# Economic Applications of Nonparametric Methods

By
Giovanni Baiocchi

SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF YORK
DEPARTMENT OF ECONOMICS AND RELATED STUDIES
June 2006

*To Mauro Cuppo*

# Abstract

This thesis deals with the subject of nonparametric methods, focusing on application to economic issues.

Chapter 2 introduces the basic nonparametric methods underlying the applications in the subsequent chapters.

In Chapter 3 we propose some basic standards to improve the use and reporting of nonparametric methods in the statistics and economics literature for the purpose of accuracy and reproducibility. We make recommendations on four aspects of the application of nonparametric methods: computational practice, published reporting, numerical accuracy, and visualization.

In Chapter 4 we investigate the effect of life-cycle factors and other demographic characteristics on income inequality in the UK. Two conditional inequality measures are derived from estimating the cumulative distribution function of household income, conditional upon a broad set of explanatory variables. Estimation of the distribution is carried out using a semiparametric approach. The proposed inequality estimators are easily interpretable and are shown to be consistent. Our results indicate the importance of inter-family differences in the analysis of income distribution. In addition, our estimation procedure uncovers higher-order properties of the income distribution and non-linearities of its moments that cannot be captured by means of a "standard" parametric approach. Several features of the conditional distribution of income are highlighted.

Chapter 5 we reexamine the relationship between openness to trade and the environment, controlling for economic development, in order to identify the presence of multiple regimes in the cross-country pollution-economic relationship. We first identify the presence of multiple regimes by using specification tests which entertain a single regime model as the null hypothesis. Then we develop an easily interpretable measure, based on an original application of the Blinder-Oaxaca decomposition, of the impact on the en-

vironment due to differences in regimes. Finally we apply a nonparametric recursive partitioning algorithm to endogenously identify various regimes. Our conclusions are threefold. First, we reject the null hypothesis that all countries obey a common linear model. Second, we find that quantitatively regime differences can have a significant impact. Thirdly, by using regression tree analysis we find subsets of countries which appear to possess very different environmental/economic relationships.

In Chapter 6 investigate the existence of the so called *environmental kuznets curve* (EKC), the inverted-U shaped relationship between income and pollution, using nonparametric regression and a threshold regression methods.We find support for threshold models that lead to different reduced-form relationships between environmental quality and economic activity when early stages of economic growth are contrasted with later stages. There is no evidence of a common inverted U-shaped environment/economy relationship that all country follow as they grow. We also find that changes that might benefit the environment occur at much higher levels of income than those implied by standard models. Our findings support models in which improvements are a consequence of the deliberate introduction of policies addressing environmental concerns. Moreover, we find evidence that countries with low-income levels have a far greater variability in emissions per capita than high-income countries. This has the implication that it may be more difficult to predict emission levels for low-income countries approaching the turning point.

A summary of the main findings and further research directions are presented in Chapter 7 and in Chapter 8, respectively.

# Contents

# List of Figures

# List of Tables

# Acknowledgements

A large number of people played an important part in the completion of this thesis. My gratitude goes to my supervisor Huw Dixon. His personal attitude and enthusiasm contributed so much in shaping my own ideas about research and academic profession at large. I thank him for his patience, constant support, and availability. Other members of staff at the Department of Economics and Related Studies in York deserve to be mentioned. First of all the members of the Thesis Advisory Group. John Hutton and Karim Abadir helped me improve the earlier versions of the chapters and provided a challenging and stimulating supervision. I also benefited from comments and suggestions from my colleague, coauthor, and dear friend Walter Distaso.

Writing this thesis has taken me many years. I am grateful to all my fellow post-graduate colleagues in the Economics Department for contributing to a very pleasant and productive atmosphere. In a special way, Antonio Giuffrida, Mariella Cabizza, Maria Vittoria Levati, Paolo Lupi, Fabio Manenti, Borislava Mihaylova, Maria Teresa Monteduro, Francesca Perrone, Gennaro Scarfiglieri and Ernesto Somma, for their friendship since early days.

During my time at York I made very good friends in the Environment Department where I worked as a lecturer for many years. I have fond memories of late night sessions with Jan Minx discussing about science, economics, and other subjects. During that period, Fabian Capitanio, Chiung-Ting Chang, Silvana Dalmazzone, Caterina De Lucia, Cinzia Faiella, Katia Martin, Giuseppe Nocella, Pasquale Pazienza, and Riccardo Scarpa, became very good friends and helped me in several respects. I also want to thank my colleague and friend Ashar Aftab, from the University of Durham, for making my current job there as a lecturer more bearable.

Finally, I want to thank the most important persons in my life: my mother Nevia, my sister Michela, all my relatives, and Ikuko for all their constant support.

# Declaration

Background chapters 2 and 3 have been used as notes in the undergraduate course *Computing in Statistics and Econometrics* taught in the Department of Economics in the years 1999–2003, in the course for research students in economics and finance *Computing Skills for Economists*, thought in the Economics Department of the University of York in the summer of 2004, and in the course for research students in environmental sciences *Advanced Research Methods*, thought in the Environment Department of the University of York in the Summer of 2005.

A extended version generalized to computational economics of Chapter 3 has been published in Baiocchi (2007).

Applied chapter 4 is the result of a joint effort with Walter Distaso (Tanaka Business School, Imperial College London). The authors contributed equally to the final product. Chapter 4 has been presented at the International Conference (April 10-11, 2003) On the Wealth of Nations - Extending the Tinbergen Heritage. The chapter is based on the project "A Conditional Density Estimation Approach to Polarization in the EU", prepared together with Walter Distaso, that was funded by IRISS (Integrated Research Infrastructure in the Socio-Economic Sciences) at the CEPS/INSTEAD (Centre d'Études de Populations, de Pauvreté et de Politiques Socio-Économiques / International Networks for Studies in Technology, Environment, Alternatives, Development), based in Luxembourg. The paper has been submitted to a peer reviewed economic journal for publication.

Chapter 5 is an extension of the papers:

- 'Economic Growth, Trade policies, and the Environment: An Endogenous Determination of Multiple Cross-Country Regimes" and
- "Towards Explaining the "Pollution" Gap Between Rich and Poor Countries: Testing the Pollution Displacement Hypothesis."

The first paper was accepted for presentation at Thirteenth Annual Conference (2004) of the European Association of Environmental and Resource Economists (EAERE) in Budapest, Hungary, and has been presented in the Environment Department of the York University, UK, (21st November 2003) and the Department of Environmental Studies (1st December 2003) in Dublin, Ireland.

The latter paper was presented at the Second World Congress (2002) of the Association of Environmental and Resource Economists (AERE), and the European Association of Environmental and Resource Economists (EAERE), in Monteray, California. It was submitted to a peer reviewed economic journal for publication under the name "A Decomposition of the Pollution Gap Between Rich and Poor Countries."

Chapter 6 is based on two papers, namely:

- "Investigating the Shape of the EKC: A Nonparametric Approach," and
- "The Relationship Between Growth and Environment: Should we be Looking for Turning or Break Points?"

The first is based on a paper together with Salvatore Di Falco. That paper has been published as a working paper by Fondazione Eni Enrico Mattei (Nota di Lavoro 66, 2001). The paper has also been presented in the First World Congress (2000) of the Association of Environmental and Resource Economists (AERE), and the European Association of Environmental and Resource Economists (EAERE).

The latter was presented at a research workshop in Economics held in Department of Economics and Finance, University of Durham, on the 10th March 2005, will be presented at the Third World Congress of Environmental and Resource Economists in Kyoto, Japan (3rd -7th July 2006), and has been submitted to a peer reviewed economic journal for publication.

Chapter 1

# Introduction

This thesis is about the practice, and visualisation of nonparametric econometrics. The primary objective is to to apply nonparametric and semiparametric methods to relevant economic issues. Though nonparametric and semiparametric models have received considerable attention from theoretical econometricians, they were still used only sparingly by applied economists until recently. There are a few possible explanations for this apparent initial lack of interest from practitioners. In comparison with constructing an histogram or fitting a linear model, nonparametric and semiparametric methods can be theoretically more advanced and often, especially in the past, require relatively more advanced computer programming skills. Also, because of the nature of the estimated functional relationships, traditional tabular formats used to report econometric results have become less useful. More often, computational results can be communicated accurately and clearly only by means of graphs. Because of the nature of the computed results visualization has become an essential part of nonparametric econometrics. A different set of tools coming from a variety of disciplines is needed to apply these methods to the solution of economic problems. This thesis acknowledges the multidisciplinarity of the subject by drawing on research from economics, mathematical statistics, numerical analysis, computer programming, and computer graphics.

The topic of nonparametric and semiparametric methods is too vast and and complex to be given an exhaustive treatment in a doctoral thesis, indeed many have been already written on the topic. In this thesis emphasis will be given to methods that enable the inclusion of multiple explanatory variables without suffering of the so called "curse of dimensionality" problem. Also, more traditional parametric based methods will be used to support and strengthen nonparametric results. As Scott (1992) points out: "there is a natural flow among the parametric, exploratory, and nonparametric procedures that represent a rational approach to statistical data analysis. Begin with a fully exploratory point of view in order to obtain an overview of the data. If a probabilistic structure is present, estimate that structure nonparametrically and explore it visually. Finally, if a linear model appears adequate, adopt a fully parametric approach." We will attempt to follow this precept

as closely as possible.

The dataset used in the examples and in the main applications are another contribution of this thesis. They were all prepared from the original sources and took a considerable amount of time to prepare.

This thesis is organized in three main parts.

Part I introduces the fundamental concepts underlying the analyses of subsequent chapters. Chapter 2 purports to provide an introduction to the basic nonparametric methods. In this Chapter we introduce the distinction between parametric and nonparametric models. We also highlight the importance of visualization when applying nonparametric methods. Several original applications are provide to illustrate the use of nonparametric methods in economics. For instance, Example 4 introduces an original methodology to estimate a conditional density with an application to labor economics. Also, Example 3 was contributed to the forthcoming book by Li & Racine (2006), Nonparametric Econometrics: Theory and Practice, Princeton University Press. In Chapter 3 we propose some basic standards to improve the use and reporting of nonparametric methods in the statistics and economics literature for the purpose of accuracy and reproducibility. In this Chapter we make recommendations on four aspects of the process: computational practice, published reporting, numerical accuracy, and visualization.

Part II presents the main economic applications of nonparametric methods. In Chapter 4 we investigate the effect of demographic and socio-economic characteristics of households on income inequality in the UK. We propose the use of a semiparametric method to estimate conditional measures of inequality from an estimate of a conditional distribution, in order to control for different determinants of income inequality. To estimate the conditional distribution, we resort to the semiparametric method developed by Foresi & Peracchi (1995). Conditional quantiles are obtained by inverting the estimated conditional distribution and conditional measures of income inequality are derived from the conditional quantiles.

The data used in the analysis have been taken from the database produced by the Consortium of Household panels for European socio-economic

Research (CHER).[1] The CHER database for United Kingdom (UK) is based upon the results of the British Household Panel Survey (BHPS), which is carried out in the UK annually over a target sample size of 5000 households.

Our approach is novel in at least four respects. First, by estimating the entire conditional distribution of income over a broad set of determinants, our estimation procedure uncovers higher-order properties of the income distribution and non-linearities of its moments that cannot be captured by means of a "standard" parametric approach. For example, similar to the results obtained in the previous literature, we find that the shape of the age-income profiles agrees with the observable prediction of the life-cycle model, which assumes that resources are accumulated at a faster rate at a young age. Also, we find that income of families during the period of child rearing is higher than income in the retirement stage of the life-cycle, when economic responsibility is greatly reduced. In addition, we find that the age-income profiles peak later for the wealthier households and appear considerably non-linear, declining rapidly after the age of 50. Besides having important consequences for the policy maker as such, this asymmetry might also indicate the presence of different factors affecting the upward and downward branches of the age-income profile that have not been included in our and previous analysis. For instance, factors that determine a loss in earning capacity at retirement age of individuals, like deterioration of health and increasing aversion towards risk, could help in explaining the observed asymmetry.

Second, by estimating the whole distribution we are able to identify where in the distribution of income the various determinants exert their greatest impact. This detailed analysis can provide further insight into the determinants of inequality, of great importance to researchers as well as policy makers. For example, we find that the impact of employment status is spread over the

[1]The aim of CHER is to create an international comparative micro database containing longitudinal datasets from many national household panels and from the European Household panel study (ECHP). This will provide the basis to facilitate comparative cross-national and longitudinal research and to study processes and dynamics of policy issues related to family structures, educational aspects, labour force participation, income distribution, poverty, etc. Access to the (beta version of the) database has been granted while visiting the Integrated Research Infrastructure in the Socio-Economic Sciences (IRISS) at CEPS/INSTEAD.

entire income distribution. This finding seems to agree with results obtained by Nolan (1988-89) using 1977 Family Expenditure Surveys (FES) data in his analysis of the impact of UK economic conditions on income inequality. However, in addition, we find that the impact on income is substantially greater for lower income families.

Third, we devise a method for obtaining nonparametric conditional inequality measures by inverting the estimated conditional distribution. Our estimates indicate that, for instance, if the household size increases from 2 to 4, households in the top 90th percentile of the income distribution move from earning 3.2 times more then households in the 10th percentile to earning about 2.5 times more. This amounts to a 20 per cent fall in inequality. This increase in inequality is obtained controlling for other important determinant of inequality, such as the age structure, the presence of a retired head, and young children. Previous approaches, based on the "standardization" of inequality series, inequality decomposition by population sub-groups, or nonparametric methods, have not been to identify the contribution of individual factors on inequality, except for very simple cases.

Finally, our approach allows us to establish consistency and to estimate asymptotic variances of the proposed inequality estimators, which is useful for inference purposes. It provides a visually clear representation of both the substantive and statistical impact of each individual factor on income inequality, keeping all others constant. For instance, we find that for the UK sample, household size, number of young children, age of head, and employment status, have a large substantive and statistical impact on inequality. Factors such as years of education, marital status, and urban versus rural households, on the other hand, do not significantly impact inequality.

Chapter 5 reexamines the relationship between openness to trade and the environment, controlling for economic development, in order to identify the presence of multiple regimes in the cross-country pollution-economic relationship.

The data used in this Chapter consists of 2,294 observations representing 74 countries, 23 OECD and 51 non-OECD members, spanning the years 1960-1990. The dataset was constructed using data from various sources.

For the sulfur emissions, we took the data from the *Historical Global Sulfur Emissions* data set of A.S.L and Associates (1997), which includes the sulfur dioxide emissions from burning hard coal, brown coal, and petroleum, and sulfur emissions from mining and related activities for most of the countries of the world during the period 1850-1990 (Allen S. Lefohn 1999). The carbon dioxide emissions data come from the 1998 World Bank *World Development Indicators* CD-ROM. Most macroeconomic data is derived from the *Penn World Tables*(PWT) Mark 5.6 which compiles data for 152 countries on 29 subjects for the period 1950-1992. Foreign Direct Investment data are taken from the UN *World Trade Data Base* discussed in Feenstra, Lipsey, and Bowen (1997).

In this Chapter we first identify the presence of multiple regimes by using specification tests which entertain a single regime model as the null hypothesis. We then develop an easily interpretable measure, based on an original application of the Blinder-Oaxaca decomposition, of the quantitative impact on the environment due to differences in regimes.

We reject the linear model commonly used in the previous empirical literature in favor of a multiple regime alternative in which different countries obey different models when grouped according to income, trade policies, factor endowment, and other relevant variables. We also find that as much as 40 per cent of the pollution gap between developed and developing countries can be attributed to regime differences rather than economic activity. Applying a recursive partitioning method, we find that the impact of openness to foreign markets on sulfur and carbon dioxide emissions varies according to the level of development, trade policies, and the productive structure of the economy. Our result also show there is substantial geographic homogeneity within each regime, giving some support to findings by geographical factors (see, e.g., Neumayer, 2002). Our finding also highlight the importance of democracy (see, e.g., Torras & J.K., 1998; Harbaugh et al., 2002), corruption (see, e.g., Lopez & Mitra, 2000), and civil and political liberties (see, e.g., Barrett & Graddy, 2000; Torras & J.K., 1998). We find support for studies that based on the poor environmental performance of Soviet economies and dictatorships established in Latin America, Asia and Africa, have been

advocating democratic reforms as a way to promote both economic and environmental welfare (see, e.g., McCloskey, 1983; Payne, 1995). Income turning point estimates of the relationship between income and emissions agree with previous empirical studies on similar local impact pollutants. Only for the for the high-income countries the turning point is within the sample range at $16,000. For medium and low income countries, the turning point is either non-existent or the curve is monotone increasing over the sample range. For poorest countries the income variables are not statistically significant. For the poorer countries with low capital intensity, the turning point is outside the sample range, whereas for the countries with higher capital-per-worker, the curve is U-shaped with very low turning point so that the curve is monotone increasing over the sample range. Our results for sulfur emissions seem to give some support to the pollution haven hypothesis. The impact of openness to trade on pollution is almost 4 times higher than it is for rich countries then for poor countries. We find that turning points for $CO_2$ emissions tend to be higher than those for $SO_2$ emissions. For instance, The turning point for the rich country group was $9,679 *per capita*, whereas its $23,420 *per capita* for the high capital intensity high income) group for $CO_2$ emissions. A higher turning point for $CO_2$ is consistent with the environmental economics literature suggesting that inverted-U type relationships relationships are more likely to be found for certain types of environmental indicators, particularly those with a more short-term and local impact rather than those with a more global and long-term impacts (see, e.e, Arrow et al., 1995; Cole et al., 1997; Selden & Song, 1994). This finding also agrees with Dijkgraaf & Melenberg (2005) which finds that the inverted-U for $CO_2$ is likely to exist for several, but not all, countries. In particular, our findings could explain the sensitivity of their estimated emissions income relationships for $CO_2$, even with a relatively homogeneous sample of OECD countries.

Chapter 6 investigates the existence of the so called *environmental kuznets curve* (EKC) using nonparametric regression methods. The EKC empirical law features two variables of considerable interests to economists and policy makers, namely an indicator of environmental quality and the level of per capita income. The link between these variables takes the form of an

"inverted-U" shaped curve in the pollutant/income space. Several *ad hoc* explanations have been proposed to justify this empirical law. A simple and frequently used explanation for the EKC is that its inverted-U shape reflects changes in the demand for environmental quality as income increases. Assuming that environmental quality is a normal good, pollution will rise in the early stages of economic development, to decline later as income continues to rise. Several papers explain the Kuznets curve by using models with threshold effects in either pollution abatement, (see, e.g., Jones and Manuelli, 1995), or environmental policy regulation (see, e.g., Stokey, 1988). Threshold effects lead to a very different relationship between environmental quality and income during early stages of economic development as opposed to later stages. The threshold-effect predicts a period of long inactivity in private sector responses to ever tightening pollution policy: the income-effect theory predicts that the abatement intensity rises continuously as policy tightens.

Using nonparametric regression methods we have also estimated the nonparametric elasticity with respect of per capita income. The flexible nature of nonparametric estimation allows as to find evidence of an asymmetric behaviour of the curve before and after the turning point, consistent with threshold-effect models. This finding is also consistent with the empirical evidence found by Vincent (1997) and Carson (1997) concerning the existence of a Kuznets curve within individual countries as summarised by Panayotou (2000). We test the nonparametric findings using Hansen's (2000) threshold model. Threshold models can be viewed as parsimonious strategies for nonparametric estimation. Our estimates suggest that there might be a sample split based on per capita income. No evidence of a split based on trade variables was found. The income turning point of the global sample is much lower than the threshold income that divides the two regimes. Changes that might benefit the environment occur at much higher levels of income than those implied by standard EKC models. The turning point of the global sample is much lower than the threshold income that divides the two regimes. We find that the impact of income on pollution is greater in regime of richer countries than in the poorer regime. This is consistent with the nonparametric findings. Moreover, we find that regime differences are also

apparent from the estimated error variance. The estimated error variance of the poorer countries regime is more than twice that of the richer countries regime. This result supports claims made previously in the literature. For instance, Panayotou (2000) after examining the evidence form Vincent (1997) and Carson et al. (1997a) concerning the existence of a Kuznets curve within individual countries concludes that: "whereby rising incomes result in a more effective regulatory structure by changing public preferences and making resources available to regulatory agencies. States with low-income levels have a far greater variability in emissions per capita than high-income states suggesting more divergent development paths. This has the implication that it may be more difficult to predict emission levels for low-income countries approaching the turning point." We also verify this hypothesis with a formal test.

Finally Part III presents a short summary (Chapter 7), some direction for further research (Chapter 8), and concludes.

# Part I

# Background

Chapter 2

# Basic Nonparametric Methods with Economic Applications

## 2.1 Introduction

This chapter provides an introduction to the basic nonparametric methods underlying the applications in the subsequent applied chapters.

