also has a measure based on the 2nd eigenvalue in the $\mathbf{X}^T\mathbf{X}$ of the subset that is used to determine when the next cluster would be split. Eigenvalues may be used to depict the fit of the variables to a subspace of the flat's dimensionality in the matroid technique. However, this particular option was not implemented in the matroid work as the simplicity of the analysis was the primary concern. The method was developed principally as a tool for *all* researchers, not only statisticians. Eigenvalues, VIFs and R_x^2 values could have complicated the interpretation from the analysis. It would also be useful to test the results of the matroid, VARCLUS and EFA directly using CFA. One of the benefits of the matroid and VARCLUS procedures is that the non-overlapping flats and cluster components respectively arrange in a structure primed for such an analysis. The results of a range of CFA for the different thresholds could also be added to the matroid diagram to demonstrate the strengths of the relations between flats. Model fit statistics would aid with the decision of which threshold to choose and subsequently how many flats are required to best describe the structure.

Finally, the problem of analysing perfectly collinear covariates in the lifecourse and APC analyses was considered. In chapter 7 a lifecourse example was used to demonstrate statistical results associated with latent variable methods and an APC example in chapter 8 to see where the methods would fit within the existing literature of methods. In particular the focus was on highlighting the links between a 'modern' method such as PLS to established mathematical results such as PCR, the MP g-inverse, the IE and curvature/drift. The continuous predictors of the lifecourse example provided an ideal starting point for the perfect collinearity work with one exact dependency preventing the estimation of the regression parameters. An important distinction was made between the arbitrary constraints associated with the infinite g-inverses that exist and the MP-inverse. In the former, the constraint is imposed on the estimation and in the latter it is maintained from the natural relationship in the data. The flexibility of these methods to reduce the dimensionality of the data provided a useful extension to the analysis of the problem to reduce the persistent effects associated with collinear data. Methods such as the D-index could provide a useful guide in the future as to the selection of these components based on interpretive rather than predictive measures (such as PRESS and Q^2).

The work in this thesis has been intended to explore novel ideas in the field of biostatistics and epidemiology. In these chapters only a starting point or an extension to current work in collinearity has been provided which leaves many areas unexplored. It would be interesting to view the results of the D-index as part of the matroid framework. The features of the index, such as the angles and components, would appear ideally suited to this method, which does require additional descriptive statistics. This challenge would certainly be one worth exploring once both approaches have been exposed in the public domain. The latent variable methods of PLS and PCA have been discussed throughout the work. It was not surprising to find links between PLS and the D-index with the goals to maximize and measure covariance with the response respectively. The use of the D-index to identify an 'optimal' number of components to enter would help to change the focus of component selection from a predictive to interpretive guide. This would also look to be a promising extension to the current work. Finally, the application of PLS to perfect collinearity problems is interesting. However, there is only so much information the tabulated data can provide. A comparison of approaches and analysing what assumptions are made with each would appear most beneficial to this field. This was the purpose of the final two chapters of the thesis. Searching for some optimal estimation procedure for lifecourse and APC would appear to be an impossible aim.

Abbreviations

ATP III	National Cholesterol Education Adult Treatment Panel III	MP	Moore-Penrose
BLUE	Best Linear Unbiased Estimator	NID	Normally and Independently Distributed
BMI	Body Mass Index	NIPALS	Non-linear Iterative PLS
BW	Birth Weight	OLS	Ordinary Least Squares
CD	Chronic Disease	PAF	Principal Axis Factoring
CFA	Confirmatory Factor Analysis	PCA	Principal Components Analysis
CI	Condition Index	PC Ins	Post-challenge Insulin
CVD	Cardiovascular Disease	PCR	Principal Components Regression
CW	Current Weight	PD	Path Diagram
DAG	Directed Acyclic Graph	PLS	Partial Least Squares Regression
DBP	Diastolic Blood Pressure	RCT	Randomized Controlled Trial
DOHaD	Developmental Origins of Health and Disease	SBP	Systolic Blood Pressure
EFA	Exploratory Factor Analysis	SEM	Structural Equation Model
FA	Factor Analysis	SES	Social Economic Status
FOAD	Foetal Origins of Adult Disease	SMK	Smoking Status
Glu	Glucose	SSR	Sum of Squared Residuals
HDL	High Density Lipoprotein	SVD	Singular Value Decomposition
Ins	Insulin	Trig	Triglycerides
IDF	International Diabetes Federation	T2-DM	Type-2 Diabetes
IE	Intrinsic Estimator	VC	Variance-covariance
ISI	Web of Knowledge	VDF	Variance Deflation Factor
MDE	Mean Dispersion Error	VDP	Variance Decomposition Proportions
MetS	Metabolic Syndrome	VIF	Variance Inflation Factor
MVUE	Minimum Variance Unbiased Estimator	WC	Weight Change
ML	Maximum Likelihood	WHR	Waist/Hip Ratio
		WHO	World Health Organization

References

- Adair, L.S. & Popkin, B. 2001. The Cebu Longitudinal Health and Nutrition Survey: history and major contributions of the project. *Philippine quarterly of culture and society*, 29, (1/2) 5-37
- Adair, L.S., Popkin, B.M., Akin, J.S., Guilkey, D.K., Gultiano, S., Borja, J., Perez, L., Kuzawa, C.W., McDade, T., & Hindin, M.J. 2011. Cohort profile: the Cebu longitudinal health and nutrition survey. *International Journal of Epidemiology*, 40, (3) 619-625
- Adair, L.S. & Prentice, A.M. 2004. A critical evaluation of the fetal origins hypothesis and its implications for developing countries. *Journal of Nutrition*, 134, (1) 191-193
- Alberti, G. 2005. Introduction to the metabolic syndrome. *European Heart Journal Supplements*, 7, (D) D3-D5
- Allen, D.M. 1974. Relationship Between Variable Selection and Data Augmentation and A Method for Prediction. *Technometrics*, 16, (1) 125-127
- Anderberg, M.R. 1973. *Cluster analysis for applications* San. Francisco, California, Academic Press
- Armstrong, J.S. 1967. Derivation of Theory by Means of Factor Analysis Or Swift,T and His Electric Factor Analysis Machine. *American Statistician*, 21, (5) 17-21
- Avanzolini, G., Barbini, P., Gnudi, G., & Grossi, A. 1991. Cluster-Analysis of Clinical-Data Measured in the Surgical Intensive-Care Unit. *Computer Methods and Programs in Biomedicine*, 35, (3) 157-170
- Barker, D.J.P. 1992. The Fetal Origins of Adult Hypertension. *Journal of Hypertension*, 10, S39-S44
- Barker, D.J.P. 2004. The developmental origins of adult disease. *Journal of the American College of Nutrition*, 23, (6) 588S-595S
- Barker, D.J.P., Gluckman, P.D., Godfrey, K.M., Harding, J.E., Owens, J.A., & Robinson, J.S. 1993. Fetal Nutrition and Cardiovascular-Disease in Adult Life. *Lancet*, 341, (8850) 938-941
- Barker, D.J.P., Osmond, C., Forsen, T.J., Kajantie, E., & Eriksson, J.G. 2005. Trajectories of growth among children who have coronary events as adults. *New England Journal of Medicine*, 353, (17) 1802-1809
-

-
- Belsley, D. 1991. *Conditioning Diagnostics: Collinearity and Weak Data in Regression* New York, Wiley-Interscience
- Ben-Shlomo, Y. & Kuh, D. 2002. A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. *International Journal of Epidemiology*, 31, (2) 285-293
- Bland, J.M. & Altman, D.G. 1994. Statistics Notes .2. Regression Towards the Mean. *British Medical Journal*, 308, (6942) 1499
- Burt, C. 1909. Experimental Tests of General Intelligence. *British Journal of Psychology*, 3, 94-177
- Burt, C. 1952. Tests of Significance in Factor Analysis. *British Journal of Psychology-Statistical Section*, 5, 109-133
- Canoy, D. 2010. Coronary Heart Disease and Body Fat Distribution. *Current Atherosclerosis Reports*, 12, (2) 125-133
- Carstensen, B. 2007. Age-period-cohort models for the Lexis diagram. *Statistics in Medicine*, 26, (15) 3018-3045
- Cattell, R.B. 1978. *The Scientific use of Factor Analysis in Behavioral and Life Sciences* New York, Plenum
- Cattell, R.B. 1983. Citation Classic - the Scree Test for the Number of Factors. *Current Contents/Social & Behavioral Sciences* (5) 16
- Chen, G., Liu, C., Chen, F., Yao, J., Jiang, Q., Chen, N., Huang, H., Liang, J., Li, L., & Lin, L. 2011. Body fat distribution and their associations with cardiovascular risk, insulin resistance and beta-cell function: are there differences between men and women? *International Journal of Clinical Practice*, 65, (5) 592-601
- Child, D. 1990. *The Essentials of Factor Analysis* London, Cassell Educational Ltd
- Clayton, D. & Schifflers, E. 1987. Models for Temporal Variation in Cancer Rates .2. Age Period Cohort Models. *Statistics in Medicine*, 6, (4) 469-481
- Cohen, J. 2003. *Applied Multiple Regression/Correlation Analysis for Behavioral Sciences* New York, Psychology Press
- Cole, T.J. 2010. Can Partial Least Squares Regression Separate the Effects of Body Size and Growth on Later Blood Pressure? Partial Least Squares Regression. *Epidemiology*, 21, (4) 449-451
- Cole, T.J. 2004. Modeling Postnatal Exposures and Their Interactions with Birth Size. *The Journal of Nutrition*, 134, (1) 201-204
- Cornfield, J. 1966. Sequential Trials, Sequential Analysis and Likelihood Principle. *American Statistician*, 20, (2) 18-23
-