First, we briefly introduce the distinction between parametric and nonparametric models. Then we introduce the basic univariate and multivariate nonparametric kernel density estimators, the fundamental nonparametric building blocks of subsequent applications. Several original applications are provided to illustrate to illustrate the practical relevance of nonparametric methods in economics.[1]

This chapter is based mostly on class notes for courses attended in the 90's in the Virginia Polytechnic Institute and state University in Blacksbourg, Virginia, USA. More specific references will be provided for selected topics.

This material should serve as a brief introduction to Chapter 3 on reporting nonparametric computational-based results. Also, several later chapters will make use of the estimators presented here.

The univariate kernel density estimator, besides serving as the building block for the multivariate kernel and the conditional kernel estimator presented in Section 2.7, was also used in Chapter 3, Section 3.2, to produce the estimate of the household income density in the two panels of Figure 3.10 on page 89, Figure 3.2 on page 61, the two panels in Figure 3.3 on page 62, Figure 3.4 on page 63, and Figure 3.5 on page 67 in Section 3.4. In Chapter 4 on income inequality, a univariate gaussian kernel was used to produce the income density estimate shown Panel 4.1(c) in Figure 4.1 on page 102 in Section 6.10. In Chapter 6, the univariate kernel density estimator features in Section 6.3 on page 6.3 to derive the nonparametric kernel regression estimator of Nadaraya (1964) and Watson (1964) and the local linear regression estimator. It also used, for example, in Section 6.7 to produce the nonparametric regression estimates of the environmental Kuznets curve in Figure 6.8 on page 184 and Figure 6.9 on page 185.

[1]For instance, Example 2 and 4. Example 4 introduces an original methodology to estimate a conditional density with an application to labor economics. Example 3 was contributed to the book by Li & Racine (2006), Nonparametric Econometrics: Theory and Practice, Princeton University Press.

The bivariate gaussian product kernel is used, for instance, in Chapter 4 to produce the estimate of the joint density of household income and age of head, displayed in panel 4.1(d) of Figure 4.1 on page 102 in Section 6.10.

Together with the univariate kernel, the bivariate kernel was used in Chapter 4 to estimate the density of household income conditional on age of the head in Panel 4.2(a) of Figure 4.2, and the density of household income conditional on household size in Panel 4.3(b) of Figure 4.3 on page 103 in Section 6.10.

In Section 2.2 we review the basic definitions of parametric families and models against which the nonparametric equivalent are later contrasted and defined. Section 2.3 presents the basic definitions of nonparametric and semi-parametric models. In Section 2.4 we present the advantages and disadvantages of using nonparametric methods in economics. Examples are used to illustrate the usefulness of the nonparametric methods. Some problems with nonparametric methods are presented in Section 2.5. In section 2.6 the ideas behind the construction of the univariate kernel density estimator of a density function are introduced. In Section 2.7 few of the main issues associated with multivariate kernel density estimation are addressed. Section 2.8 concludes.

## 2.2 Populations, Samples, and Parametric Models

In statistical inference, a data set is viewed as a *realization* or *observation* of a random element defined on a probability space $(\Omega, \mathscr{F}, P)$ related to a random experiment. The probability measure $P$ is called the *population.*

As the population $P$ is unknown, to simplify the analysis, a set of assumptions on it are usually made in the form of a statistical model.

**Definition 1 (Parametric family and model)** Given a measurable space, $(\Omega, \mathscr{F})$, a set of probability measures defined on that space, $\mathscr{P}$, the triplet $(\Omega, \mathscr{F}, \mathscr{P})$ is known as a *statistical model.*

If a set of probability measure $\mathscr{P}_\theta$, indexed by a parameter $\theta \in \Theta$, is said to be a *parametric family* iff $\Theta \subseteq \mathbb{R}^d$ for some fixed positive integer $d$, and

each $P_\theta$ is a known probability measure when $\theta$ is known. The index set $\Theta$ is referred to as the *parameter space* and $d$ is called its *dimension.* □

**Example 1** Consider estimating a density function, $f$. The parametric methods specify the form of $f(x;\boldsymbol{\theta})$. If we assume that $f(x;\boldsymbol{\theta})$ is the normal density, with $\boldsymbol{\theta} = (\mu, \sigma^2)^T$, the parametric normal family is then

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

where the mean $\mu$ and the variance $\sigma^2$ are the parameters of $f$. The problem of completely describe the distribution function is reduced to the problem of estimating $\boldsymbol{\theta} = (\mu, \sigma^2)^T$. A parametric estimator of $f$ is then

$$\hat{f}(x) = \frac{1}{\widehat{\sigma}\sqrt{2\pi}} e^{-(x-\widehat{\mu})^2/2\widehat{\sigma}^2},$$

where $\mu$ and $\sigma$ are estimated from a sample using the well known sample mean and sample variance formulae, respectively

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

and

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.$$

Using regional Italian data on capita GDP (constant prices 1990) for the period 1951-1998 provided by ISTAT (Istituto di Statistica Nazionale)[2], Figure 2.1 presents three views, 1955, 1975, and 1995, of the evolution of real GDP per capita (millions of 1990 Lire). The sample mean and standard deviations for the estimates are: $\widehat{\mu}_{1955} = 6.6827$, $\widehat{\sigma}_{1955} = 2.237202$, $\widehat{\mu}_{1975} = 12.3468$, $\widehat{\sigma}_{1975} = 3.288181$, $\widehat{\mu}_{1995} = 18.4447$, and $\widehat{\sigma}_{1955} = 4.447884$. □

[2]Except form the 1951-1963 period (CRENoS) and the 1996-1998 period (SVIMEZ).

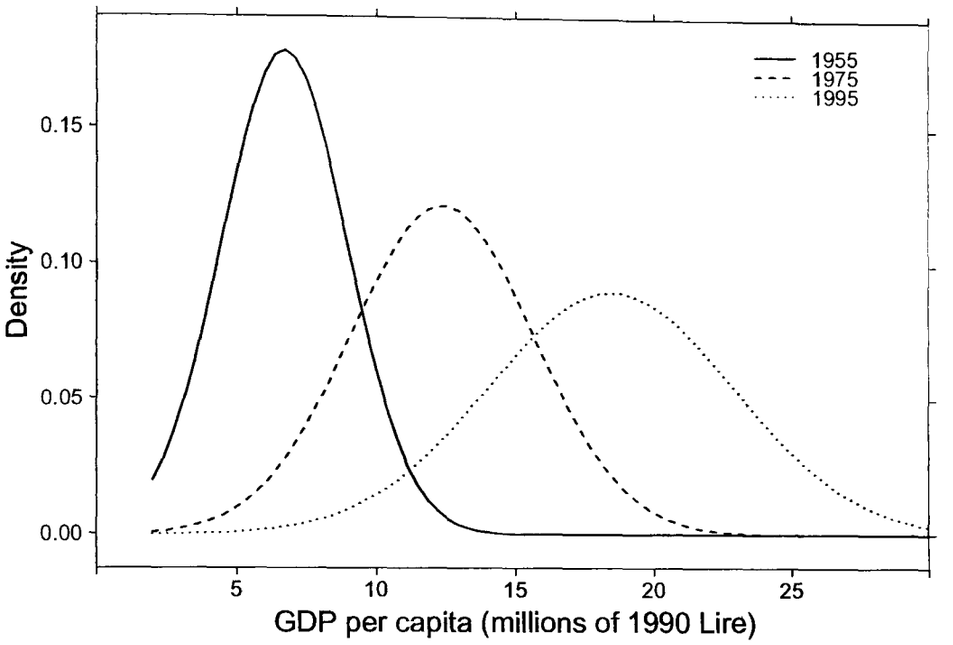



Figure 2.1: Evolution of Italian GDP per capita, 1951-1988

## 2.3 Nonparametric and Semiparametric Models

According to David (1995), the term nonparametric applied to estimation and statistical inference has been first used in (Wolfowitz, 1942, p. 264): "We shall refer to this situation [where the knowledge of the parameters, finite in number, would completely determine the distributions involved] as the parametric case, and denote the opposite case, where the functional forms of the distributions are unknown, as the non-parametric case."

The term "nonparametric" has acquired over time two separate more narrow meanings, one roughly in the statistical literature and another in the econometric literature. The older use of the term refers to tests, such as the Kolmogorov-Smirnov test, Wald-Wolfowitz runs test, Mann-Whitney U test, etc., (see, e.g, Hollander & Wolfe, 1973) that do not assume normality, and that are often based on rank transformed data. More recently, the term has been used to refer to "smoothing techniques," as in Simonoff (1999). We will take the more modern use of "nonparametric" to refer mostly to

density estimation and regression smoothing. (Scott, 1992, p. 44) provides an interesting discussion on when is an estimator nonparametric.

Excellent general surveys of nonparametric methods written for statisticians include Simonoff (1999) and Loader (1999). Other excellent surveys focusing on kernel and local regression methods include Bowman & Azzalini (1997), Wand & Jones (1995) and Fan & Gijbels (1996). Survey dealing with some issues of key interest in econometrics include Pagan & Ullah (1999), Yatchew (1999), and Li & Racine (2006).

For the purpose of this thesis we need the following definitions.

**Definition 2 (Nonparametric family and model)** A family of probability measures is said to be *nonparametric* if it is not parametric according to Definition 1

A nonparametric model refers to the assumption that the population $P$ is a nonparametric family. □

**Remark 1** Nonparametric families are probability measures indexed by an infinite-dimensional parameter set. Another name is "families with large parameter space." □

**Remark 2** In principle nonparametric families are not restricted by any assumption. In most applications though, assumption on the support of the distributions, on the existence of moments, on the shape of the distributions, and on the smoothness of the distributions, are made. □

**Definition 3 (Semiparametric family and model)** Semiparametric families are usually characterized by two components, a component with a finite dimensional parameter set and a component with an infinite-dimensional parameter set, i.e., a function.

A semiparametric model refers to the assumption that the population $P$ is a semiparametric family. □

## 2.4 Nonparametric Vs. Parametric Models

We argue that nonparametric and semiparametric methods can provide information of considerable value to economists. This information would be

difficult to detect using parametric models. Particular features appearing in the data can be fitted only through *ad hoc* assumptions with parametric models. A few examples shall illustrate these points.

The following two example apply nonparametric regression to economic problems.

**Example 2 (Environmental Kuznets Curve)** There exists an extensive parametric literature in environmental economics, where an indicator of environmental quality is generally modeled as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i$$

where $y_i$ is the natural logarithm of pollution emissions *per capita*, $x_i$ is the natural log of income *per capita*, and $u_i$ the standard error term. Sometimes a cubic term is added to the basic regression equation. Researchers are interested in determining whether an inverted-U relationship between environmental quality and economic growth. For this purpose estimating a polynomial function appears adequate. However, since polynomial functions possess all orders of derivatives everywhere, this property might smooth out important features that are present in the data, such as an asymmetric behavior around the turning points. For example, in the estimation of the relationship between *per capita* GDP and an environmental indicator, researchers might be interested not only in determining the existence and location of turning points but also whether the behavior of an up swing following a down swing is symmetric. Asymmetric behavior around a turning point, besides having important consequences for the policy maker as such, might also indicate the presence of different factors affecting the downward and the upward branch of the curve. Figure 2.2 suggests that such an asymmetric behavior is supported by the data. Stern and Common (2000) have pointed out that trade might play an important role in explaining the downward part of the EKC for developed countries. Asymmetries might also indicate the presence of irreversibilities. □

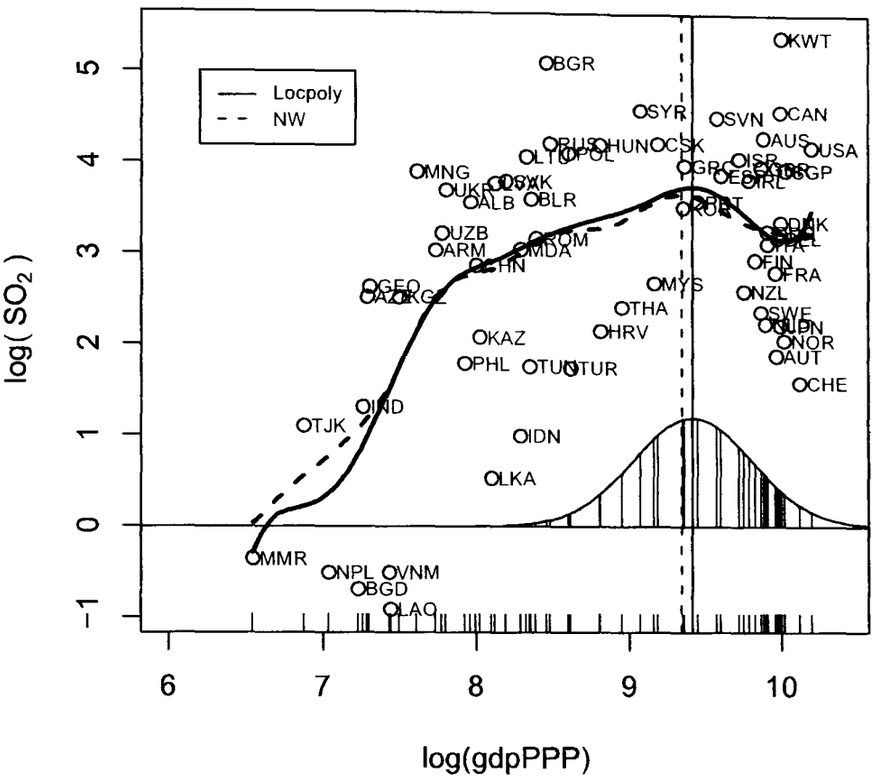



Figure 2.2: Local Polynomial and Nadaraya-Watson estimate for $SO_2$.

As another example, Pagan & Ullah (1999) consider the relationship between the natural income and age, using data on a sample of 205 Canadian workers from a 1971 Canadian Census Public Tapes (Ullah, 1985). The standard approach in labor economics is to assume a quadratic relationship in age, estimated by OLS. The nonparametric approach makes no assumptions about the functional form of the relationship. The nonparametric specification finds a flatter peak than the quadratic curve and indicates the presence of a "dip" around the mean age of 40. Pagan & Ullah (1999) argue that a possible explanation lies in the generations effect. The dip is produced by the overlap of earning trajectories of different generations. They conclude that "only if the sociopolitical environment of the economy has remained stable intergenerationally can we assume these trajectories to be the same" (Pagan & Ullah, 1999, p. 154). This result is robust to bandwidth choice, and is

observed whether using simple rules of thumb or data-driven methods such as likelihood cross-validation. Figure 2.3 shows the stacked density estimates from 1951 to 1988.[3]

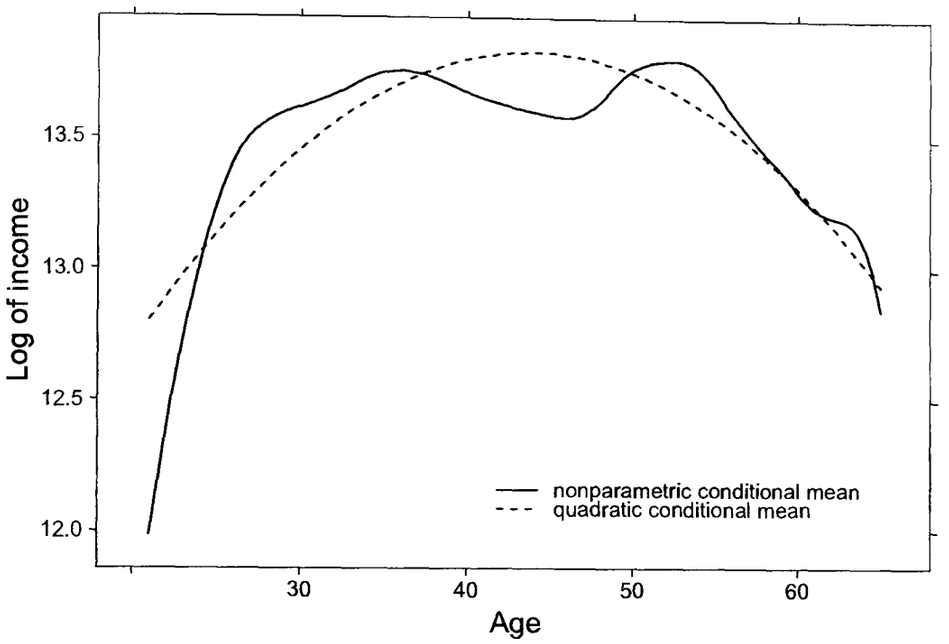



Figure 2.3: Nonparametric and quadratic fit, income/age profile (Canadian workers data)

Though it is true in both cases that parametric specifications could be used to fit these complex models, parametric model would find the detection of these features problematic. The parametric specification (say through mixtures, dummies, etc.) would require *ad-hoc* assumptions.

Moreover semiparametric models and estimation methods, where unknowns are a finite dimensional set of parameters and functions, retain the flexibility of nonparametric methods, whilst, mitigating most of the problems with nonparametric methods.

[3]The figure was obtained using univariate gaussian kernel evaluated on 100 equally spaced points in the interval [21,65] with bandwidth selected using the plug-in method for local linear regression described in Ruppert et al. (1995b) as implemented in the *spill* function provided by R's `sm` library by Bowman & Azzalini (1997). This result is robust to the choice of kernel and bandwidth selection method.

The following example illustrate the use of nonparametric density estimation to shed light on important economic problems and can reveal features not identifiable by parametric means.

**Example 3 (Italian income distribution evolution)** Using the data described in Example 1, Figure 2.4 shows the stacked density estimates from 1951 to 1988[4].

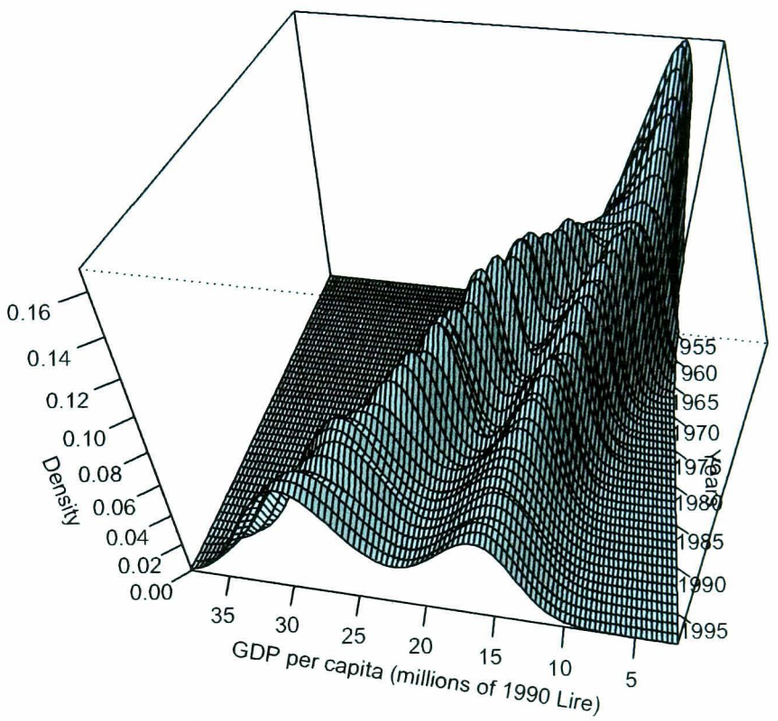



Figure 2.4: Evolution of Italian GDP per capita, 1951-1988

[4]The figure is composed of 48 stacked kernel density estimates using univariate gaussian kernel evaluated on 100 equally spaced points in the interval [2,38] with bandwidth selected using the plug-in method described in Sheather & Jones (1991) as implemented in the *sm* R library by Bowman & Azzalini (1997). This result is robust to the choice of kernel and bandwidth selection method.

It is clear from the Figure, that the Italian distribution of per capita GDP displays an interesting dynamics: it starts as a unimodal distribution in the 50s and becomes bimodal in the 60s. The two mode tend to diverge during the 90-98 period. This is also illustrated in Figure 2.5 □

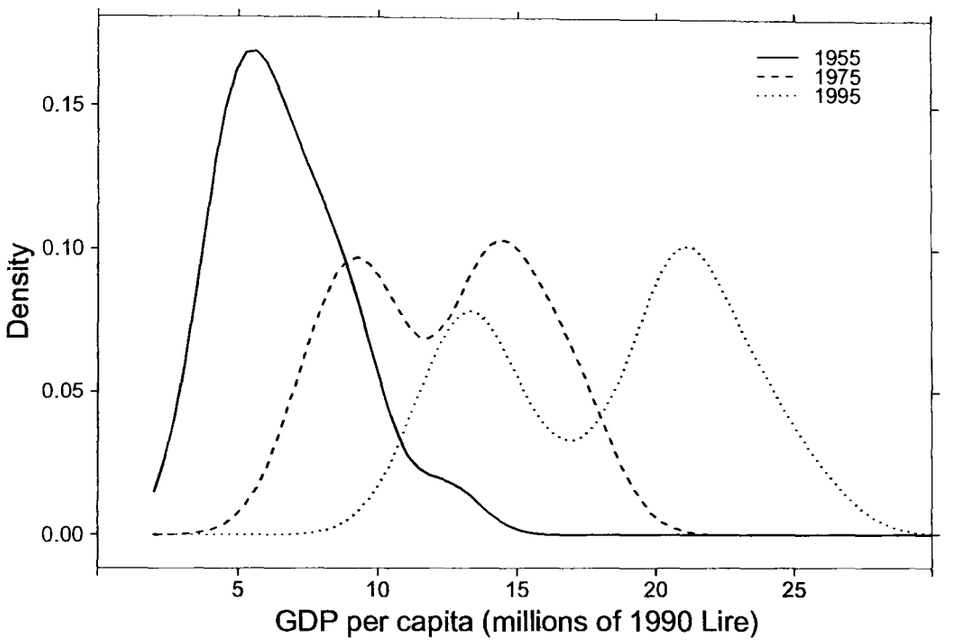



Figure 2.5: Evolution of Italian GDP per capita, 1951-1988

## 2.5 Limits of Nonparametric Models

### 2.5.1 Curse of Dimensionality

Curse of dimensionality (Bellman, 1961, see) refers to the exponential growth of hypervolume as a function of dimensionality. An example in Hastie & Tibshirani (1990, p. 84) clearly illustrates the problem.

Consider two hypercubes cubes with identical orientation, both centered on the origin of a cartesian coordinate system. Suppose that one cube has sides of length $\ell$ and the other has slightly smaller sides of length $\ell - \epsilon$.

The volume of the $d$-dimensional hypercube of side length $\ell$ is given by the formula

$$V_d = \ell^d.$$

Consider the fraction of the volume of the larger cube in between the cubes. Then

$$\lim_{d\to\infty} \frac{V_d(\ell) - V_d(l-\epsilon)}{V_d(\ell)} = \lim_{d\to\infty} \frac{\ell^d - (\ell-\epsilon)^d}{\ell^d} = \lim_{d\to\infty} 1 - \left(1 - \frac{\epsilon}{\ell}\right)^d = 1$$

Hence, the content of an hypercube tends to concentrate toward its surface, a $d-1$-dimensional subspace, as the number of dimensions increase. This conversely, implies that the center becomes less and less important, as the dimension increases. This "space distortion" has potentially serious practical consequences for data analysis. For example, in parametric linear regression, the fact that the data tends to concentrate in a lower dimensional space, renders the method prone to the problem of multicollinearity.

In nonparametric estimation this problem limits the applicability of the technique low-dimensional cases only. Most nonparametric methods employ the concept of local neighborhood to compute estimates. For example in nonparametric regression analysis, to calculate a conditional mean at a particular point, only the $k$-nearest points are included in the averaging (hard neighborhood) or the data are weighted according to their distance from the conditioning value (soft neighborhood).

Consider constructing a cube-shaped neighborhood of a point, say the origin, that should include all, $p \cdot 100$ per cent, of the data, assumed to be uniformly distributed within a unit hypercube. The cubic neighborhood should have side length $\ell = p^{1/d}$. This signifies that to include 10 per cent of the data, i.e $p = 0.1$, when $d = 1$, the length of the side of the cube-shaped neighborhood should be $\ell = 0.1$. With $d = 10$, $\ell \approx 0.8$. This example illustrates the idea of "local," in high dimension cannot be readily understood using intuition developed within much simpler low-dimensional geometry. Figure 2.6 represents the side length of the hypercube needed to capture a pre-specified proportion of the data for dimensions $d = 1, 2, 3, 10, 20$, from

the bottom upwards, respectively.

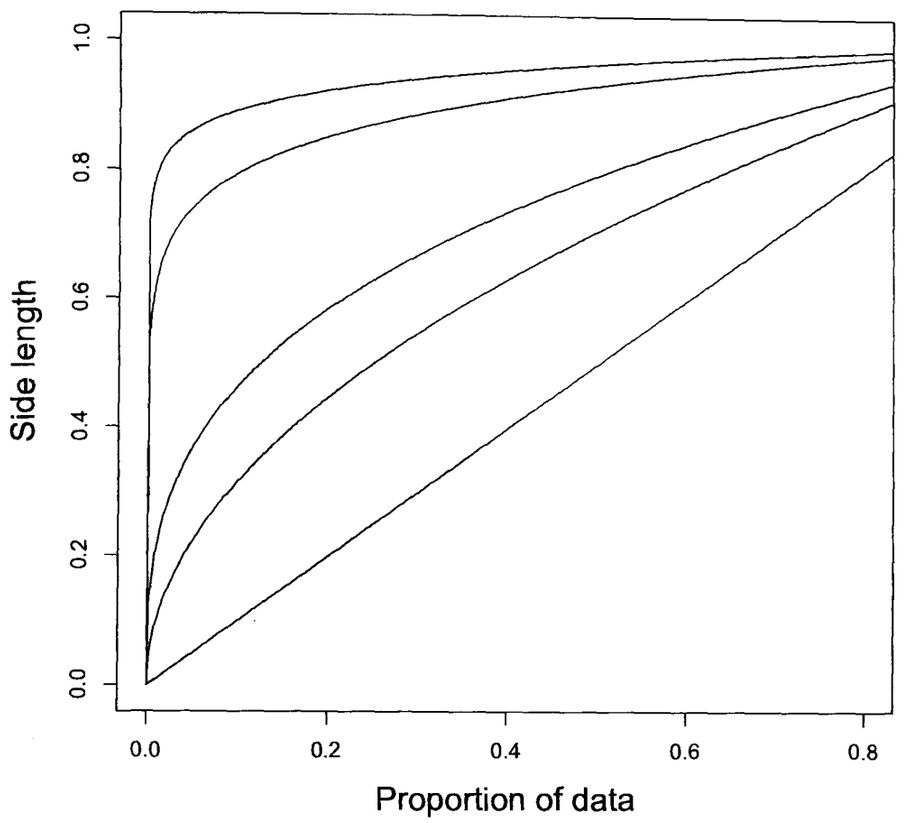



Figure 2.6: Curse of dimensionality illustration

The curse of dimensionality problem affects also density estimation methods. Consider the case of density estimation. Conceptually, estimating a density function nonparametrically appears to be simple. The most basic nonparametric method of density estimation is the histogram. Because of the curse of dimensionality problem, as the dimension of the data increases, the complexity of estimating a density via an histogram increases exponentially, the number of histogram grid cells increases exponentially as the dimensions

increase. This effect cannot be avoided, even by other, more complicated, nonparametric estimation methods.

### 2.5.2 Interpretability

A problem with nonparametric methods is the difficulty in presenting and interpreting results in a multivariate setting. As the number of dimensions increases, only a lower dimensional projection can be displayed and interpreted. Several graphical devices may be needed to display and to highlight important features in the estimates. Moreover, the curse of dimensionality problem makes interpreting multi-dimensional problems difficult, as intuition acquired in low-dimensional geometry can be of no help when we move beyond the three dimensions.

### 2.5.3 Forecasting

A further problem with nonparametric methods is that they do not readily permit extrapolation. In the case of $E[Y|x]$, it does not provide predictions at points $x$ that are not in the support of $X$. This could be a serious problem when analyzing policies and making forecasts, whose main purpose is to make statements about what could happen under conditions that do not exist under the data available. A parametric model, in which $E[Y|x]$ is known up to a finite-dimensional parameter, provides predictions a all values of $x$.

## 2.6 Univariate Kernel Density Estimation

Much work has been done on the problem of density estimation. One of the most popular methods is that of kernel smoothing. We refer to Watson (1964), Nadaraya (1964), Silverman (1986), Wand & Jones (1995), and Simonoff (1999) and the references given therein.

If we consider the definition of $f(x)$:

$$f(x) \equiv \frac{dF(x)}{dx} \equiv \lim_{h \to 0} \frac{F(x + \frac{h}{2}) - F(x - \frac{h}{2})}{h}$$

If we replace $F(x)$ with the empirical CDF

$$\widehat{F}_n(x) \equiv \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(x_i)$$

we get

$$\widehat{f}(x) = \frac{\widehat{F}(x+\frac{h}{2}) - \widehat{F}(x-\frac{h}{2})}{h} = \frac{1}{nh}\sum_{i=1}^{n} 1_{(x-h/2,x+h/2]}(x_i)$$

which can be rewritten as

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

where

$$K(t) = 1_{(-\frac{1}{2},\frac{1}{2}]}(t).$$

The Kernel density estimator is the central finite-difference approximation to the derivative of the ECDF.

The problem is this estimator is not smooth. If We choose $K$ to be the standard normal we obtain the classical density estimator.

Figure 2.7 shows the components of a kernel density estimate based on a Normal kernel. The four data points are marked by crosses on the horizontal axis. The data are represented by. Centered at each data point are the broken curves represent the normal components, namely, $\frac{1}{nh}\mathrm{K}\left(\frac{x-X_i}{h}\right)$ (i.e., 1/n times a normal density with mean $X_i$ and standard deviation $h$). The solid curve represents the kernel density estimate.

Figures 2.8 and 2.9 illustrate the impact of the choice of bandwidth on the shape of the estimated kernel. in Section 2.7 starting on on page 35. A

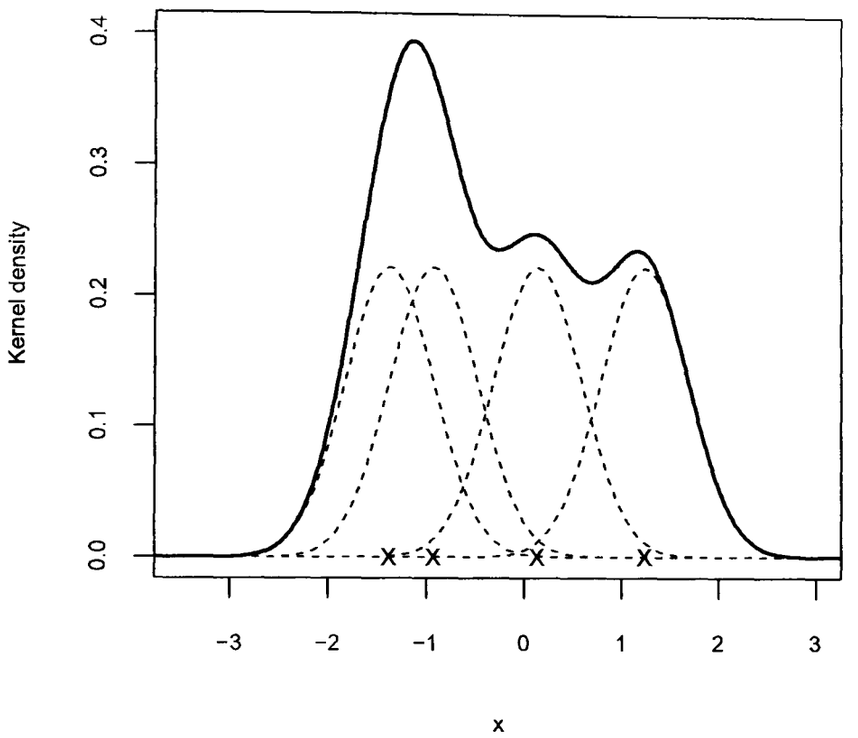



Figure 2.7: Construction of kernel density estimate

Java applet we have developed,[5] that allows the user to watch the effects of changing the bandwidth and the shape of the kernel function on the resulting density estimate, was cited in a survey of density estimation by (Sheather, 2004, p. 589).

Next we introduce an original methodology to estimate a conditional density with an application to labor economics, that makes use of kernel density estimation.

[5]The applet can be found at http://www-users.york.ac.uk/~jb35/mygr2.htm.

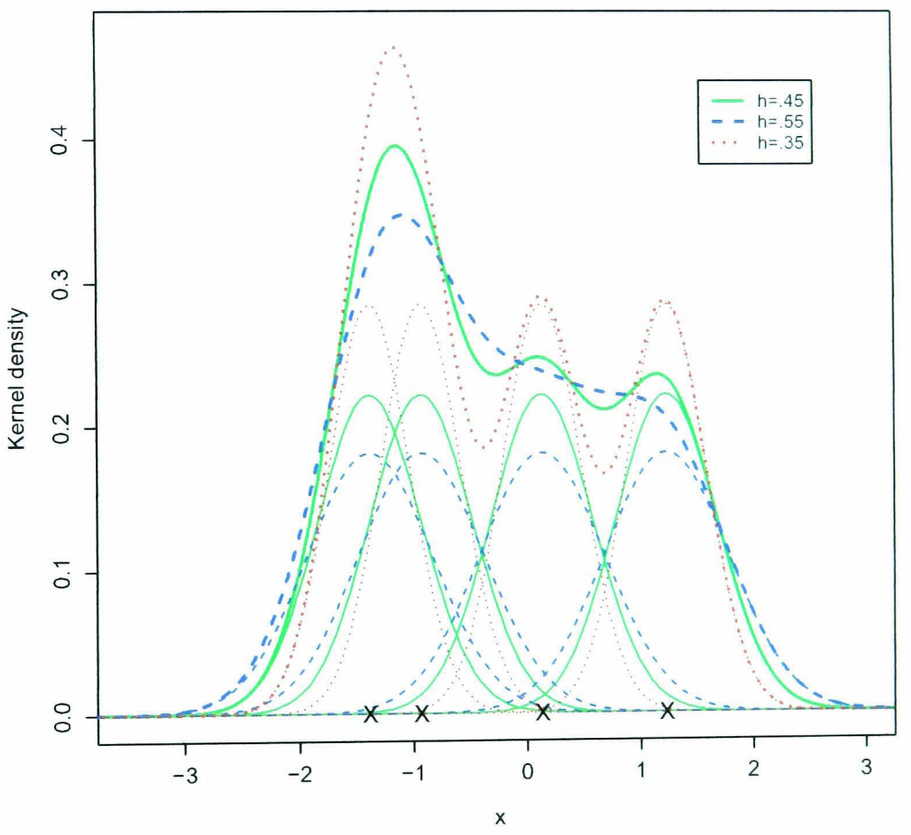



Figure 2.8: Influence of the window width

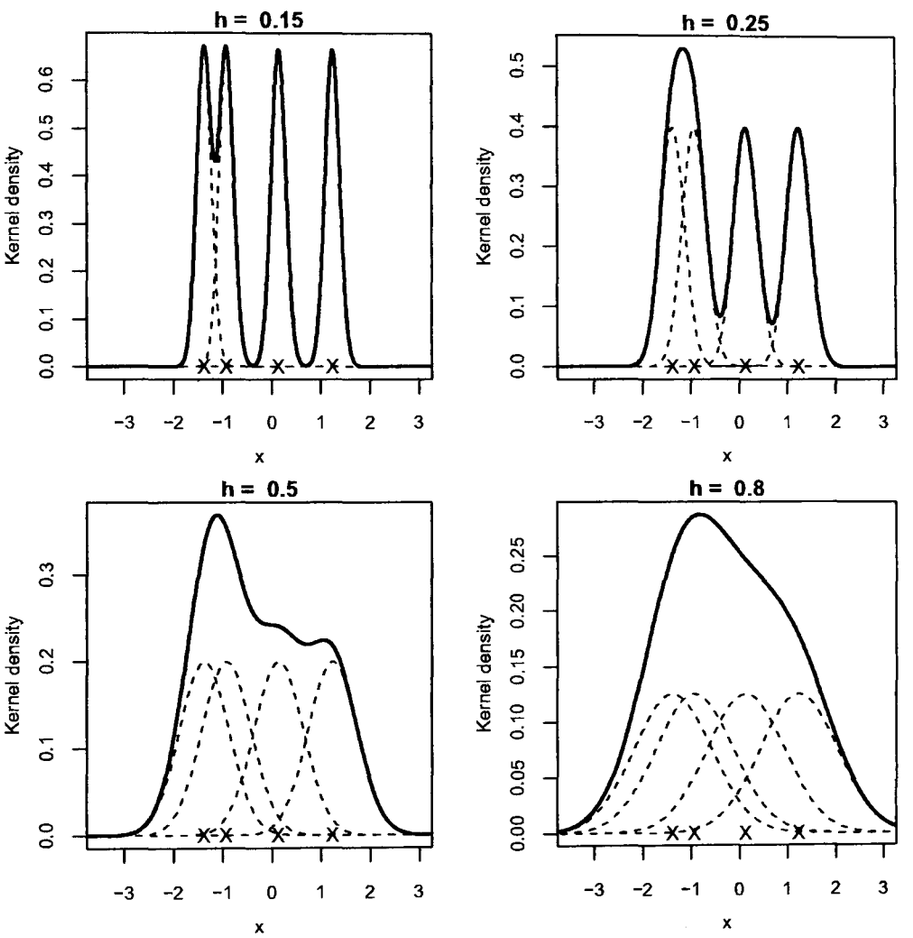



Figure 2.9: Influence of the window width

**Example 4 (Impact of Unionization)** Consider the case of density estimation. We want to determine the impact of unionization on the distribution of wages conditional on many determinants. The data are assumed to be a realisation of a strictly stationary stochastic process $\left\{\left(\vec{X}_i, Y_i\right)\right\}_{i=0}^{\infty}$, where $Y_i$ is a scalar and $\vec{X}_i$ is a $d$-dimensional vector, usually of individual attributes. This general framework includes the particular case were the pairs $\left(\vec{X}_i, Y_i\right)$ are independent and identically distributed. Let $f(y|\vec{x})$ be the conditional density of $Y_i$ given $\vec{X}_i = \vec{x}$, which we assume to be smooth in both $\vec{x}$ and $y$. We are interested in estimating $f(y|\vec{x})$ from the data $\left\{\left(\vec{X}_i, Y_i\right)\right\}_{i=0}^{\infty}$. The kernel density estimator for ordinary data can be written as the convolution product

$$\hat{f}_Y(y) = \left(\hat{F}_Y * K_h\right)(y) \equiv \int_{-\infty}^{\infty} K_h(y-u)\, d\hat{F}_Y(u) \tag{2.1}$$

where the integral is a Stieltjes integral, $\hat{F}_Y$ is an estimate of the cumulative distribution function of $Y$, and $K_h(u) = h^{-1}K(u/h)$. The kernel function $K$ will be taken to be the Gaussian distribution throughout the paper. The smoothing parameter $h$ will be taken to be the asymptotically optimal for estimating a density function when the underlying distribution is Normal. Equation 2.2 uses the ideas of convolving a kernel with the density estimate induced by an estimate of the cumulative distribution function. When $\hat{F}_Y$ is the empirical cumulative distribution function, $\hat{F}_n(x) \equiv n^{-1}\sum_{i=1}^{n} 1[Y_i \leq x]$, Equation 2.2 can be rewritten as

$$\hat{f}_Y(y) = n^{-1}\sum_{i=1}^{n} K_h(y - Y_i) \tag{2.2}$$

which is the usual way to represent the kernel density estimator. By analogy, the kernel estimator of $\hat{f}_{Y|X}(y|x)$ induced by the conditional distribution function $\hat{F}_{Y|X}$ is then

$$\hat{f}_{Y|\vec{X}}(y|\vec{x}) = \int_{-\infty}^{\infty} K_h(y-u)\, d\hat{F}_{Y|\vec{X}}(u) \tag{2.3}$$

where $\hat{F}_{Y|X}$ is an appropriate estimator for the conditional distribution function of $Y$ given $\vec{X} = \vec{x}$. A similar approach has been followed to estimate a *hazard function* by using an "empirical cumulative hazard function" and densities with right-censored data by using *Kaplan-Meier*'s generalization of the ECDF (see Wand and Jones, 1995. Equation 2.3 can be rewritten as

$$\hat{f}_{Y|\vec{X}}(y|\vec{x}) = \sum_{i=1}^{n} w_i K_h(y - Y_i)$$

where $w_i$ is the size of the jump of $\hat{F}_{Y|\vec{X}}$ at $Y_i$. Once an estimate for the weights $w$ is obtained the conditional density can be estimated by weighted kernel methods.

To obtain the weights $w$ we need from an estimate of the conditional distribution function. The following paragraph describes the simple semiparametric approach used by Foresi and Peracchi (1995) for estimating $F_{Y|X}(y \mid \vec{x})$.