Cornfield, J., Halperin, M., & GREENHOU.SW 1969. An Adaptive Procedure for Sequential Clinical Trials. *Journal of the American Statistical Association*, 64, (327) 759-770

Costello, A.B. & Osborne, J.W. 2005. Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research and Evaluation*

Cudeck, R. & Odell, L.L. 1994. Applications of Standard Error-Estimates in Unrestricted Factor-Analysis - Significance Tests for Factor Loadings and Correlations. *Psychological Bulletin*, 115, (3) 475-487

Cureton, E.E. & Mulaik, S.A. 1975. Weighted Varimax Rotation and Promax Rotation. *Psychometrika*, 40, (2) 183-195

Dahlquist, G. & Mustonen, L. 2000. Analysis of 20 years of prospective registration of childhood onset diabetes - time trends and birth cohort effects. *Acta Paediatrica*, 89, (10) 1231-1237

Dahlquist, G.G., Nystrom, L., & Patterson, C.C. 2011. Incidence of Type 1 Diabetes in Sweden Among Individuals Aged 0-34 Years, 1983-2007 An analysis of time trends. *Diabetes Care*, 34, (8) 1754-1759

Darsow, T., Kendall, D., & Maggs, D. 2006. Is the metabolic syndrome a real clinical entity and should it receive drug treatment? *Current Diabetes Reports*, 6, (5) 357-364

Davey-Smith, G. 2009. Smoking and lung cancer: causality, Cornfield and an early observational meta-analysis. *International Journal of Epidemiology*, 38, (5) 1169-1171

Davey-Smith, G. & Kuh, D. 2001. Commentary: William Ogilvy Kermack and the childhood origins of adult health and disease. *International Journal of Epidemiology*, 30, (4) 696-703

Dawkins, R. 2009. *The Greatest Show on Earth: The Evidence for Evolution*, First Edition ed. New York, Bantam Press

Dejong, S. 1993. Simpls - An Alternative Approach to Partial Least-Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, 18, (3) 251-263

Derksen, S. & Keselman, H.J. 1992. Backward, Forward and Stepwise Automated Subset-Selection Algorithms - Frequency of Obtaining Authentic and Noise Variables. *British Journal of Mathematical & Statistical Psychology*, 45, 265-282

Derrick, V.P.A. 1927. Observations on (1) Errors of Age in the Population Statistics of England and Wales, and (2) the Changes in Mortality indicated by the national records. *Journal of the Institute of Actuaries*, 58, (2) 117-159

Eriksson, J., Forsen, T., Tuomilehto, J., Osmond, C., & Barker, D. 2000. Fetal and childhood growth and hypertension in adult life. *Hypertension*, 36, (5) 790-794

Everitt, B.S., Gourlay, A.J., & Kendell, R.E. 1971. Attempt at Validation of Traditional Psychiatric Syndromes by Cluster Analysis. *British Journal of Psychiatry*, 119, (551) 399-412