In general, if we define a new random variable using the indicator function $Z_i = 1[Y_i \leq y]$, then $E[Z_i \mid \vec{X}_i = \vec{x}] = F_{Y|X}(y| \vec{x})$. In order to estimate $F_{Y|X}(y \mid \vec{x})$ we propose to use the simple semiparametric approach used by Foresi & Peracchi (1995). A summary of other analogous nonparametric methods that could be employed is provided by Hyndeman et al. (1996) and by Hall et al. (1999).

The simplest approach is to fit a logistic binary regression model to $Z_i$. By estimating $J$ distinct functions $P_1(\vec{x}), \ldots, P_J(\vec{x})$ where $P_j(\vec{x}) = F(y_j|\vec{x})$ and $-\infty < y_j \cdots < y_J < \infty$ are distinct points in the support of $Y_i$. By fitting $J$ distinct logistic binary regressions to each binary variable $Z_{j,i} = I_{(-\infty,y_j]}(Y_i)$, $j = 1, \ldots, J$, where $I_A(\cdot)$ denotes the indicator function of the event $A$, we can approximate the cumulative distribution, $F(y|\vec{x})$. The logit model, besides being simple to implement and available in most econometric packages, also ensures that the estimated functions are bounded between 0 and 1. However, this method does not guarantee the monotonicity property of the conditional distribution function.[6]

To illustrate the effectiveness of the new approach we are going to apply

[6]For more details, see Foresi and Peracchi, 1995.

it to the wage dataset from Johnston and DiNardo (1997). Figure 2.10 shows the conditional empirical CDF and figure 2.11 plots the conditional distributions for union and non union workers . The estimates suggest that for men unions have an equalizing effect. The density center is shifted to the right when union=1. The density for union=0 has less weight at its center and more on its lower half. Lower wage workers are the ones that benefit the most from unionization. There is a suggestion that at relatively high wages union have a negative impact. □

Figure 2.10: Conditional CDF of log(wages)

Figure 2.11: Conditional Density of log(wages)

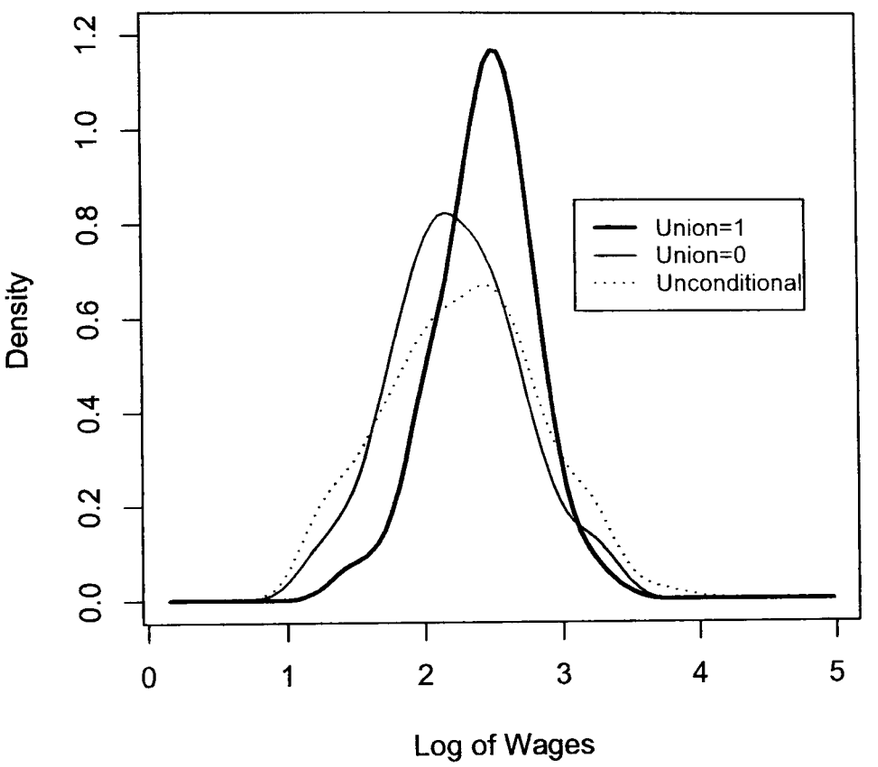


Figure 2.12: Conditional Density

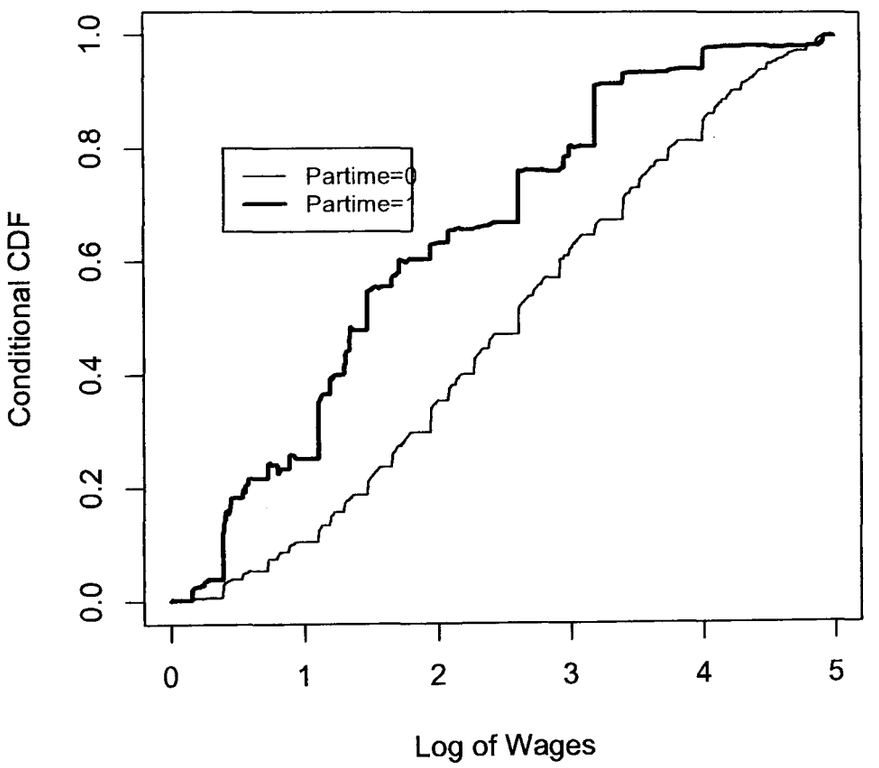
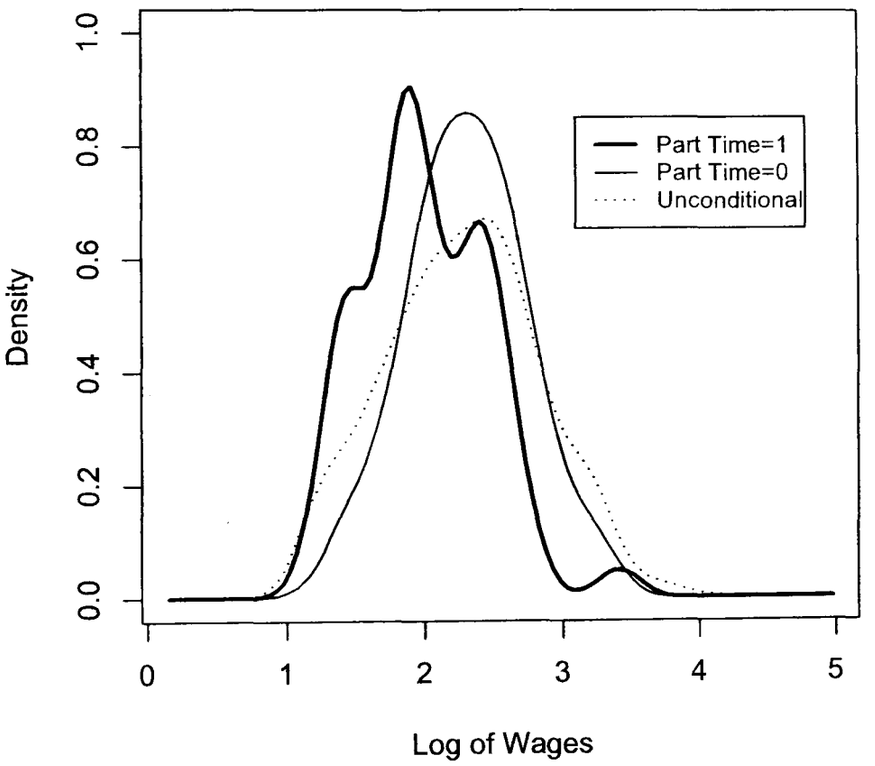


Figure 2.13: Conditional Density



# 2.7 Multivariate Kernel Density Estimation

## 2.7.1 Introduction

In this section we will investigate how kernel density estimation can be extended to include multivariate settings. Multivariate kernel density estimation is a prerequisite for conditional density estimation. A comprehensive treatment of the argument can be found in the monographs by Wand & Jones (1995), Scott (1992), and Fan & Gijbels (1996).

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a $\mathbb{R}^d$-valued random sample from an unknown $F$ with Lebesgue density $f$. The most general kernel density estimator of $f$ is given by

$$\widehat{f}(\boldsymbol{x}; \boldsymbol{H}) = \frac{1}{n} \sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{X}_i) \tag{2.4}$$

where $\boldsymbol{H} = \{h_{ij}\}$ is a $d \times d$ positive definite matrix of bandwidths,

$$K_{\boldsymbol{H}}(\boldsymbol{x}) = |\boldsymbol{H}|^{-1/2} K(\boldsymbol{H}^{-1/2}\boldsymbol{x}),$$

and $K$ is a $d$-variate kernel function satisfying the condition

$$\int K(\boldsymbol{x})\, d\boldsymbol{x} = 1.$$

Typically $K$ is taken to be a $d$-variate density function. Using the standard $d$-variate gaussian kernel function

$$\Phi(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\boldsymbol{x}^T\boldsymbol{x}}{2}\right) \tag{2.5}$$

then $K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{X}_i)$ becomes the joint probability density of the multivariate-normal vector random variable $\boldsymbol{x}$ with mean vector $\boldsymbol{X}_i$ and positive-definite variance-covariance matrix $\boldsymbol{H}$, $N(\boldsymbol{x}, \boldsymbol{X}_i)$,[7] in which case (2.4) becomes

[7]

$$K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{X}_i) = |\boldsymbol{H}|^{-1/2} K\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x} - \boldsymbol{X}_i)\right) \tag{2.6}$$

$$\widehat{f}(\boldsymbol{x};\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2}|\boldsymbol{H}|^{1/2}} \cdot \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{X}_i)^T \boldsymbol{H}^{-1}(\boldsymbol{x}-\boldsymbol{X}_i)}{2}\right) \quad (2.10)$$

Another popular $d$-variate kernel is a generalization of the univariate Epanechnikov kernel

$$K(\boldsymbol{x}) = \frac{(d+2)\Gamma(d/2+1)}{(2\pi)^{d/2}}\left(1-\boldsymbol{x}^T\boldsymbol{x}\right) I_{(\boldsymbol{x}^T\boldsymbol{x}\leq 1)} \quad (2.11)$$

### 2.7.2 Smoothing Parametrisation Selection

In general $\boldsymbol{W}$ belongs to the class of positive definite (and therefore symmetric) matrices

$$\mathscr{W}_p = \left\{ \boldsymbol{W} = \begin{pmatrix} \boldsymbol{w}_1^2 & \boldsymbol{w}_{12} & \cdots & \boldsymbol{w}_{1d} \\ \boldsymbol{w}_{12} & \boldsymbol{w}_2^2 & \cdots & \boldsymbol{w}_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{w}_{d1} & \boldsymbol{w}_{d2} & \cdots & \boldsymbol{w}_d^{-1} \end{pmatrix} : \forall \text{ nonzero } \boldsymbol{x}, \boldsymbol{x}^T\boldsymbol{W}\boldsymbol{x} > 0 \right\}.$$

If $\boldsymbol{W} \in \mathscr{W}_p$, then it has $\frac{1}{2}d(d+1)$ distinct smoothing parameters. The number of parameters to be chosen or estimated can be drastically reduced if $\boldsymbol{W}$ is restricted to the subclass of diagonal positive definite $d$-dimensional matrices

$$\mathscr{W}_d = \left\{\text{diag}(\boldsymbol{w}_1^2, \ldots, \boldsymbol{w}_d^2) : \boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d > 0\right\}.$$

$$= \frac{1}{(2\pi)^{d/2}} \cdot \exp\left(-\frac{1}{2}\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{X}_i)\right)^T\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{X}_i)\right)\right) \quad (2.7)$$

$$= \frac{1}{(2\pi)^{d/2}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{X})^T\boldsymbol{H}^{-1/2}\boldsymbol{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{X}_i)\right) \quad (2.8)$$

$$= \frac{1}{(2\pi)^{d/2}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{X})^T\boldsymbol{H}^{-1}(\boldsymbol{x}-\boldsymbol{X}_i)\right) \quad (2.9)$$

since $\boldsymbol{H}$ is positive definite, and therefore has a square root, such that

$$\boldsymbol{H}^{-1} = \boldsymbol{H}^{-1/2}\boldsymbol{H}^{-1/2}$$

This parametrisation allows different degrees of smoothing in each coordinate direction. Then for $\boldsymbol{W} \in \mathscr{W}_d$, (2.4) can be written as[8]

$$\widehat{f}(\boldsymbol{x}, \boldsymbol{H}) = \frac{1}{n}\left(\prod_{k=1}^{d} h_k\right)^{-1} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{i1}}{h_1}, \frac{x_2 - X_{i2}}{h_2}, \ldots, \frac{x_d - X_{id}}{h_d}\right). \tag{2.12}$$

If the $h$'s are assumed all equal, i.e., $\boldsymbol{H}$ belongs to the subclass

$$\mathscr{H}_i = \left\{h_1^2 \boldsymbol{I} : h_1 > 0\right\}.$$

the kernel estimator simplifies to

$$\widehat{f}(\boldsymbol{x}; h) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h}\right). \tag{2.13}$$

We note that $\mathscr{H}_i \subseteq \mathscr{H}_d \subseteq \mathscr{H}_p$, and that each of theses classes represent multivariate estimators with, 1, $d$, and $\frac{1}{2}d(d+1)$ independent bandwidth parameters.

The same principle guiding the choice of bandwidths for the univariate case apply to the multivariate setting. Following Silverman (1986), if we define the constants $\alpha = \int t_1^2 K(\boldsymbol{t})\, d\boldsymbol{t}$ and $\beta = \int K(\boldsymbol{t})^2\, d\boldsymbol{t}$, using the multivariate form of Taylor's theorem, yields the approximations for the bias and

[8] Let $\boldsymbol{W} = \operatorname{diag}(\boldsymbol{w}_1^2, \ldots, \boldsymbol{w}_d^2)$, then

$$\begin{aligned}
K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{X}_i) &= |\boldsymbol{W}|^{-1/2} K\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x} - \boldsymbol{X}_i)\right) \\
&= |\boldsymbol{W}|^{-1/2} K\left(\operatorname{diag}(\boldsymbol{w}_1^2, \boldsymbol{w}_2^2, \ldots, \boldsymbol{w}_d^2)^{-1/2}(\boldsymbol{x} - \boldsymbol{X}_i)\right) \\
&= |\boldsymbol{W}|^{-1/2} K\left(\begin{pmatrix} \boldsymbol{w}_1^{-1} & 0 & \cdots & 0 \\ 0 & \boldsymbol{w}_2^{-1} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \boldsymbol{w}_d^{-1} \end{pmatrix} \begin{pmatrix} x_1 - X_{i1} \\ x_2 - X_{i2} \\ \vdots \\ x_d - X_{id} \end{pmatrix}\right) \\
&= \left(\prod_{k=1}^{d} \boldsymbol{w}_k\right)^{-1} K\left(\frac{x_1 - X_{i1}}{\boldsymbol{w}_1}, \frac{x_2 - X_{i2}}{\boldsymbol{w}_2}, \ldots, \frac{x_d - X_{id}}{\boldsymbol{w}_d}\right)
\end{aligned}$$

Note that for $\boldsymbol{W}$ diagonal, $|\boldsymbol{W}| = \boldsymbol{w}_1 \boldsymbol{w}_2 \cdots \boldsymbol{w}_d = \prod_{k=1}^{d} \boldsymbol{w}_k$.

the variance

$$\text{bias}_h(\boldsymbol{x}) \approx \frac{1}{2}h^2\alpha\nabla^2 f(\boldsymbol{x}) \tag{2.14}$$

and

$$\text{var}\,\widehat{f}(\boldsymbol{x}) \approx \frac{1}{n}h^{-d}\beta f(\boldsymbol{x}). \tag{2.15}$$

Combining (2.14) and (2.15) yields the approximate mean integrated square error

$$\frac{1}{4}\alpha^2 \int \{\nabla^2 f(\boldsymbol{x})\}^2 d\boldsymbol{x} + \frac{1}{n}h^{-d}\beta. \tag{2.16}$$

The derivation of the mean squared error and the mean integrated squared error is analogous to the one-dimensional case. We will sketch the asymptotic expansions and concentrate on the asymptotic mean integrated squared error. As usual, has a bias part and a variance part. The bias of $\widehat{f}(\boldsymbol{x}; \boldsymbol{H})$ is defined as $\mathrm{E}\,\widehat{f}(\boldsymbol{x}; \boldsymbol{H}) - f(\boldsymbol{x})$ and the integrated squared bias

$$IB = \int \left\{\mathrm{E}\,\widehat{f}(\boldsymbol{t}; \boldsymbol{H}) - f(\boldsymbol{t})\right\}^2 d\boldsymbol{t} \tag{2.17}$$

The asymptotic integrated squared bias **AIB** is the first order term of **IB**, i.e.

$$\frac{IB - AIB}{AIB} = o(1) \tag{2.18}$$

as $|\boldsymbol{X}| \to 0$, $n \to \infty$, and $n|\boldsymbol{X}| \to \infty$. Define now the integrated variance

$$IV = \int \mathrm{E}\left\{\widehat{f}(\boldsymbol{t}; \boldsymbol{H}) - \mathrm{E}\,\widehat{f}(\boldsymbol{t}; \boldsymbol{H})\right\}^2 d\boldsymbol{t} \tag{2.19}$$

and the asymptotic integrated variance AIV analogous to AIB. Then the asymptotic mean integrated squared error, AMISE, can be calculated as

$$AMISE = AIB + AIV.$$

Here and in the following we denote with $\nabla_f$ the gradient of f and with $\boldsymbol{X}$ the Hessian matrix of second order partial derivatives of f. Then the Taylor

expansion of around $\boldsymbol{x}$ is

$$f(\boldsymbol{x}+\boldsymbol{u}) = f(\boldsymbol{x}) + \boldsymbol{u}^T\nabla_f(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{u}^T\mathscr{H}_f(\boldsymbol{x})\boldsymbol{u} + o(\boldsymbol{x}^T\boldsymbol{x}).$$

This leads to the expression[9]

$$\begin{aligned}
\mathrm{E}\,\widehat{f}(\boldsymbol{x};\boldsymbol{W}) &= \int K_{\boldsymbol{W}}(\boldsymbol{x}-\boldsymbol{u})f(\boldsymbol{u})d\boldsymbol{u} \\
&= |\boldsymbol{W}|^{-1/2}|\boldsymbol{W}^{-1/2}|\int K(\boldsymbol{s})f(\boldsymbol{x}-\boldsymbol{W}^{1/2}\boldsymbol{s})d\boldsymbol{s} = \int K(\boldsymbol{s})f(\boldsymbol{x}-\boldsymbol{W}^{1/2}\boldsymbol{s})d\boldsymbol{s} \\
&\approx \int K(\boldsymbol{s})\left\{f(\boldsymbol{x}) - \boldsymbol{s}^T\boldsymbol{W}^{1/2}\nabla_f(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{s}^T\boldsymbol{W}^{1/2}\mathscr{H}_f(\boldsymbol{x})\boldsymbol{W}^{1/2}\boldsymbol{s}\right\}d\boldsymbol{s} \\
&= f(\boldsymbol{x}) - \int \boldsymbol{s}^T\boldsymbol{W}^{1/2}\nabla_f K(\boldsymbol{s})\,d\boldsymbol{s} + \frac{1}{2}\int \boldsymbol{s}^T\boldsymbol{W}^{1/2}\mathscr{H}_f(\boldsymbol{x})\boldsymbol{W}^{1/2}\boldsymbol{s}K(\boldsymbol{s})\,d\boldsymbol{s} \\
&= f(\boldsymbol{x}) + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{W}^{1/2}\mathscr{H}_f(\boldsymbol{x})\boldsymbol{W}^{1/2}\int \boldsymbol{s}\boldsymbol{s}^T K(\boldsymbol{s})\,d\boldsymbol{s}\right) \\
&= \frac{1}{2}\mu_2\operatorname{tr}\left(\boldsymbol{W}\mathscr{H}(\boldsymbol{x})\right)
\end{aligned}$$

Assuming

$$\int \boldsymbol{u}K(\boldsymbol{u})\,d\boldsymbol{u} = \boldsymbol{0}_d,$$

and

$$\boldsymbol{u}\boldsymbol{u}^T K(\boldsymbol{u})\,d\boldsymbol{u} = \mu_2(K)\boldsymbol{I}_d,$$

then (2.7.2) yields

$$\mathrm{E}\,\widehat{f}(\boldsymbol{x};\boldsymbol{W}) - f(\boldsymbol{x}) \approx \frac{1}{2}\mu_2(K)\operatorname{tr}\left\{\boldsymbol{W}^T\mathscr{H}_f(\boldsymbol{x})\boldsymbol{W}\right\}$$

and therefore the asymptotic integrated bias is

$$AIB = \frac{1}{4}\mu_2^2(K)\int\left\{\operatorname{tr}\left\{\boldsymbol{W}\mathscr{H}_f(\boldsymbol{x})\right\}\right\}^2$$

[9]Note that $\int g(\boldsymbol{A}\boldsymbol{x})\,d\boldsymbol{x} = |A|\int g(y)\,dy$ and $\boldsymbol{z} = \boldsymbol{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y})$ and therefore $\boldsymbol{y} = \boldsymbol{x}-\boldsymbol{H}^{1/2}\boldsymbol{z}$.