-
- Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, (3) 272-299
- Fienberg, S.E. & Mason, W.M. 1982. *Specification and Implementation of Age, Period and Cohort Models* New York, Springer-Verlag
- Fisher, H. 2011. *A History of the Central Limit Theorem*, 1st Edition ed. New York, Springer
- Fisher, R.M. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10, 507-521
- Floyd, F.J. & Widaman, K.F. 1995. Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7, (3) 286-299
- Ford, E.S. & Giles, W.H. 2003. A comparison of the prevalence of the metabolic syndrome using two proposed definitions. *Diabetes Care*, 26, (3) 575-581
- Ford, J.K., MacCallum, R.C., & Tait, M. 1986. The Application of Exploratory Factor-Analysis in Applied-Psychology - A Critical-Review and Analysis. *Personnel Psychology*, 39, (2) 291-314
- Forouhi, N.G., Luan, J., Hennings, S., & Wareham, N.J. 2007. Incidence of Type 2 diabetes in England and its association with baseline impaired fasting glucose: The Ely study 1990-2000. *Diabetic Medicine*, 24, (2) 200-207
- Fox, J. & Monette, G. 1992. Generalized Collinearity Diagnostics. *Journal of the American Statistical Association*, 87, (417) 178-183
- Freund, R.J. & Wilson, W.J. 1998. *Regression Analysis: Statistical Modeling of a Response Variable* London, Academic Press
- Frost, W.H. 1939. The age selection of mortality from tuberculosis in successive decades. *American Journal of Hygiene*, 30, (1/3) 91-96
- Galton, F. 2003. A diagram of heredity (Reprinted). *Annals of Noninvasive Electrocardiology*, 8, (2) 171-172
- Gami, A.S., Witt, B.J., Howard, D.E., Erwin, P.J., Gami, L.A., Somers, V.K., & Montori, V.M. 2007. Metabolic syndrome and risk of incident cardiovascular events and death - A systematic review and meta-analysis of longitudinal studies. *Journal of the American College of Cardiology*, 49, (4) 403-414
- Gatignon, G. 2003. *Statistical Analysis of Management Data* New York, Springer
- Gillham, N.W. 2002. *A Life of Sir Francis Galton: From African Exploration to the Birth of Eugenics* New York, Oxford Univeristy Press
- Glantz, S.A. & Slinker, B.Y. 2001. *Applied Regression and Analysis of Variance* New York, McGraw-Hill
-

-
- Gluckman, P.D., Hanson, M.A., & Pinal, C. 2005. The developmental origins of adult disease. *Maternal and Child Nutrition*, 1, (3) 130-141
- Goodman, C.H. 1943. Factorial Analysis of Thurstone'S Seven Primary Abilities. *Psychometrika*, 8, (2) 121-129
- Gordis, L. 2000. *Epidemiology*, 2nd Edition ed. Philadelphia, W.B. Saunders Company
- Greene, T. 1990. The Depiction of Linear Association by Matroids. *Computational Statistics & Data Analysis*, 9, (3) 251-269
- Greene, T. 1991. Descriptively Sufficient Subcollections of Flats in Matroids. *Discrete Mathematics*, 87, (2) 149-161
- Gregg, E.W., Cheng, Y.L.J., Cadwell, B.L., Imperatore, G., Williams, D.E., Flegal, K.M., Narayan, K.M.V., & Williamson, D.F. 2005. Secular trends in cardiovascular disease risk factors according to body mass index in US adults (vol 293, pg 1868, 2005). *Jama-Journal of the American Medical Association*, 294, (2) 182
- Gujarati, D.N. 2002. *Basic Econometrics* McGraw-Hill Higher Education
- Hemingway, P. & Brereton, N. 2009. What is a Systematic Review? *Haywood Medical Communications*
- Henriksen, T. & Clausen, T. 2002. The fetal origins hypothesis: placental insufficiency and inheritance versus maternal malnutrition in well-nourished populations. *Acta Obstetrica et Gynecologica Scandinavica*, 81, (2) 112-114
- Herr, D.G. 1980. History of the Use of Geometry in the General Linear-Model. *American Statistician*, 34, (1) 43-47
- Hobcraft, J., Menken, J., & Preston, S. 1982. Age, Period, and Cohort Effects in Demography - A Review. *Population Index*, 48, (1) 4-43
- Hocking, R.R. 2003. *Methods and Applications of Linear Models* New Jersey, Wiley-Blackwell
- Hoerl, A.E. 1962. Application of Ridge Analysis for Characterizing and Solving Almost Indeterminate Linear Simultaneous Equations. *Siam Review*, 4, (2) 179-&
- Holford, T.R. 1985. An Alternative Approach to Statistical Age-Period-Cohort Analysis. *Journal of Chronic Diseases*, 38, (10) 831-836
- Holford, T.R. 1992. Analysing the temporal effects of age, period and cohort. *Statistical Methods in Medical Research*, 1, (3) 317-337
- Holford, T. R. 2005, "Age-Period-Cohort Analysis," *In Encyclopedia of Biostatistics*, John Wiley & Sons, Ltd.
- Hortobagyi, T., Israel, R.G., Houmard, J.A., Mccammon, M.R., & Obrien, K.F. 1992. Comparison of Body-Composition Assessment by Hydrodensitometry, Skinfolds, and
-

Multiple Site Near-Infrared Spectrophotometry. *European Journal of Clinical Nutrition*, 46, (3) 205-211

Humphreys, L.G. & Montanelli, R.G. 1975. An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, 10, 193-205

Huxley, R., Neil, A., & Collins, R. 2002. Unravelling the fetal origins hypothesis: is there really an inverse association between birthweight and subsequent blood pressure? *Lancet*, 360, (9334) 659-665

Jackson, J.E. 2003. *A User's Guide to Principal Components* New Jersey, Wiley-Interscience