The variance of $\widehat{f}(\boldsymbol{x};\boldsymbol{W})$ is given by

$$\begin{aligned}
\operatorname{var}(\widehat{f}(\boldsymbol{x};\boldsymbol{W})) &= \mathrm{E}\left(\widehat{f}^2(\boldsymbol{x};\boldsymbol{W})\right) - \left(\mathrm{E}\left(\widehat{f}(\boldsymbol{x};\boldsymbol{W})\right)\right)^2 \\
&= \frac{1}{n}|\boldsymbol{W}|^{-1/2}\int K(\boldsymbol{s})^2 f(\boldsymbol{x}-\boldsymbol{W}^{1/2}\boldsymbol{s})\,d\boldsymbol{s} - \frac{1}{n}\left(\int K(\boldsymbol{s}) f\left(\boldsymbol{x}-\boldsymbol{W}^{1/2}\boldsymbol{s}\right)\right)^2 \\
&\approx \frac{1}{n}|\boldsymbol{W}|^{-1/2}\int K(\boldsymbol{s})^2\left(f(\boldsymbol{x})\boldsymbol{s}^T\boldsymbol{W}^{1/2}\nabla_f\right)\,d\boldsymbol{s} \\
&= \frac{1}{n}|\boldsymbol{W}|^{-1/2}\int K(\boldsymbol{s})^2 f(\boldsymbol{x})\,d\boldsymbol{s} - \frac{1}{n}|\boldsymbol{W}|^{-1/2}\int K(\boldsymbol{s})^2\boldsymbol{s}^T\,d\boldsymbol{s} \\
&= \frac{1}{n}|\boldsymbol{W}|^{-1/2} f(\boldsymbol{x})\int K(\boldsymbol{s})^2\,d\boldsymbol{s} \\
&= \frac{1}{n}|\boldsymbol{W}|^{-1/2} f(\boldsymbol{x})\|K\|_2^2
\end{aligned}$$

where $\|K\|$ denotes the $d$-dimensional $L_2$-norm of $K$.

Combining the asymptotic integrated bias (AIB) and the asymptotic integrated variance (AIV) to get the AMISE for the multivariate kernel density estimator

$$AMISE = \frac{1}{4}\mu_2^2(K)\int\left\{\operatorname{tr}\left\{\boldsymbol{W}\mathscr{H}_f(\boldsymbol{x})\right\}\right\}^2 + \frac{1}{n}|\boldsymbol{W}|^{-1/2} f(\boldsymbol{x})\|K\|_2^2 \qquad (2.20)$$

If we define a scalar $w > 0$ and a $d \times d$ matrix $\boldsymbol{A}$ such that

$$\boldsymbol{W} = w^2\boldsymbol{A} \text{ where } |\boldsymbol{A}| = 1$$

then (2.20) can be written as

$$AMISE = \frac{1}{4}w^4\mu_2^2(K)\int\left\{\operatorname{tr}\left\{\boldsymbol{A}\mathscr{H}_f(\boldsymbol{x})\right\}\right\}^2 + \frac{1}{n}w^{-q}\|K\|_2^2 \qquad (2.21)$$

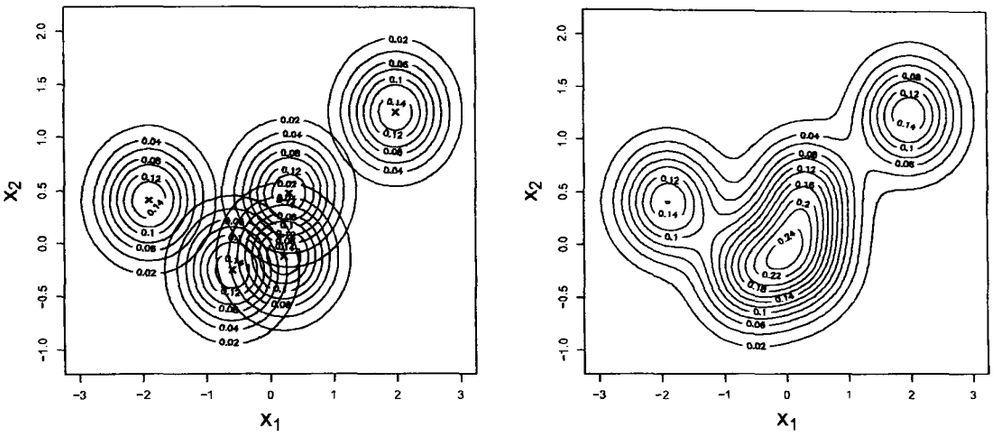
Allowing changes in $\boldsymbol{w}$ the optimal orders for the smoothing parameter $\boldsymbol{w}$ and AMISE are

$$\boldsymbol{w}_0 = O\left(n^{-1/(q+4)}\right), \text{ and}$$

$$AMISE = O\left(n^{-4/(q+4)}\right).$$

Analytic expression for the AMISE optimal bandwidth matrix are not

Figure 2.14: Construction of a bivariate kernel density estimate.




(a) Normal kernel density mass centered at each observation.

(b) Contour view of the resulting kernel density estimate. The bandwidths are $\boldsymbol{H} = \text{diag}(0.5185813, 0.3135906)$ .

available for the general multivariate case. Explicit formulae can be derived for some special cases, as it is illustrated in the next section.

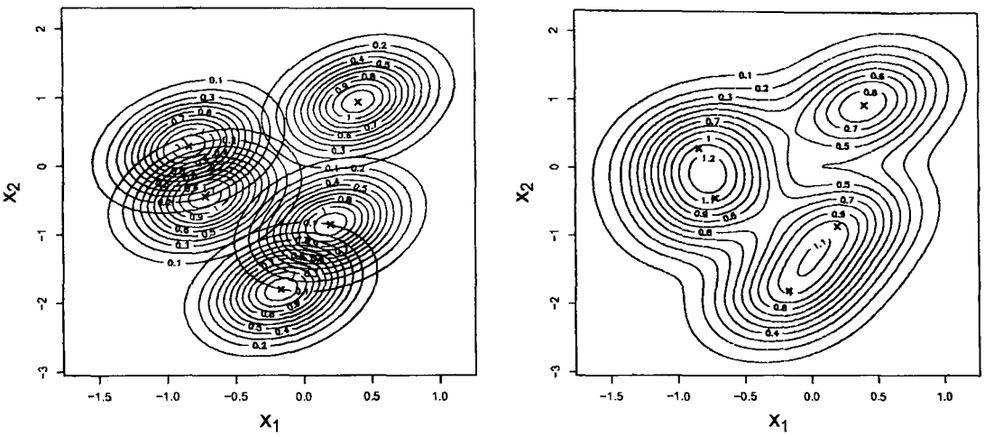
### 2.7.3 Rule-of-thumb Bandwidth Selection

Rule-of-thumb bandwidth selection provides a formula arising from a reference distribution. The obvious candidate for a reference distribution in the multivariate case is the pdf of a multivariate normal distribution, $N(\mu, \Sigma)$. Suppose that the kernel is Gaussian, i.e., the pdf of $N(\mathbf{0}, \boldsymbol{I})$. In this case, $m_2(K) = 1$ and $\|K\|_2^2 = \frac{1}{2^d \pi^{d/2}}$. From (2.20), since,

$$\int \text{tr}^2 \left( \boldsymbol{W} \mathscr{H}_f(\boldsymbol{x}) \right) \, d\boldsymbol{x} = \frac{1}{2^{d+2} \pi^{d/2} |\Sigma|^{1/2}} \left( 2 \, \text{tr}(\boldsymbol{W} \Sigma^{-1})^2 + \text{tr}^2(\boldsymbol{W} \Sigma^{-1}) \right)$$

Figure 2.14 shows a parametrization with independent normals. Figure 2.15 shows a parametrization with independent normals. More discussion on bandwith selection can be found in Scott (1992).

Figure 2.15: Construction of a bivariate kernel density estimate.




(a) Normal kernel density mass centered at each observation.

(b) Contour view of the resulting kernel density estimate. The bandwidths are $\boldsymbol{H} = \left(\begin{smallmatrix} 0.4043552 & 0.1530887 \\ 0.1530887 & 0.7748501 \end{smallmatrix}\right)$.

### 2.7.4 Kernel Selection

$K$ can also be generated from univariate kernels, $\kappa$, through a product kernel

$$K^p(\boldsymbol{x}) = \prod_{j=1}^{d} \kappa(x_j). \tag{2.22}$$

Using a product kernel 2.12 simplified further to

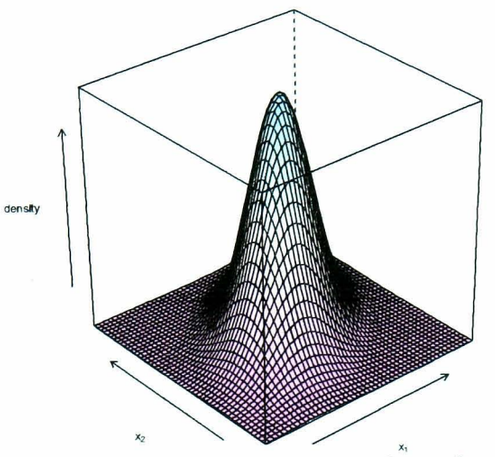
$$\widehat{f}(\boldsymbol{x}; h) = \frac{1}{nh^d} \sum_{i=1}^{n} \prod_{j=1}^{d} \kappa\left(\frac{x_j - X_{ij}}{h_j}\right). \tag{2.23}$$

The most frequently used multivariate kernel is the product kernel density estimator with normal kernels and bandwidth parametrisation $\boldsymbol{H} = \text{diag}(d_1, d_2)$, i.e.,

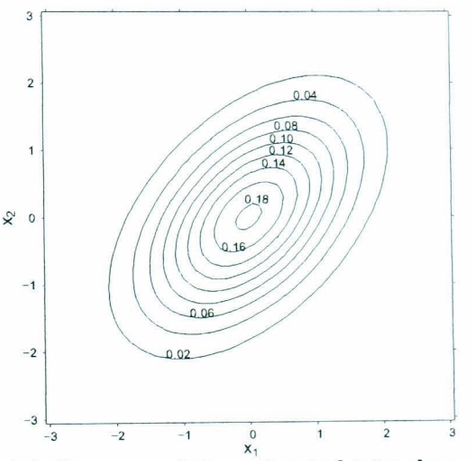
$$\widehat{f}(x_1, x_2; h_1, h_2) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} \phi\left(\frac{x_1 - X_{i1}}{h_1}\right) \phi\left(\frac{x_2 - X_{i2}}{h_2}\right).$$

Figure 2.16 shows a perspective and contour view of a bivariate normal

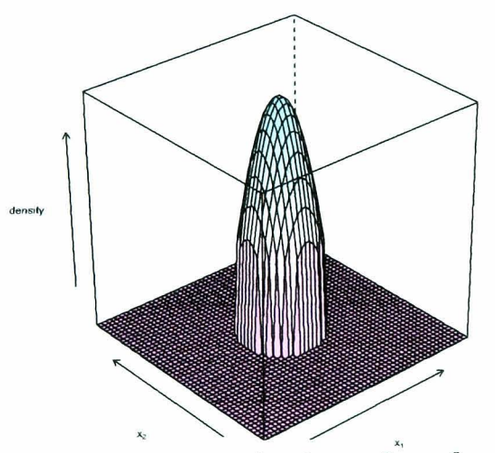
Figure 2.16: Perspective and contour plot of a bivariate normal kernel with dependent and independent normals.



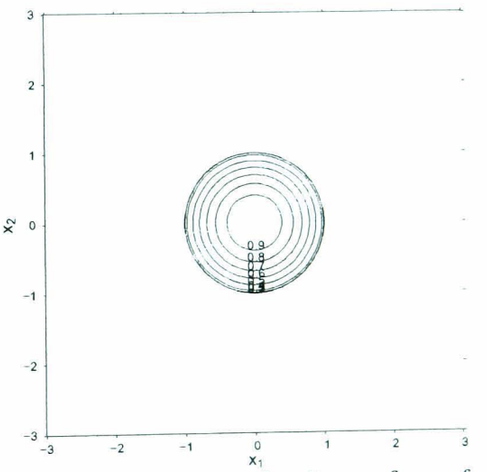

(a) Estimated joint density surface of a bivariate normal kernel.




(b) Contours of the estimated joint density surface of a bivariate normal kernel with dependent normals.




(c) Estimated joint density surface of a bivariate normal kernel with independent normals.




(d) Estimated joint density surface of a bivariate normal kernel with independent normals,

kernel using dependent and independent normals.

This version of the multivariate kernel will be the one used in subsequent chapters. Properties of this multivariate kernel are discussed in details in Scott (1992).

The next example will look at an application of the multivariate kernel to economics.

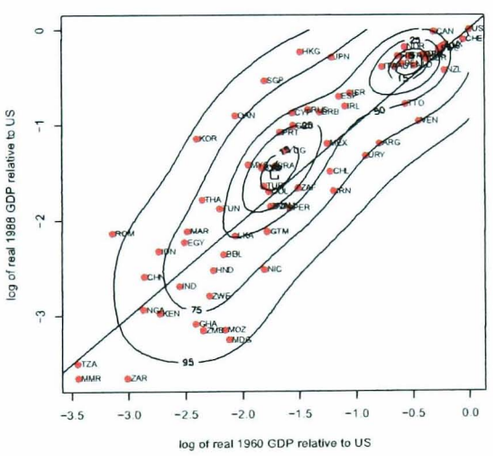
**Example 5 (Stochastic kernel)** Figure 2.17 shows a contour and perspective view of a bivariate kernel density estimate of the natural log of Real GDP per worker relative to the USA in 1960 and in 1988. The dataset used in this example comes from the Penn World Table. This estimate can be viewed as a continuous version of transition probability matrix where the number of distinct cells tend to infinity. The peaks in the perspective plot represent "basins of attraction", as countries close to one of the peaks have high probability to remain there. On the other hand, countries located in "valleys" will have a small probability to remain in the same income range. This serves as a representation of the vanishing middle class phenomenon. Also, the fact that most of the probability mass lies on the 45 degree line suggests that mobility among countries is low. □

The next example uses the univariate and bivariate kernel to produce a conditional density/distribution estimator.

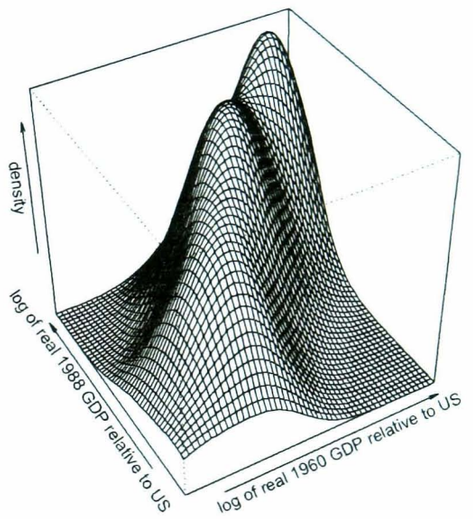
**Example 6 (Conditional density)** Using univariate and multivariate kernel, we can construct a nonparametric conditional quantile estimator as developed by Samanta (1989). A similar approach has been used by Trede (1998b) to illustrate and compare income mobility in Germany and the US. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed two dimensional random variables with joint density $f(x, y)$ and a joint distribution function $F(y, x) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)\, du\, dv$. The marginal density of $X$ is $g(x) = \int_{-\infty}^{\infty} f(y, x)\, dy$. The conditional density and distribution function of $Y$ given $X = x$ are

$$f(y|x) = \frac{f(y, x)}{g(x)}.$$

Figure 2.17: Bivariate kernel estimate of stochastic kernel.




(a) Contours of the estimated joint density surface of GDP in 1960 and in 1988.




(b) Perspective view of the estimated joint density surface of GDP in 1960 and in 1988. The bandwidth are 0.4338565 and 0.5055188 respectively for income in 1960 and in 1988.

and

$$F(y|x) = \int_{-\infty}^{y} f(u|x)\, du = \frac{\int_{-\infty}^{y} f(y,x)\, du}{g(x)}.$$

respectively. A product kernel estimate of $f(x,y)$ is

$$\hat{f}_n(x,y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h_x}\right) k\left(\frac{y - Y_i}{h_y}\right),$$

while the kernel density estimate of $g(x)$ is

$$\hat{g}_n(x) = \frac{1}{n\, h_x} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h_x}\right).$$

so that

$$\hat{f}_n(y|x) = \frac{\hat{f}_n(y,x)}{\hat{g}_n(x)},$$

and the kernel estimator of the conditional distribution function is give by,

$$\hat{F}_n(y|x) = \int_{-\infty}^{y} \hat{f}_n(u|x)\, du = \frac{B_n(y,x)}{\hat{g}_n(x)}$$

where

$$B_n(x,y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_y}\right) k\left(\frac{x - X_i}{h_x}\right)$$

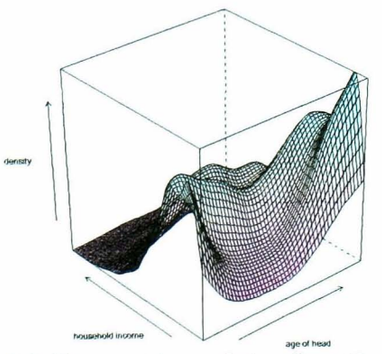
with $K(y) = \int_{-\infty}^{y} k(u)\, du$.

More derivation details, a proof of strong consistency, and of asymptotic normality can be found in Samanta (1989).

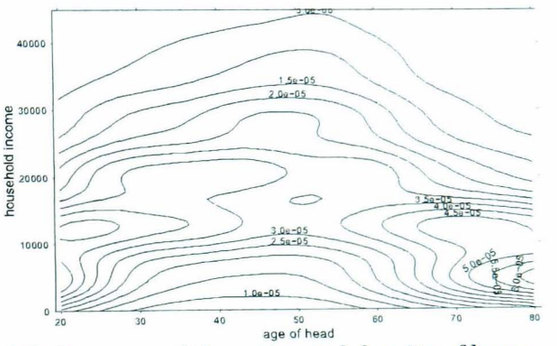
This approach is used in Chapter 4 to estimate the density of household income conditional on age of the head in Panel 4.2(a) of Figure 4.2, and the density of household income conditional on household size in Panel 4.3(b) of Figure 4.3 on page 103 in Section 6.10.

Panel 4.2 displays the contours of the estimated density of household income conditional on age of the head. The relationship between mean income and age appears to be non-linear, increasing up to the age of 50 and declining

afterwards. The contours also suggest that inequality in the distribution of household income could be functions of life-cycle factors. Income inequality seems also to increase up to the age of 50 and decline, more sharply, afterwards. Moreover, the contour view seems also to suggests that inequality is lower for older household heads than for younger ones, as the contour lines are more closely bunched together for older household head than for younger ones. □

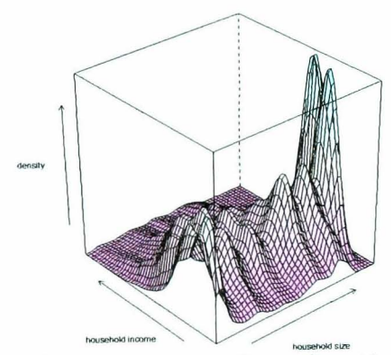



(c) Perspective plot of estimated density of household income conditional on age of head.