Johnson, D.E. & Wichern, D.W. 2002. *Applied Multivariate Statistical Analyses*, 5th Edition ed. NJ, Prentice Hall

Joliffe, I.T. 2002. *Principal Component Analysis*, 2nd Edition ed. New York, Springer-Verlag

Kahn, R. 2007. Is the metabolic syndrome a real syndrome? *Circulation*, 115, (13) 1806-1811

Kaiser, H.F. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23, 187-200

Kaplan, G.A. & Keil, J.E. 1993. Socioeconomic-Factors and Cardiovascular-Disease - A Review of the Literature. *Circulation*, 88, (4) 1973-1998

Kaufman, L. & Rosseeuw, P.J. 1990. *Finding Groups in Data* California, Wiley

Kermack, W.O., McKendrick, A.G., & McKinlay, P.L. 1934. Death-rates in Great Britain and Sweden - Some general regularities and their significance. *Lancet*, 1, 698-703

Keyes, K.M., Utz, R.L., Robinson, W., & Li, G.H. 2010. What is a cohort effect? Comparison of three statistical methods for modeling cohort effects in obesity prevalence in the United States, 1971-2006. *Social Science & Medicine*, 70, (7) 1100-1108

Khattree, R. & Naik, D.N. 2000. *Multivariate data reduction and discrimination with SAS software* Cary, NC, SAS Institute Inc.

Kirkwood, B. & Stern, J.A.C. 2003. *Essential Medical Statistics* Oxford, Blackwell

Kupper, L.L., Janis, J.M., Karmous, A., & Greenberg, B.G. 1985a. An Alternative Approach to Statistical Age-Period-Cohort Analysis - Reply. *Journal of Chronic Diseases*, 38, (10) 837-840

Kupper, L.L., Janis, J.M., Karmous, A., & Greenberg, B.G. 1985b. Statistical Age-Period-Cohort Analysis - A Review and Critique. *Journal of Chronic Diseases*, 38, (10) 811-830

Kupper, L.L., Janis, J.M., Salama, I.A., Yoshizawa, C.N., & Greenberg, B.G. 1983. Age-Period-Cohort Analysis - An Illustration of the Problems in Assessing Interaction in One

Observation Per Cell Data. *Communications in Statistics-Theory and Methods*, 12, (23) 2779-2807

Kylin, E. 1923. Studien uber das Hypertonie-Hyperglyka 'mie-Hyperurika' miesyndrom. *Zentralbl Inn Med*, 44, 105-127

Lafortuna, C.L., Adorni, F., Agosti, F., & Sartorio, A. 2008. Factor analysis of metabolic syndrome components in obese women. *Nutrition Metabolism and Cardiovascular Diseases*, 18, (3) 233-241

Law, C.M. & Shiell, A.W. 1996. Is blood pressure inversely related to birth weight? The strength of evidence from a systematic review of the literature. *Journal of Hypertension*, 14, (8) 935-941

Lawlor, D.A., Ebrahim, S., & Davey-Smith, G. 2002. Socioeconomic position in childhood and adulthood and insulin resistance: cross sectional survey using data from British women's heart and health study. *British Medical Journal*, 325, (7368) 805-807

Lawlor, D.A., Ebrahim, S., May, M., & Davey-Smith, G. 2004. (Mis)use of factor analysis in the study of insulin resistance syndrome. *American Journal of Epidemiology*, 159, (11) 1013-1018

Lee, R.D. & Carter, L.R. 1992. Modeling and Forecasting United-States Mortality. *Journal of the American Statistical Association*, 87, (419) 659-671

Lee, W.C. & Lin, R.S. 1996a. Autoregressive age-period-cohort models. *Statistics in Medicine*, 15, (3) 273-281

Lee, W.C. & Lin, R.S. 1996b. Modelling the age-period-cohort trend surface. *Biometrical Journal*, 38, (1) 97-106

Leng, C.L. & Wang, H.S. 2009. On General Adaptive Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 18, (1) 201-215

Li, J.S.H., Hardy, M.R., & Tan, K.S. 2009. Uncertainty in Mortality Forecasting: An Extension to the Classical Lee-Carter Approach. *Astin Bulletin*, 39, (1) 137-164

Lindsay, R.S., Dabelea, D., Roumain, J., Hanson, R.L., Bennett, P.H., & Knowler, W.C. 2000. Type 2 diabetes and low birth weight - The role of paternal inheritance in the association of low birth weight and diabetes. *Diabetes*, 49, (3) 445-449