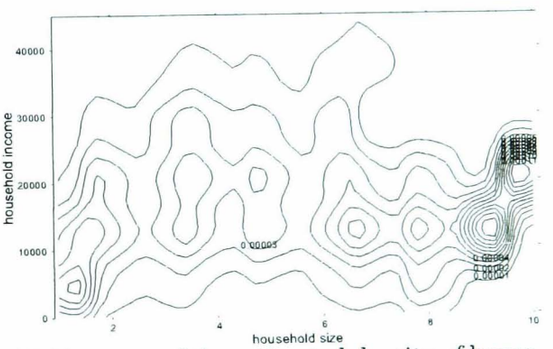



(d) Contours of the estimated density of household income conditional on age of head. The bandwidths are 3475 and 4.571 respectively for income and age.

Figure 2.18: Estimated density of household income conditional on age of head.




(a) Perspective plot of estimated density of household income conditional on household size.




(b) Contours of the estimated density of household income conditional on household size. The bandwidths are 3475 and 0.3156 respectively for income and household size.

Figure 2.19: Estimated density of household income conditional on household size.

## 2.8 Conclusion

We have provided several original examples to illustrate how these methods can be used to detect interesting features that would be harder to detect by standard parametric specifications alone. These methods can be used in conjunction with parametric methods to mutually support each others findings. Once a probabilistic structure has been identified by nonparametric means, we can adopt, if appropriate (and on an independent sample!), a fully parametric approach, to "buttress" the nonparametric results and to test relevant economic hypothesis. See the last paragraph in Appendix D on page 239 for further comments on the appropriateness of nonparametric methods.

Chapter 3

# Reporting Nonparametric Computational-Based Results

## 3.1 Introduction

Nonparametric smoothing methods have become increasingly popular among economists and statisticians in recent years and have firmly established themselves as important applied tools. Their increase in popularity can be attributed in part to their flexible nature but also to the ever growing computational power, the availability of more powerful graphic devices, and their implementation many in off-the-shelf software. Many statistical and econometrics software application offer nonparametric density and regression estimators that can be accessed with few click of a mouse or with a simple function call at a prompt. This simplicity is only apparent as important implementation details are hidden from the user's point of view. Nonparametric methods are inherently computationally intensive and rely on a plethora of implementation details that can be built-in the software application, fixed as default settings, or determined by the researcher. The control available over these implementation details is a function of both the sophistication of the software and the user. More knowledgeable users and better designed software can give greater control over the nonparametric estimation procedure. Detailed control over the estimation procedure is often required to achieve more accurate results, for correct model selection strategy, for efficiency in computation, and to facilitate reproducibility and further research. Understanding many implementation details requires knowledge of computational disciplines such as numerical analysis, computer programming, and computer graphics. Few published published papers and books report nonparametric results accurately and extensively: they often refer to published methodology and only present the graphical output. This makes assessing the quality and robustness of the result at best difficult. Lack of detailed documentation can also make nonparametric computation-based results hard to reproduce.

Hoaglin & Andrews (1975) provided a list of items that should accompany any computation-based result in statistics. In principle, any information useful to assess the accuracy of the results and to facilitate their reproduction, should be supplied. These recommendation have been echoed in many statistics and econometrics papers and books and have been incorporated in style

guides for authors of statistical journals such as the Journal of Statistical Software. Even so, almost after ten years from the publication of Hoaglin and Andrews' recommendations, Hauck & Anderson (1984) found little evidence of improvement in the reporting of computational-based results by statisticians. Recommendations focussing on more specific methods used in statistics and econometrics have also appeared in the literature. For instance, guidelines on how to present Monte Carlo results, appeared in Gentle (2003), Geweke (1996), and Baiocchi (2005). No recommendations specific to nonparametric smoothing methods have been proposed.

In this chapter we propose some basic standards to improve the use and reporting of nonparametric methods in the statistics and economics literature for the purpose of accuracy and reproducibility. We will make recommendations in four aspects of the process: computational practice, published reporting, numerical accuracy, and visualization. Section 3.2 discusses some important practical issues in nonparametric estimation. Practical aspects of nonparametric methods concern the speed of the algorithm, the ease of implementation, their numerical accuracy, reproducibility, and the availability of portable implementations. In Section 3.3 we provide guidelines for reporting computation-based nonparametric results in published research. Researchers should provide information useful to assess the accuracy of the results and to facilitate their independent reproduction. In Section 3.4 we propose a methodology to assess the numerical reliability of software implementation of nonparametric methods. Because of the nature of the estimated function visualization of nonparametric estimated curves becomes an essential part of nonparametric estimation. Section 3.5 focusses on the reproducibility of computational results. Section 3.6 discusses guidelines for the graphical presentation of estimated nonparametric curves. Section 3.7 presents an example of reporting applied to financial data. Section 3.8 concludes.

## 3.2 Computational Practice

Computing nonparametric estimates should conform to best practice from other disciplines engaged in computing. It is important to avoid reinventing

the wheel when writing software. For basic nonparametric routines there exist well written and documented routines which are implemented in many applications because of their ease of application. Table 3.1 presents a selection of R[1] can be used to perform the experiment on most computers in merely a fraction of a second. packages and functions available for nonparametric estimation. Even from a quick inspection of the table, it is apparent that different modules can provide overlapping functionality. Indeed, as an example the methods *width.SJ*, *hsj*, *dpik*, *sjpi*, respectively provided by the `MASS`, `sm`, `KernSmooth`, and `locfit` packages, can all be used to select a bandwidth for kernel density estimation using method described in Sheather & Jones (1991). From an implementation point of view, packages providing similar functionality might differ in their interface, the algorithms implemented, design, and so on. From a user point of view, the choice of which software or package to employ will depend upon several factors, such as the field of application, ease of use, efficiency, and sophistication of the user. To complicate matters further, often modules are not well documented or have only an incomplete documentation. We will show that this latter issue can produce unexpected results. Using well-established software is always recommended but has its own risks associated with it. Often, knowledge of the estimators as well as numerical and visualization methods is required for a successful use of the software.[2]

Table 3.1: R packages and functions for nonparametric density and regression estimation

[1]R is an open-source implementation of the S language (see, e.g., Ihaka and Gentleman, 1996).

[2]The hardware used in this Chapter was a Dual Intel Pentium IV (Prestonia) Xeon Processors 3.06 GHz with HT Technology with 4 GB of RAM running on Microsoft Windows XP/2002 Professional (Win32 x86) 5.01.2600 (Service Pack 2). We used R release 2.1.0, the standard Win32 release available at the time of writing the present chapter.

| Package | Function | Description |
|---|---|---|
| `stats (base)` | *density* | Computes kernel density |
| | *hist* | computes a histogram of the given data |
| | *smooth.spline* | Fits a cubic smoothing spline as described in Chambers & Hastie (1991) |

*Continued on next page*

| Package | Function | Description |
|---|---|---|
| | *ksmooth* | The Nadaraya-Watson kernel regression estimate as described in Wand & Jones (1995) |
| | *loess* | Scatter Plot with Smooth Curve Fitted by Loess as described in Cleveland et al. (1992) |
| `Graphics (base)` | *nclass.Sturges* | Computes kernel density |
| | *nclass.scott* | computes a histogram of the given data |
| | *smooth.spline* | Fits a cubic smoothing spline as described in Chambers & Hastie (1991) |
| | *nclass.FD* | The Nadaraya-Watson kernel regression estimate as described in Wand & Jones (1995) |
| `car`: Data and functions for econometrics as in Fox (2002) | *n.bins* | Computes number of bins for histograms with different rules. Implementing option "freedman.diaconis" $(n^{1/3} \cdot range)/(2 \cdot IQR)$ as described in Freedman & Diaconis (1981). "sturges" $\lceil \log_2 n + 1 \rceil$. implementing Sturges' rule Sturges (1926). "scott" $\lceil n^{1/3} \cdot range/(3.5 \cdot s) \rceil$ as in Scott (1979). and "simple" implementing $\lfloor 10 \log_{10}(n) \rfloor$ for $n > 100$. or $\lfloor 2/\sqrt{(n)} \rfloor$ for $n <= 100$. Venables & Ripley (1999) where $n$ is the number of observations. *range* is the range of $x$. $IQR$ is the inter-quartile range of $x$. and $s_X$ is the sample standard deviation of $x$. $\lfloor x \rfloor$. the floor function, denotes the integer part of $x$ while $\lceil x \rceil$. the ceiling function. denotes the smallest integer $m$ such that $m \geq x$. |



*Continued on next page*

| Package | Function | Description |
|---|---|---|
| MASS: Functions for density estimation for Venables & Ripley (1999) | *bandwidth.nrd* | A well-supported rule-of-thumb for choosing the bandwidth of a Gaussian kernel density estimator |
| | *hist.scott* | Plot a histogram with automatic bin width selection, using the Scott formula |
| | *hist.FD* | Plot a histogram with automatic bin width selection, using the Freedman-Diaconis formula |
| | *kde2d* | Two-dimensional kernel density estimation with an axis-aligned bivariate normal kernel, evaluated on a square grid. |
| | *width.SJ* | Uses the method of Sheather & Jones (1991) to select the bandwidth of a Gaussian kernel density estimator |
| | *bcv* | Uses biased cross-validation to select the bandwidth of a Gaussian kernel density estimator. |
| | *ucv* | Uses unbiased cross-validation to select the bandwidth of a Gaussian kernel density estimator. |
| KernSmooth: Functions for kernel smoothing for Wand & Jones (1995) | *bkde* | Compute a binned kernel density estimate using the fast Fourier transform as described in Silverman (1982) |
| | *bkde2D* | Compute a two-dimensional binned kernel density as described in Wand (1994) |

*Continued on next page*

| Package | Function | Description |
|---|---|---|
| | *dpik* | Select a Bandwidth for kernel density estimation using method described in Sheather & Jones (1991) |
| | *dpill* | Select a bandwidth for local linear regression using method described in Ruppert et al. (1995b) |
| | *locpoly* | Estimates a probability density function, regression function or their derivatives using local polynomials. A fast binned implementation over an equally-spaced grid is used. |
| `sm`: Functions for kernel smoothing for Bowman & Azzalini (1997) | *sm.density* | Nonparametric density estimation in one, two or three dimensions |
| | *sm.regression* | Nonparametric regression with one or two covariates |
| | *hnorm* | Normal optimal smoothing parameter |
| | *hcv* | Cross-validatory choice of smoothing parameter |
| | *hsj* | Sheather-Jones choice of smoothing parameter for density estimation |
| `locfit`: Functions for fitting local regression and likelihood models for Loader (1999) | *locfit* | Function for fitting local regression and likelihood models |
| | *sjpi* | Computes a bandwidth via the plug-in SJ method |
| | *kdeb* | Function to compute kernel density estimate bandwidths |

*Continued on next page*

| Package | Function | Description |
|---|---|---|
| `ash`: David Scotts ASH routines as in Scott (1992) | *ash1* | Computes univariate averaged shifted histogram |
| | *ash2* | Compute bivariate ASH estimate |
| `GenKern`: Computes generalised KDEs as in Lucy & Pollard (2002) | *KernSec* | Computes univariate kernel density estimate using Gaussian kernels which can also use non-equally spaced ordinates and adaptive bandwidths and local bandwidths |
| | *KernSur* | Compute bivariate kernel density estimate using five parameter Gaussian kernels which can also use non equally spaced and adaptive bandwidths |

Ueberhuber (1997) identifies four sources of uncertainty of numerical computations resulting from the use of ready-made software:

1. the risk of selecting a program that is not suitable to solve the problem at hand,
2. incorrect results due to the inadequate use of software,
3. software bugs, including design errors, code errors, and shortcomings in the documentation, and
4. bugs and incorrect use of compilers and operating systems.

Selection of the appropriate use of software for nonparametric estimation requires careful consideration and knowledge. For instance, when estimating a density using available software several parameter require careful consideration. In density estimation, bandwidth selection, number and location of grid points used to evaluate the nonparametric estimate, data transformation used, binning and other speed-enhancing approximation method used, etc. It is an established fact that different implementations of the same statistical or econometric procedure can produce different results. This is indeed the case for nonparametric methods. As more sophisticated nonparametric procedures become available in statistics

and econometric packages, particular care is required to identify defaults and options of various implementation before the procedure can be applied fruitfully and accurately.

**Bandwidth** Plug-in bandwith selectors are based on the analogy principle whereby unknown functionals that appear in the formulae of asymptotically optimal bandwidth are replaced by their sample nonparametric analogue. Many choices of bandwidth estimators and implementations are available computational methods can be made when implementing these selectors. Different implementation of the same estimator often produce different numerical values. It is important that not only the method used to select the bandwidth, but also the numerical values as well be reported for reproducibility purposes.

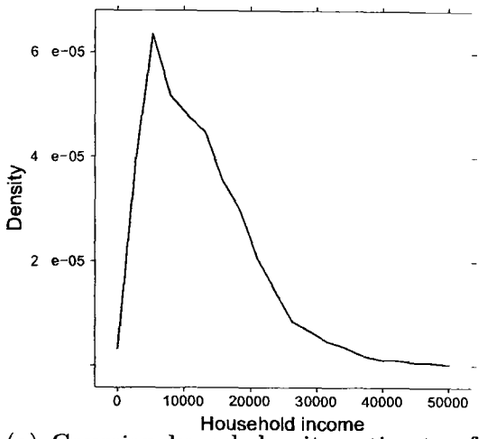
**Grid points** Figure 3.10 displays univariate kernel density estimates of income in UK for the year 1991 using 4571 observation on British households from the British Household Panel Survey (BHPS). The main features characterizing the British income distribution are positive skewness and some degree of bimodality. Panel (a) displays a density estimate of the income distribution that uses 20 grids points, whereas Panel (b) displays a density estimate with 30 grid points.[3] Note that the density in Panel (a) has no modes whereas Panel (b) displays another mode also described in Jenkins (1995b) and Schmitz & Marron (1992), where arguments in favor of a bimodal distribution of the density of household income in Great Britain are discussed.

**Binning** Applying nonparametric methods requires serious

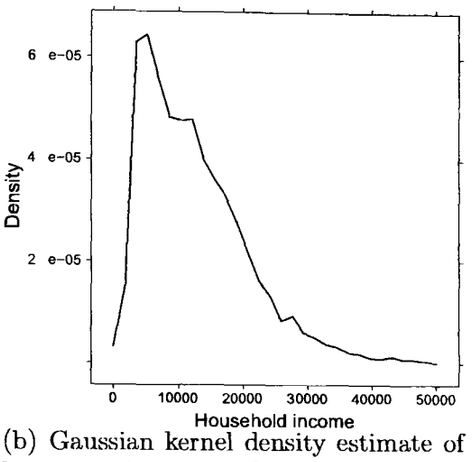
Panel (a) of Figure 3.3 seems to suggest that a larger bandwidth is needed to smooth what appear as spurious feature in the estimated density. "True" modes are masked by "spurious" modes which are an artifacts caused by the discretization of the data. If we increase the bandwidth, or adopt the normal rule for bandwidth selection, we obtain

[3]Gaussian kernel density estimate of household income evaluated on a grid of 20 equally spaced grid points in the interval $[0, 50000]$ using a bandwidth of 713.989 calculated using method described in Sheather & Jones (1991) as implemented in the R package KernSmooth.

Figure 3.1: Gaussian kernel estimate of income.




(a) Gaussian kernel density estimate of household income evaluated on a grid of 20 equally spaced grid points.




(b) Gaussian kernel density estimate of household income evaluated on a grid of 30 equally spaced grid points.

## 3.3 Published Reporting

Results based on nonparametric methods should be reported as carefully as any other computation result. Hoaglin & Andrews (1975) provided a list of items that should accompany any computation-based result. In principle, any information useful to assess the accuracy of the results and to facilitate their reproduction, should be supplied. As a minimum, taking into account recent development, the study should provide:

- information on the nonparametric estimator, including the underlying kernel, the bandwith selector used, convergence properties, etc. which should be fully adequate for the needs of the study,
- details on any transformation applied to the data to reduce boundary bias such as the logarithmic transformation and other boundary adjustments,
- number and location of grid points used for estimation,

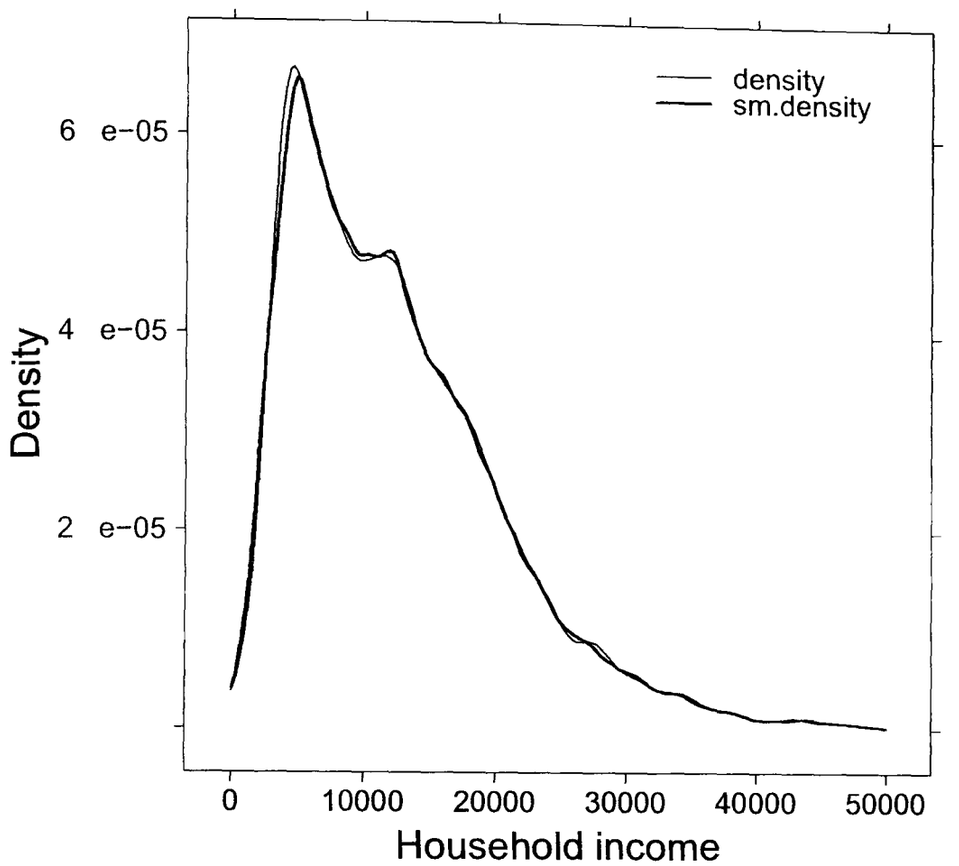
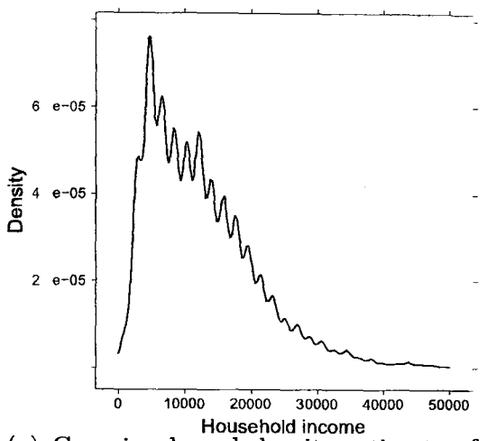



Figure 3.2: Different implementation of density estimator comparison

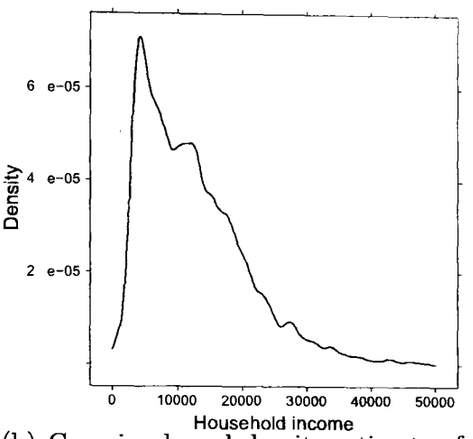
- interpolating algorithm used for the display of the results,
- details on any measure employed to speed computations such as binning or the fast fourier transform,
- detailed information of programming languages or software applications used, vendor, version, serial number, alternative platforms on which it runs, etc.,
- information on the computer used, including details on the CPU, and operating system,[4] moreover
- any published result should be checked for robustness with respect to the

[1]It is worth remembering that in the fall of 1994, a serious design flaw was discovered in the Intel Pentium processor, commonly referred to as the "Pentium floating-point-division bug" or "Pentium bug" in short. As a consequence, certain floating-point division operations performed by the Pentium processor produced incorrect results.