Lithell, H.O., McKeigue, P.M., Berglund, L., Mohsen, R., Lithell, U.B., & Leon, D.A. 1996. Relation of size at birth to non-insulin dependent diabetes and insulin concentrations in men aged 50-60 years. *British Medical Journal*, 312, (7028) 406-410

Loucks, E.B., Magnusson, K.T., Cook, S., Rehkopf, D.H., Ford, E.S., & Berkman, L.F. 2007. Socioeconomic position and the metabolic syndrome in early, middle, and late life: Evidence from NHANES 1999-2002. *Annals of Epidemiology*, 17, (10) 782-790

Lucas, A., Fewtrell, M.S., & Cole, T.J. 1999. Education and debate - Fetal origins of adult disease - the hypothesis revisited. *British Medical Journal*, 319, (7204) 245-249

-
- Mackie, J.L. 1974. *The Cement of the Universe: A Study of Causation* Oxford, Oxford University Press
- Mackinnon, D. P., Krull, J. L., and Lockwood, C. M. 2000. Equivalence of the mediation, confounding and suppression effect. *Prevention science the official journal of the Society for Prevention Research*, 1, (4) 173-181
- Maison, P., Byrne, C.D., Hales, C.N., Day, N.E., & Wareham, N.J. 2001. Do different dimensions of the metabolic syndrome change together over time? Evidence supporting obesity as the central feature. *Diabetes Care*, 24, (10) 1758-1763
- Manne, R. 1987. Analysis of 2 Partial-Least-Squares Algorithms for Multivariate Calibration. *Chemometrics and Intelligent Laboratory Systems*, 2, (1-3) 187-197
- Mannucci, E., Monami, M., & Rotella, C.M. 2007. How many components for the metabolic syndrome? Results of exploratory factor analysis in the FIBAR study. *Nutrition Metabolism and Cardiovascular Diseases*, 17, (10) 719-726
- Marquardt, D.W. & Snee, R.D. 1975. Ridge Regression in Practice. *American Statistician*, 29, (1) 3-20
- Mason, K.O. & WINSBORO.HH 1973. Some Methodological Issues in Cohort Analysis of Archival Data. *American Sociological Review*, 38, (2) 242-258
- Mayr, E. 1970. *Populations, Species and Evolution: An Abridgment of Animal Species and Evolution* Belknap Press of Harvard University Press
- Mazumdar, S., Li, C.C., & Bryce, G.R. 1980. Correspondence Between A Linear Restriction and A Generalized Inverse in Linear-Model Analysis. *American Statistician*, 34, (2) 103-105
- McNeill, A.M., Katz, R., Girman, C.J., Rosamond, W.D., Wagenknecht, L.E., Barzilay, J.I., Tracy, R.P., Savage, P.J., & Jackson, S.A. 2006. Metabolic syndrome and cardiovascular disease in older people: The cardiovascular health study. *Journal of the American Geriatrics Society*, 54, (9) 1317-1324
- Meyer, C.D. 2001. *Matrix Analysis and Applied Linear Algebra and Solutions Manual* Philadelphia, Society for Industrial and Applied Mathematics
- Mortimer, J.T. & Shanahan, M.J. 2006. *Handbook of the Life Course* New York, Springer Science
- Mundi, M.S., Karpyak, M.V., Koutsari, C., Votruba, S.B., O'Brien, P.C., & Jensen, M.D. 2010. Body Fat Distribution, Adipocyte Size, and Metabolic Characteristics of Nondiabetic Adults. *Journal of Clinical Endocrinology & Metabolism*, 95, (1) 67-73
- Myers, R.H. 1990. *Classical and Modern Regression with Applications*, 2 ed. PWS-KENT
- O'Brien, R.M. 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41, (5) 673-690
-

-
- O'Brien, R.M. 2011. Constrained Estimators and Age-Period-Cohort Models. *Sociological Methods & Research*, 40, (3) 419-452
- O'Brien, R.M., Hudson, K., & Stockard, J. 2008. A mixed model estimation of age, period, and cohort effects. *Sociological Methods & Research*, 36, (3) 402-428
- Ong, K.K.L., Ahmed, M.L., Emmett, P.M., Preece, M.A., & Dunger, D.B. 2000. Association between postnatal catch-up growth and obesity in childhood: prospective cohort study. *British Medical Journal*, 320, (7240) 967-971
- Osmond, C. 1996. An appreciation of "cohort analysis of mortality rates as an historical or narrative technique" (RAM Case). *Journal of Epidemiology and Community Health*, 50, (2) 125-126
- Osmond, C. & Gardner, M.J. 1989. Age, Period, and Cohort Models - Non-Overlapping Cohorts Dont Resolve the Identification Problem. *American Journal of Epidemiology*, 129, (1) 31-35
- Osmond, C. & Hardy, R. 2004, "Ischemic heart disease and cerebrovascular disease mortality trends with special reference to England and Wales: are there cohort effects?," In *A LIFE COURSE APPROACH TO chronic disease epidemiology*, 2 ed. D. Kuh & Ben-Schlomo.Y, eds., New York: Oxford University Press, pp. 116-143.
- Pearl, J. 2000. *Causality: models, reasoning and inference* Cambridge, Cambridge University Press
- Penrose, K.W., Nelson, A.G., & Fisher, A.G. 1985. Generalized Body-Composition Prediction Equation for Men Using Simple Measurement Techniques. *Medicine and Science in Sports and Exercise*, 17, (2) 189
- Phatak, A. & DeJong, S. 1997. The geometry of partial least squares. *Journal of Chemometrics*, 11, (4) 311-338
- Phatak, A. & Dejong, S. 1997. The geometry of partial least squares. *Journal of Chemometrics*, 11, (4) 311-338
- Pitkaniemi, J., Onkamo, P., Tuomilehto, J., & Arjas, E. 2004. Increasing incidence of Type 1 diabetes - role for genes? *Bmc Genetics*, 5,
- Pladevall, M., Singal, B., Williams, L.K., Brotons, C., Guyer, H., Sadurni, J., Falces, C., Serrano-Rios, M., Gabriel, R., Shaw, J.E., Zimmet, P.Z., & Haffner, S. 2006. A single factor underlies the metabolic syndrome - A confirmatory factor analysis. *Diabetes Care*, 29, (1) 113-122
- Preacher, K. J. and MacCallum, R. C. 2003. Repairing Tom Swift's Electric Factor Analysis Machine. *Understanding Statistics*, 2, (1) 13-43
- Rao, C.R. 1999. *Linear Models: Least Squares and Alternatives* New York, Springer
- Rawlings, J.O., Pantula, S.G., & Dickey, D.A. 1998. *Applied Regression Analysis: A Research Tool* New York, Springer-Verlag
-

- Rich-Edwards, J.W., Stampfer, M.J., Manson, J.E., Rosner, B., Hankinson, S.E., Colditz, G.A., Willett, W.C., & Hennekens, C.H. 1997. Birth weight and risk of cardiovascular disease in a cohort of women followed up since 1976. *British Medical Journal*, 315, (7105) 396-400
- Robertson, C. & Boyle, P. 1986. Age, Period and Cohort Models - the Use of Individual Records. *Statistics in Medicine*, 5, (5) 527-538
- Robertson, C., Gandini, S., & Boyle, P. 1999. Age-period-cohort models: A comparative study of available methodologies. *Journal of Clinical Epidemiology*, 52, (6) 569-583
- Rodgers, W.L. 1982. Estimable Functions of Age, Period, and Cohort Effects. *American Sociological Review*, 47, (6) 774-787
- Romero-Corral, A., Somers, V.K., Sierra-Johnson, J., Korenfeld, Y., Boarin, S., Korinek, J., Jensen, M.D., Parati, G., & Lopez-Jimenez, F. 2010. Normal weight obesity: a risk factor for cardiometabolic dysregulation and cardiovascular mortality. *European Heart Journal*, 31, (6) 737-746
- Rothman, K.J. & Greenland, S. 2005. Causation and causal inference in epidemiology. *American Journal of Public Health*, 95, S144-S150
- Rothman, K.J., Greenland, S., & Lash, T.L. 1998. *Modern Epidemiology*, 2nd Edition ed. Philadelphia, Lippincott Williams & Wilkins
- Sandhu, M.S., Heald, A.H., Gibson, J.M., Cruickshank, J.K., Dunger, D.B., & Wareham, N.J. 2002. Circulating concentrations of insulin-like growth factor-I and development of glucose intolerance: a prospective observational study. *Lancet*, 359, (9319) 1740-1745
- Scheutz, F. & Poulsen, S. 1999. Determining causation in epidemiology. *Community Dentistry and Oral Epidemiology*, 27, (3) 161-170
- Shah, S., Novak, S., & Stapleton, L.A. 2006. Evaluation and comparison of models of metabolic syndrome using confirmatory factor analysis. *European Journal of Epidemiology*, 21, (5) 343-349
- Shen, B.J., Goldberg, R.B., Llabre, M.M., & Schneiderman, N. 2006. Is the factor structure of the metabolic syndrome comparable between men and women and across three ethnic groups: The Miami Community Health Study. *Annals of Epidemiology*, 16, (2) 131-137
- Shen, B.J., Todaro, J.F., Niaura, R., McCaffery, J.M., Zhang, J.P., Spiro, A., & Ward, K.D. 2003. Are metabolic risk factors one unified syndrome? Modeling the structure of the metabolic syndrome X. *American Journal of Epidemiology*, 157, (8) 701-711
- Smith, G.D. & Kuh, D. 2001. Commentary: William Ogilvy Kermack and the childhood origins of adult health and disease. *International Journal of Epidemiology*, 30, (4) 696-703
- Spearman, C. 1904. "General intelligence " objectively determined and measured. *American Journal of Psychology*, 15, 201-292
- Stigler, S.M. 1990. *The History of Statistics* Belknap Press of Harvard University Press
-