Figure 3.3: Gaussian kernel estimate of income.




(a) Gaussian kernel density estimate of household income using the *sm.density* function provided by the `sm` R package evaluated on a grid of 200 equally spaced grid points from 0 to 50000, with bandwidth equal to 713.989.




(b) Gaussian kernel density estimate of household income using the *density* function provided by the `stats` R package evaluated on a grid of 200 equally spaced grid points from 0 to 50000, with bandwidth equal to 713.989.

choice of alternative kernels and bandwith selectors.

All the items listed above provide information to help assess the accuracy of the nonparametric computer-based results. It is assumed that computations follow the current state of the art. Preference should be given to well-known, good algorithms and software available in the public domain. Nonparametric routines not in the public domain or that have not being tested before, should be thoroughly tested empirically before use (see Section 3.4).

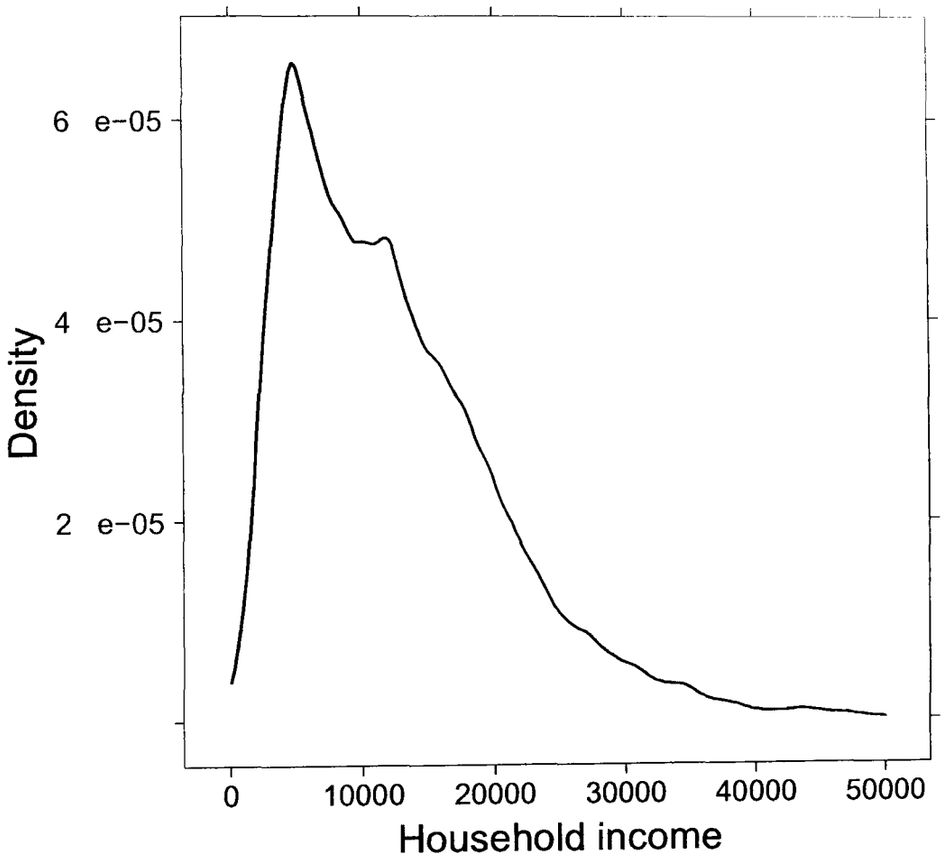



Figure 3.4: Wrong model selection strategy

## 3.4 Numerical Accuracy of Nonparametric Procedures

It is well-known that different software applications implementing a statistical or econometric procedure can produce different sets of solutions to the same estimation problem. McCullough & Vinod (1999a), McCullough (1998), McCullough (1999), Sawitzki (1994a), Sawitzki (1994b) provide examples in which the computational results obtained by several econometric and statistical packages are different. The problem of assessing accuracy is even more crucial for nonlinear procedures (see, e.g., McCullough & Renfro, 1999a; McCullough & Vinod, 2003). Sometimes the discrepancies can be attributed to implementation. In other instances, the reason for the discrepancies is less obvious. Question of accuracy can be addressed using benchmarks. However, benchmarks can be of more use than determining the accuracy of software; they can also assist in setting standard features which econometric software should possess, such as defaults and options for nonparametric procedures or bandwidth selection methods. This function of benchmarks has been highlighted in McCullough & Renfro (1999a), and is growing in importance as more computationally intensive nonlinear procedures become part of the standard researcher's toolkit.

Given the open source nature of R, considerable information about these issues can be gathered form inspecting the code directly. Even so, testing can still provide critical information for several reasons. For example, visual code inspection is not always practical as it might be too time consuming and require a considerable knowledge of R programming. Also, even though a particular algorithm is theoretically sound, it is important to assess whether it has been implemented correctly and efficiently in the software and package under scrutiny. Another potential benefit of thoroughly testing the implementations using a standard battery of tests is that it allows to make comparisons with other software packages useful for statistics that have already been tested for reviews in specialized journals.

To assess the numerical reliability of software usually the methodology proposed by McCullough (1998) is followed. This testing methodology focuses on three features of statistical software:

(i) estimation, using the Statistical Reference Datasets[5] (StRD) (Rogers et al.,

[5]Available at the web address *http://www.itl.nist.gov/div898/strd/*.

1998) from the National Institute of Standards and Technology (NIST) to evaluate the accuracy of univariate summary statistics and linear regression;

(ii) statistical distributions, using the exact values, computed with ELV (Knüsel, 1989) to verify the accuracy of statistical distributions computations; and

(iii) random number generation, using the DIEHARD (Marsaglia, 1996) Battery of Randomness tests to determine whether random numbers are seem to behave as independent samples uniformly distributed over (0,1).

To implement a benchmark a suitable reference dataset is required. Such a dataset should be well-known and easily accessible. We use data on eruptions lengths of time of Old Faithful, a well-known geyser in Yellowstone National Park in Wyoming, as the reference datatset. The Old Faithful dataset has some interesting features that make it a popular choice for examples to illustrate nonparametric methods (see, e.g., Silverman, 1986; Scott, 1992; Bowman & Azzalini, 1997; Simonoff, 1999). The version of data used is described in Azzalini & Bowman (1990) or Härdle (1991) and is provided in Appendix B. The data used consists of 272 measurements of the duration, in minutes, of an eruption of the Old Faithful geyser. The eruptions last from 1 minute and 36 seconds to 5 minutes and 6 seconds. Figure 3.5 shows a gaussian kernel density estimate obtained using a grid of 200 points in the interval $[1, 6]$ with a bandwidth of 0.15.

The figure clearly show the presence of two modes, one of "short" eruptions of 1 minute and 54 seconds and the other of longer eruptions of 4 minutes and 27 seconds. For the certified values, gaussian kernel density estimate of eruption durations evaluated on a grid of 17 equally spaced grid points in the interval $[1, 5]$ using a bandwidth of 0.15 were computed with high precision. The certified values were obtained using PARI using a precision of 100 significant decimal digits. PARI/GP is a widely used computer algebra system designed for fast computations in number theory. PARI allows fast computations with arbitrary precision arithmetic.[6]

[6]PARI is a C library, allowing fast computations with arbitrary precision arithmetic. gp is an interactive shell giving access to PARI functions, much easier to use. Pari is distributed under the terms of the GNU General Public License and is available for most commonly used computer platform. PARI-GP was originally developed in 1987 by a team led by Henry Cohen at the laboratory of number theory A2X, University of Bordeaux 1 and

| Value | Estimate *density* | Estimate *sm.density* | Estimate *kernel* | Certified value |
|---|---|---|---|---|
| 1.00 | $4{\cdot}4244466496146813e-6$ | $4{\cdot}3244018404174050e-6$ | $4{\cdot}3244018404174021e-6$ | $4{\cdot}324401840417410489e-6$ |
| 1.25 | $1{\cdot}4317615027741204e-3$ | $1{\cdot}4106366263239728e-3$ | $1{\cdot}4106366263239723e-3$ | $1{\cdot}410636626323972574e-3$ |
| 1.50 | $6{\cdot}0720458633242019e-2$ | $6{\cdot}0447409539174907e-2$ | $6{\cdot}0447409539174941e-2$ | $6{\cdot}044740953917489381e-2$ |
| 1.75 | $4{\cdot}0173446675594671e-1$ | $4{\cdot}0158966441537991e-1$ | $4{\cdot}0158966441537985e-1$ | $4{\cdot}015896644153799042e-1$ |
| 2.00 | $4{\cdot}8791375678965188e-1$ | $4{\cdot}8758384833916907e-1$ | $4{\cdot}8758384833916912e-1$ | $4{\cdot}875838483391690262e-1$ |
| 2.25 | $2{\cdot}9350465936165387e-1$ | $2{\cdot}9321389962726524e-1$ | $2{\cdot}9321389962726524e-1$ | $2{\cdot}932138996272652723e-1$ |
| 2.50 | $1{\cdot}1860145325603355e-1$ | $1{\cdot}1839925036257792e-1$ | $1{\cdot}1839925036257790e-1$ | $1{\cdot}183992503625778919e-1$ |
| 2.75 | $4{\cdot}3640929256470035e-2$ | $4{\cdot}3566855467890193e-2$ | $4{\cdot}3566855467890200e-2$ | $4{\cdot}356685546789018129e-2$ |
| 3.00 | $3{\cdot}2501747531810676e-2$ | $3{\cdot}2446841658661262e-2$ | $3{\cdot}2446841658661275e-2$ | $3{\cdot}244684165866129053e-2$ |
| 3.25 | $5{\cdot}8066251820211086e-2$ | $5{\cdot}7974448038795250e-2$ | $5{\cdot}7974448038795222e-2$ | $5{\cdot}797444803879524222e-2$ |
| 3.50 | $1{\cdot}3391036154321939e-1$ | $1{\cdot}3378481985161042e-1$ | $1{\cdot}3378481985161045e-1$ | $1{\cdot}337848198516104283e-1$ |
| 3.75 | $2{\cdot}3359237533776925e-1$ | $2{\cdot}3329325994125560e-1$ | $2{\cdot}3329325994125549e-1$ | $2{\cdot}332932599412556658e-1$ |
| 4.00 | $4{\cdot}1205538906156358e-1$ | $4{\cdot}1169672698426613e-1$ | $4{\cdot}1169672698426624e-1$ | $4{\cdot}116967269842663951e-1$ |
| 4.25 | $5{\cdot}3931955480530891e-1$ | $5{\cdot}3883539992365226e-1$ | $5{\cdot}3883539992365193e-1$ | $5{\cdot}388353999236521145e-1$ |
| 4.50 | $5{\cdot}8353432312898845e-1$ | $5{\cdot}8308556073550899e-1$ | $5{\cdot}8308556073550899e-1$ | $5{\cdot}830855607355092001e-1$ |
| 4.75 | $4{\cdot}1566358196367742e-1$ | $4{\cdot}1533449442326342e-1$ | $4{\cdot}1533449442326326e-1$ | $4{\cdot}153344944232636620e-1$ |
| 5.00 | $1{\cdot}6344012239884725e-1$ | $1{\cdot}6320444378071458e-1$ | $1{\cdot}6320444378071458e-1$ | $1{\cdot}632044437807145318e-1$ |

Table 3.2: Nonparametric estimates results for Old Faithful geyser data

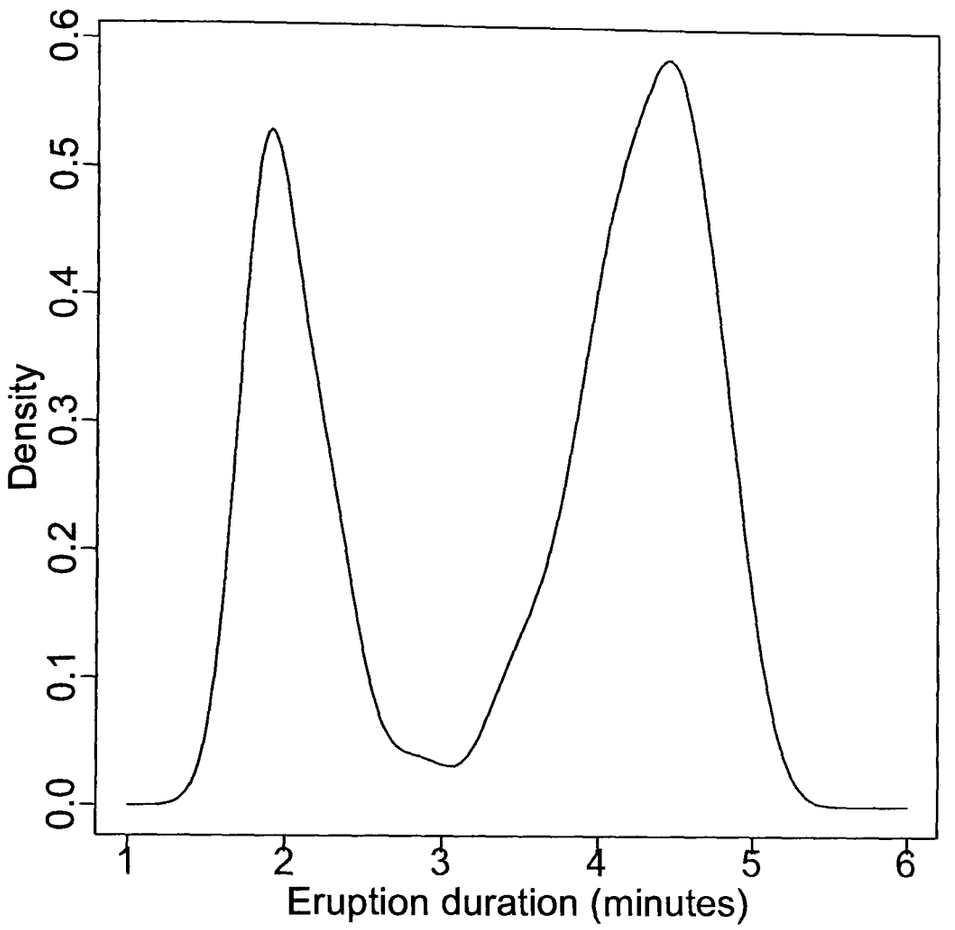



Figure 3.5: Old Faithful eruption times density estimate

The certified results are reported to 11 decimal places for each dataset. Clearly, most of these digits are not statistically significant, and we are not advocating that results should be reported to this number of digits in a statistical context. We do believe, however, that this number of digits can be useful when testing the numerical properties of a procedure. A good nonparametric density estimation procedure should be able to duplicate the certified results to at least 7 or 8 digits. There can be several reasons that a given result might not agree with the certified values. First, the code might be wrong. More probably, in this case, there might be several default assumption made that result in different estimates, different

is now maintained by Karim Belabas at the Mathematics department of the University of Paris-Sud 11 with the help of many volunteer contributors. `Math::Pari` (version 2.010603) is a Perl interface for PARI.

approximations, etc.

Table 3.2 reports the estimation results for the *density* function, the *sm.density* function provided by the sm package, a simple implementation in R, *ker*, by the author, and the certified values computed with PARI.

```
ker <- function( x, y, h ) {
       n   <- length(y)
       sum <- 0
       for (i in 1:n) { sum = sum + 1/(n*h) * dnorm( (x-y[i])/h) }
       return( sum )
}
```

Differences are generally small. The main difference is in the results for *density* which implements a binned kernel density estimator which cannot be changed (as opposed to *sm.density*).

Table 3.3 reports the accuracy of the functions under scrutiny. The accuracy of the estimates is measured by the base-10 logarithm of the relative error (LRE) given by the formula

$$lre(q, c) = -\log_{10}\left(\frac{|q - c|}{|c|}\right), \tag{3.1}$$

where $q$ represents the estimated value and $c$ the correct value. When the two values are sufficiently close, the LRE is a measure of the number of correct significant digits. The implementation in Perl used for this chapter that allows for cases where *lre* function is undefined and checks for closeness of estimated and correct values, is provided in Appendix A.

The table on accuracy confirms the impressions from the estimates.

Table 3.3: Numerical accuracy of R nonparametric kernel density estimates functions

| Grid value | LRE *density* | LRE *sm.density* | LRE *kernel* |
|---|---|---|---|
| 1.00 | 1·63573 | 14·8964 | 14·7122 |
| 1.25 | 1·82462 | 15·7972 | 15·71 |
| 1.50 | 2·34514 | 15·6615 | 15·1076 |
| 1.75 | 3·44301 | 16·8444 | 15·8693 |
| 2.00 | 3·16966 | 16·0474 | 15·7162 |
| 2.25 | 3·00365 | 15·9575 | 15·9575 |
| 2.50 | 2·76756 | 15·6255 | 16·1675 |
| 2.75 | 2·76949 | 15·5707 | 15·3672 |
| 3.00 | 2·77155 | 15·0557 | 15·3197 |
| 3.25 | 2·80038 | 15·8729 | 15·4572 |
| 3.50 | 3·02762 | 16·2048 | 15·7911 |
| 3.75 | 2·89206 | 15·5492 | 15·1227 |
| 4.00 | 3·05989 | 15·191 | 15·4237 |
| 4.25 | 3·04647 | 15·5688 | 15·4654 |
| 4.50 | 3·11372 | 15·4432 | 15·4432 |
| 4.75 | 3·10109 | 15·2344 | 15·0141 |
| 5.00 | 2·84041 | 15·5306 | 15·5306 |

## 3.5 Reproducibility of Nonparametric Computation Results

Econometrics and other traditionally empirically-oriented economic models such as input-output analysis are inherently computational. More recently, the use of ever more powerful computers and the development of increasingly sophisticated software applications has allowed economists to explore economic models with less restrictive assumptions, estimate and test richer behavioral models, experiment with different complex methodologies, compare different estimation methods, etc. All these approaches have become part of the cross-disciplinary subject we now refer to as *computational economics*. In general terms, the goal of computational economics is to advance the subjects of economics, mainly through the analysis of mathematical economic models by the application of advanced computing techniques. To appreciate the wide range of economic issues were computational methods have been brought to bear, one just needs to glance at the table of content of issues of this Journal, the *Journal of Applied Econometrics*, *Journal of Economic Dynamics and Control*, or at the papers collected in books, such as, Varian (1996), Amman et al. (1996), and Judd & Tesfatsion (2006). Many of these applications rely rather heavily on computing.[7] Increasingly often, economists use computers, not only for computations on data and model simulations, but also for simple, mechanical operations such as searching for information, collecting and storing data, changing the format of data, validating data, post-processing output from statistical applications, writing reports, handling tedious and complex algebraic manipulations, collaborating with other researchers, and in disseminating the final results. The use of computers has also benefited learning and research by suggesting conjectures and enriching our understanding of abstract economic and econometric concepts by means of examples and visualizations.

On the negative side, this increasing dependence of economists on computers

[7]We could say that these methods are computationally intensive, however this expression is rather fluid as yesterday's computationally intensive methods become today's standard approaches. As an example, Leontief (1966) recounts that in 1939 to solve a system of 42 equations in 42 unknown, in what was the first effort to analyze a large economics model through computers, required several months of programming and 56 hours of computing on the Harvard Mark II computer, one of the most powerful computers available at the time. Today the same calculations can be done in a fraction of a second on a standard PC after comparatively very little programming effort.

has resulted in research that has become increasingly difficult to replicate. Implementation details of computations in economics are left out of traditional printed publication. Reproducibility relies on a plethora of implementation details that are difficult to communicate through conventional printed publications. As Dewald et al. (1986) pointed out, this lack of information can result in months of effort by researchers trying to replicate a study yielding inconclusive results regarding the validity of the original study. Program written in specialized languages such as GAUSS are often not easily portable between different platforms and versions of the program. Programs written in conventional programming languages such as FORTRAN or C++ also depend on implementation details including the vendor, version of the compiler, and the specific platform on which they run. All these factors can amount to insurmountable obstacles in the replication of computational-based results in economics, as extensively reported by Dewald et al. (1986). For instance, Dewald et al. (1986) report that they had to abandon attempts to reproduce results from a large macroeconometric model because of difficulties in transferring programs and data across computer systems. McCullough & Renfro (1999b), in a survey of GARCH estimation procedure implemented in various packages, found that often important information that affects computed results, such as parameter initialization, was not available. Buckheit & Donoho (1995) pointed out that in the field of computational experiments researchers often cannot reproduce their own work, even only a few months after its completion, that research students have difficulties in presenting their problems to their academic advisers, and that researchers cannot reproduce computational results of other researchers and other published work. There is substantial evidence that analogous problems occur also in economics (see, e.g, Dewald et al., 1986).