-
- Streiner, D.L. 1994. Figuring Out Factors - the Use and Misuse of Factor-Analysis. *Canadian Journal of Psychiatry-Revue Canadienne de Psychiatrie*, 39, (3) 135-140
- Tang, W.H., Pankow, J.S., & Arnett, D.K. 2005. Re: "(MIS) use of factor analysis in the study of insulin resistance syndrome". *American Journal of Epidemiology*, 162, (9) 921-922
- Thurstone, L.L. 1934. The Vectors of Mind. *Psychological Review*, 41, 1-32
- Thurstone, L.L. 1940. Experimental Study of Simple Structure. *Psychometrika*, 5, (2) 153-168
- Thurstone, L.L. 1947. *Multiple-Factor Analysis* Chicago, University of Chicago Press
- Thurstone, L.L. 1948. Primary Mental Abilities. *Science*, 108, (2813) 585
- Tryon, R.C. 1935. A theory of psychological components - An alternative to 'Mathematical factors'. *Psychological Review*, 42, 425-454
- Tryon, R.C. & Bailey, D.E. 1970. *Cluster Analysis* New York, McGraw-Hill
- Tu, Y.K. & Gilthorpe, M.S. 2007. Revisiting the relation between change and initial value: A review and evaluation. *Statistics in Medicine*, 26, (2) 443-457
- Tu, Y.K. & Law, G.R. 2010. Re-examining the associations between family and children's cognitive developments in early ages. *Early Child Development* 1243-1252
- Tu, Y.K., Maddick, I.H., Griffiths, G.S., & Gilthorpe, M.S. 2004. Mathematical coupling can undermine the statistical assessment of clinical research: illustration from the treatment of guided tissue regeneration (vol, 32, pg 133, 2004). *Journal of Dentistry*, 32, (4) 339-340
- Tu, Y.K., Woolston, A., Baxter, P.D., & Gilthorpe, M.S. 2010. Assessing the Impact of Body Size in Childhood and Adolescence on Blood Pressure An Application of Partial Least Squares Regression. *Epidemiology*, 21, (4) 440-448
- Tudor-Locke, C., Ainsworth, B.E., Adair, L.S., & Popkin, B.M. 2003. Physical activity in Filipino youth: the Cebu Longitudinal Health and Nutrition Survey. *International journal of obesity*, 27, (2) 181-190
- Tukey, J.W. 1958. Bias and Confidence in Not-Quite Large Samples. *Annals of Mathematical Statistics*, 29, (2) 614
- Weinkam, J.J. & Sterling, T.D. 1991. A graphical approach to the interpretation of age-period-cohort data. *Epidemiology*, 2, 133-137
- Welsh, D. 1976. *Matroid theory* London, Academic P
- Whitney, H. 1935. On the abstract properties of linear dependence. *American Journal of Mathematics*, 57, 509-533
- Wickens, T.D. 1995. *The Geometry of Multivariate Statistics* London, Psychology Press
-

Wilson, P.W.F., D'Agostino, R.B., Parise, H., Sullivan, L., & Meigs, J.B. 2005. Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus. *Circulation*, 112, (20) 3066-3072

Wonnacott, T.H. 1981. *Regression: A Second Course in Statistics* Florida, Krieger Publishing Company

Wood, J.M., Tataryn, D.J., & Gorsuch, R.L. 1996. Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods*, 1, (4) 354-365

Yang, Y., Fu, W.J.J., & Land, K.C. 2004. A methodological comparison of age-period-cohort models: The intrinsic estimator and conventional generalized linear models. *Sociological Methodology*, 2004, Vol 34, 34, 75-110

Yang, Y. & Land, K.C. 2008. Age-period-cohort analysis of repeated cross-section surveys - Fixed or random effects? *Sociological Methods & Research*, 36, (3) 297-326

Yang, Y., Schulhofer-Wohl, S., Fu, W.J.J., & Land, K.C. 2008. The intrinsic estimator for age-period-cohort analysis: What it is and how to use it. *American Journal of Sociology*, 113, (6) 1697-1736