Computational and empirical results in economics require independent verification in order to contribute to the advancement of the subject of economics. An important step in that direction is that published computational results should be reproducible by other researchers. Ideally, reproducibility implies that identical computational results should be obtainable in a short amount of time, without requiring expensive computational resources, proprietary data, licensed software, and any application-specific knowledge. Of course, insisting on "bit-by-bit" reproducibility of computational results in economics is not always practical and the definition must be interpreted in the light of the specific context of applica-

tion.[8] In applied work, it is quite frequent that a particular commercial software, dataset, or expensive equipment makes research results difficult to reproduce.[9] In practice, obtaining qualitatively similar results might be sufficient to claim that a computational result has been reproduced.

There has been an increasing interest in making research in empirical economics reproducible since the alarm raised by Dewald et al. (1986), in their Journal of Money, Credit and Banking (JMCB) project, in which they attempted and failed to replicate most empirical results published or submitted to the same journal. Based on their recommendation several journal introduced publicly available Internet archives and required the submission of data and programs from the authors of the empirical papers submitted. In a more recent investigation, Vinod (2001) found that approximately 70 per cent of articles from prestigious economic journals were not reproducible. He attributed this problem to sloppy record keeping, inaccurate software, and the lack of maintenance of software and data, in particular, after publication. McCullough et al. (2006) take stock of 23 years experience of the JMCB data and code archive. They convincingly argue that, though most empirical work could still not be reproduced, the requirement of a data and code archive should be adopted by more journals and that stricter rules that ensure compliance from the author should be introduced. Based on the experience, they provide guidelines to facilitated the reproduction of empirical research in economics. Open source software, software whose source code is made freely available to the public, enabling anyone to copy, modify and redistribute the source, is naturally conducive to reproducibility. In this section we want to highlight the potential role of open source software in organizing computational based research and in mediating researcher's interaction with each other, PhD students, and journal editors, by streamlining operations such as replication, validation, and supporting students' participation in the research process.

Claerbout (see, e.g., Buckheit and Donoho, 1995), has recently championed the issue of reproducibility in the computational sciences. Empirical research requires independent verification. An important step in that direction is that published computational results should be reproducible by other researchers. However, re-

[8]Gentle (2003) talks of Monte Carlo computation being *strictly* reproducible if the software and the seeds used for the random number generators are preserved.

[9]Stokes (2004) discusses the potential advantages of using different software to solve the same problem.

producing computation results from published work often proofs to be a difficult and daunting task. Reproducibility relies on a plethora of implementation details that are difficult to communicate through conventional printed publications. Buckheit & Donoho (1995) point out that in the field of computational experiments:

- researchers often cannot reproduce their own work, even a few months after the study has been completed,
- research students have difficulties in presenting their problems to their academic advisers, and
- researchers cannot reproduce computational results of other researchers and other published work.

Reproducibility implies that, ideally, identical results should be obtainable in a short amount of time, without requiring expensive computational resources, proprietary data, licensed software, and any application-specific knowledge. Schwab *et al.* (2003) classify their computational problems according to their degree of reproducibility in:

- **Easily reproducible** result files can be regenerated within ten minutes on a standard workstation.
- **Non-reproducible** result files, such as hand-drawn illustrations or scanned figures, cannot be recalculated by the reader.
- **Conditionally reproducible** result files require proprietary data, licensed software, or more than 10 minutes for their re-computation. The author nevertheless supplies a complete set of source files and rules to ensure that readers can reproduce the results if they possess the necessary resources.

Based on this stringent requirements, most computational results in economics would be classified under the headings of "conditionally reproducible" at best. In a recent investigation, Vinod (2001) found that approximately 70 per cent of articles from prestigious economic journals were not reproducible. He attributed this problem to sloppy record keeping, inaccurate software, and the lack of maintenance of software and data, in particular, after publication.

In their Journal of Money, Credit and Banking seminal project, Dewald et al. (1986) attempted to replicate computation results published or submitted to the

journal. Of the 92 authors asked to supply data according to the journal policy, 75 responded, and 68 submitted something. The first 35 datasets were examined and only 7 were judged to be free of problems. The authors attempted to replicate the results of 9 papers for which they had obtained data and software code; only four computational results could be reproduced closely. Based on their findings, Dewald et al. (1986) recommended that journals require the submission of data and programs from authors at the time empirical papers are submitted.

We can identify several reasons why full and easy reproducibility of computational results is a desirable goal in economics. As we already mentioned, reproducibility facilitates independent verification. Moreover, it could help the peer review Process and Supervision. Peer review is the scholarly process for quality assurance mostly used in economics in the publications of articles and in the awarding of research grants. This process ideally should assists authors of scholarly papers in meeting the standard of their disciplines. This process presumes that the article being reviewed has been honestly written and that no gross mistakes in the implementation of the methodology have been committed. Though occasionally problems can be detected from the printed results, the process is not designed to detect fraud or error. The reviewers usually do not have access to the datasets and software code used to obtain the computational results. Easily reproducible results would considerably help this process.

Research in economics is often a process of iterative refinement. Reproducible results are also easier to improve upon. Supervision can also benefit from reproducibility at least in two ways. Firstly, by using computational results that are easily reproducible and modifiable to solve other economic problems, research students can learn and get started with their own research. Secondly, reproducible results can be better monitored for quality.

The open source software development model proponents advocate unrestricted access to the source code of software and contend that this more open style of licensing allows for a superior software development process. This two basic tenets can facilitate reproducibility and independent verification. Open source software can be freely distributed making replication of computational results easier for individual researchers without subsidized access by a major university of commercial software.

Software vendors rely on the law of contracts and intellectual property to protect their softwares source code from being used by researchers for other purposes.

Typically, the software application is "purchased" through a licensee/licensor contract. Because the source code is kept secret, software is usually delivered to licensees in object code or executable form, i.e., in machine-only readable form. It is possible to identify several, actual and potential, advantages of using OS software for researchers, students, and academic institutions.

Typically, open source software can be obtained at the cost of the media (CDs or diskettes) or network bandwidth (for distribution via the world wide web). Commercial packages used by economist can be quite expensive, especially if upgrading and licensing occurs frequently. Cost considerations can discourage the adoption of a commercial package by institutions from developing countries, and also by resource constrained universities in more developed countries. Moreover, newer versions that add new features can make previous versions rapidly obsolete (it is of small consolation if, after a long wait, you manage to obtain code that "Requires version $x.x$ or greater, and library $y$"). Analogous problems arise when modification or extended functionality is required. Asking for features to be included is a long and tedious process. Some well known software producer are slow to respond, even in fixing serious bugs identified and reported on specialized journals. "Toolboxes", "modules", "packages", etc., can be extremely expensive, sometimes more than the core software itself. GNU's copy-left license guarantees the freedom to improve the program, and release the improvements to the public, so that the whole scientific community can benefit.

Uncertainties about the future development of a software application can also prevent its adoption. Under the GNU license, the software (and the option for support and development) will also be available if the software producer no longer exists.

Open source software is reputed to have a high degree of Reliability. Serious errors have been found in some econometric and statistical packages, (see, e.g., Knüsel, 1995; McCullough & Vinod, 1999b). Vendors of proprietary software rarely describe the algorithms used to implement econometric and statistical procedures, nor provide information about their reliability. This is a serious omission that makes the use of "black box" packages less attractive for academic research. Algorithm used, their implementation benefit form being open source. To ensure the highest standard of quality and degree of confidence in the results obtained, software should be subject to peer review as any other aspect of research and based on openly published and freely available algorithms and source code.

Open source software can promote efficiency and learning. Applied econometricians will sometimes have the necessity to engage in the process of crafting their own programs. Free software allows to obtain the source code and study it. The writing of code or the adapting of existing own to one's needs is thus facilitated. Free Software should avoids "re-inventing the wheel." The GNU license guarantees the freedom to redistribute copies, modified or not, so that the whole scientific community can benefit.

Disadvantages of opens source software include abandoned code and code "forking." For an example of software used in statistics and econometrics that has *de facto* been abandoned see the discussion on Xlisp-Stat in de Leeuw (2005). Forking of a project occurs when a developer takes code from a project and develops it independently of the original project. An example of forking is the Gnu-Emacs/XEmacs split. Forking is generally considered harmful in terms of wasted resources, but it can also create some beneficial competition as the EGCS (Experimental/Enhanced GNU Compiler System) which was a fork from GCC (GNU Compiler Collection) which was eventually reincorporated in the official GCC project. For a description of how GCC and other Unix-like software tools are used in economics see Racine (2000).

Software used by economists for research, learning, and teaching include econometrics, statistical, symbolic, and various simulation and optimization packages. Table 3.4 presents the current legal status of software useful to economists reviewed by the Journal of Applied Econometrics (JAE). The table also report the volume, issue, and page numbers were the review appears. An important distinction to keep in mind is the one between open source and freeware/shareware software. With open source software, the source code is bundled with the software and is free for everyone to inspect and acquire, with freeware/shareware the software is "free" to be distributed, but the source code is withheld from the public. Open source is made available under a variety of license types. The GNU General Public License (GPL), the GNU Lesser General Public License (LGPL), the Mozilla Public License (MPL), the BSD License, the Apache Software License, the MIT License, the Artistic License, and the Perl license are among the best known. For a an explanation of these different Open Source license flavors, consult St. Laurent (2004). The table clearly shows how most free software deemed useful for economists falls under the Open Source GPL agreement and includes a completely functional UNIX operating system (GNU Linux), programming lan-

guages and developing tools (GCC, and CYGWIN), powerful typesetting system (MiKTEX/TeTEX), and a high-level, cross-platform programming language with network and object-oriented programming support (Perl). Software applications that can compete with commercial applications traditionally used in Economics, include a high-level language for matrix and numerical computations (GNU Octave), which is comparable in terms of functionality to specialized applications such as GAUSS and MATLAB, a sophisticated programming environment for statistical computing and graphics based on the S programming language (GNU R), with functionality analogous to applications such as SAS, SPSS, STATA, or S-plus, and a complete econometric package (GRETL), still under development but already with features that makes it comparable to applications such as PcGive, EViews, and MicroFit.

GNU (sometimes pronounced "guh-NEW") is an acronym for "GNU's Not Unix". It is the name of a project by the Free Software Foundation (FSF) whose purpose is to promote the free exchange of software. The GNU project was started in order to develop a complete Unix-compatible operating system as well as an extensive set of software tools, all to be made freely available to the general public. The project has grown to include programs that were developed by many other people for their own purposes, which shared the same underlying philosophy of software freedom. For more details on the organization, see Stallman (1985). GNU's success as a catalyst in the production of free software is mostly attributable to the introduction of a form of software licensing, known as the GNU General Public License, or GPL, which encourages the free distribution of software.[10] In the next few paragraphs we briefly review some of the most successful OS project useful to economists.

GNU/Linux is a Unix-like computer operating system combined with libraries and tools from other GNU projects. Linux distributions incorporate large number of software applications with the core system. It was originally developed by Linus Torvalds for Intel microprocessors in 1991 but has since then considerably expanded to support a variety of computer architectures. A review of GNU/Linux from an economist's pointy of view can be found in MacKinnon (1999).

[10]The crucial difference between GNU software and software placed in the public domain, without copyright, is that the GNU GPL makes sure that anyone who redistributes the software, with or without changes, must pass along the freedom to make further copies and changes.

| Software Useful for Economists reviewed by the JAE | | | | |
|---|---|---|---|---|
| **Free/Open source** | | **Proprietary/Closed** | | |
| Public Domain | Netlib (www.netlib.org)[a] | Freeware[b] | BACC | **14** (6), 677–89 |
| | | | EasyReg | **13** (2), 203–07 |
| | | | Scilab | **16** (4), 553–59 |
| | | | ViSta | **17** (4), 405–14 |
| Open Source | GCC (GNU C++) **11** (2), 199–202 | Commercial | GAUSS | **15** (2), 211–20 |
| | CYGWIN Tools **15** (3), 331–41 | | EViews | **15** (1), 107–10 |
| | GNU/Linux **14** (4), 443–52 | | LIMDEP | **14** (2), 191-02 |
| | GNU Octave **15** (2000) (5), 531–42 | | MATLAB | **12** (6), 735-44 |
| | GNU R **14**, (3), 319–29 | | MicroFit | **13** (1), 77–89 |
| | GRETL **18**, (1), 105–10 | | Ox[c] | **12** (1), 77–89 |
| | Perl **18**, (3), 371–78 | | PcGive | **13** (4), 411–20 |
| | mikTEX/teTEX **16**, (1), 81–92 | | RATS | **12** (2), 181–90 |
| | | | Shazam | **14** (2), 191-02 |
| | | | SORITEC | **17** (1), 85–90 |
| | | | S-plus | **12** (1), 77–89 |
| | | | Stata | **16** (5), 637–46 |
| | | | TSP | **12** (4), 445–53 |
| | | | XploRe | **13** (6), 673–79 |
| | | | LISREL | **19** (1), 135–41 |
| | | | Maple | **10** (3), 329–37 |

Table 3.4: Legal status of software applications useful to Economists reviewed by the JAE

[a]Web site directing to mostly public domain Fortran and Java code for matrix computations including solving linear equations, eigenvalue problems, and linear least-squares problems.
[b]Software that can be downloaded free of charge.
[c]A less "user friendly" non-Windows version is available with no charge for academic uses.

GNU R,[11] an open-source programming environment for data analysis and graphics, has in only a decade grown to become a de-facto standard for statistical analysis against which many popular commercial programs may be measured. R's source code was initially written by Ross Ihaka and Robert Gentleman (see Ihaka & Gentleman, 1996) at the Department of Statistics of the University of Auckland in Auckland, New Zealand. Since the mid 90's there has been a core group (the "R Core Team") who can modify the R source code archive. R provides cutting-edge statistical and visualization methods. For an introduction on how R can be used in Econometrics see, e.g., Racine & Hyndman (2002).

GRETL, an acronym for *GNU Regression, Econometrics and Time-series Library*,[12] is a cross-platform software package for econometric analysis, written in the C programming language. GRETL is the first complete econometric software package to be released under the GNU software license. The software consists of a shared library, a command-line client program, and a graphical client program. It comes with many sample data files from Greene (2000) and Ramanathan (2002), which are immediately accessible from the menu. It supports several least-squares based statistical estimators (including two-stage least squares and panel data methods), time series models (including the Cochrane-Orcutt procedure and VARs), and some maximum likelihood methods (logit and probit). It also has built-in commands for several econometric tests (including the Chow, Hausman, and Dickey-Fuller tests). It calls gnuplot to generate graphs and is capable of generating output in LaTeX format. GRETL has been written by Allin Cottrell based on ESL (Econometrics Software Library) code written by Ramu Ramanathan of the University of California, San Diego. It can be obtained from the world wide web at `http://gretl.sourceforge.net/`, where the source package and binary distributions running on GNU/Linux and Microsoft Windows in the form of a self-extracting executable can be downloaded. Particularly noteworthy is the fact that the program is also distributed on CDs that accompany two popular econometrics textbooks, Ramanathan (2002) and Wooldridge (2002). These books use GRETL

[11] R is available from the WWW's Comprehensive R Archive Network (CRAN) located at `http://cran.r-project.org/`, where source code, additional libraries, documentation, and links to binaries distributions of R are available for various platforms, including Win32, Mac, and Unix/Linux.

[12] There is also an obvious reference to the classic fairy tale "Hansel and Gretel," in which Gretel is the mature and resourceful girl whose ingenuity saves her sibling's life from an evil witch who, after kidnapping them by means of gingerbread and candies, intends to fatten and eventually eat him.

extensively for their applied examples making GRETL a useful tool for practicing and teaching econometrics. An example of how GRETL can be used to analyze economic data can be found in Baiocchi & Distaso (2003).

GNU Octave is a high-level matrix-based language, primarily intended for numerical computations, available for different platforms at the following URL: `http://www.octave.org/`, that is mostly compatible with MATLAB. It provides a convenient command line interface for solving common numerical linear algebra problems, including the roots of nonlinear equations, integrating ordinary functions, manipulating polynomials, and integrating ordinary differential equations. It may also be used as a batch-oriented language. It is easily extensible and customizable via user-defined functions written in Octave's own language, or using dynamically-loaded modules written in C++, C, Fortran, or other languages. GNU Octave was originally written by James B. Rawlings of the University of Wisconsin-Madison and John G. Ekerdt of the University of Texas. Octave is free software distributed under the terms of the GNU General Public License (GPL) as published by the Free Software Foundation (FSF). For a survey on how Octave can be used in economics see Eddelbuettel (2000).

GNU Emacs, a program written by Richard Stallman of the Free Software Foundation, can serve as an integrated environment in which to run applications useful to economists. There is an Emacs package called ESS, an acronym for *Emacs Speaks Statistics*, (2004) which provides a standard interface between statistical and econometric programs and statistical processes. It is intended to provide assistance for interactive statistical and econometrics programming and data analysis. Languages supported include: S dialects (S-Plus , and R), LispStat dialects (XLisp-Stat, ViSta), SAS, Stata, and SPSS dialect (SPSS, PSPP).

A complete computing environment that includes all the above mentioned applications and many more is Quantian Eddelbuettel (2003). Quantian is a Linux based system that is a directly bootable and self-configuring from a single cdrom/dvdrom. Quantian comprises Knoppix (Knopper, 2003) from which it takes its base system software, along with automatic hardware detection and configuration, and scientific software such as the above mentioned applications and many more including, general purpose computer algebra systems such as Axiom, Maxima, PARI/GP , etc., numerical matrix oriented applications such as Scilab, Numeric Python, Euler, and PDL, optimization software such as lp-solve, GNU Scientific Library, programmable editors such as GNU Emacs with support for

econometric and statistical applications including Stata, SAS, S-PLUS, and GNU R, and so on.

One of the features that make many open source projects so successful is their modular nature. The functionality of modular application can be easily extended to cover more specialized areas of application. Modular application make it particularly easy to create, install, update, and access the optional code and data, with accompanying documentation, within the main application. Functions, data, and documentation provided by extra modules are easily made available to the user without the need of any application-specific knowledge typically with just one statement (`\usepackage{...}`, `library(...)`, `use ...`). This allows code written to satisfy the need of a particular researcher to be easily reused and modified by others. For instance, modules (in Perl), libraries (in R), macro packages (LaTeX) useful to economists are continuously added. Modules are made available in the main Web site were the software is distributed. So called *package managers* (for instance the MiKTeX Package Manager and the Perl Package Manager) allow the installation or update on demand of additional packages. Other applications, such as R, allow installation and updating to occur making appropriate selections from the main menu bar.

In the next section, we review some of the main advantages and disadvantages of open source software.

## 3.6 Visualization

Traditionally, results in empirical economics are presented in the form of tables. The advantage of tables is that information can be clearly organized and they show exact numerical values. However, tables can only be practically used only when the computational results can be represented or summarized as a small finite set of numbers. Often, to manage large numbers of results resulting from changes in experimental conditions, response surfaces are fitted (see, e.g., Davidson & MacKinnon, 1993, chap. 2). More often, computational results can be communicated accurately and clearly only by means of graphs. Nonparametric curves are made of linearly interpolated values of the nonparametric estimates computed on a equi-spaced fine grid of points Because of the nature of the computed results visualization has becomes an essential part of nonparametric econometrics. However no attention has been given to best practices in the visualization of computational results. Visualization should display data accurately and clearly, and should help to highlight important characteristics. A well designed graph should be able to facilitate exploration, communication, as well as calculation and processing of the computational results.

Some methods of visualization, such as kernel density estimation used to present monte carlo results in econometrics, are themselves computational methods and depend on a plethora of implementation details that can be built-in the software application, fixed as default settings, or determined by the researcher. Given the importance of visualization in nonparametric estimation, the same high standard for obtaining the computational result should be applied to the production of figures.

Function visualization methods can draw on the relevant literature in the fields of scientific visualization, psychology, and computer graphics. Several graphics parameter can affect the presentation of computational results. Excellent reference for guidelines on how to produce good quality graphs. In particular, Cleveland (1980, 1993) or Tufte (1983) should serve as a useful guides. In view of the above considerations, we feel that the nonparametric results that are displayed graphically should should be accompanied by detailed information on any interpolating, smoothing, or other algorithm used for the display of the results.

Several graphics parameter can affect the presentation of nonparametric results. For instance, the aspect ratio is a critical factor in the judgment of slope

changes.

In graphical displays of nonparametric estimates we judge the orientations of line segments to decode information about the rate of change of one variable with respect to another.

Consider, as an example, Figure 6.8 where the Nadaraya-Watson and the Local polynomial estimates of an environmental Kuznets curve for $SO_2$ emissions are shown.

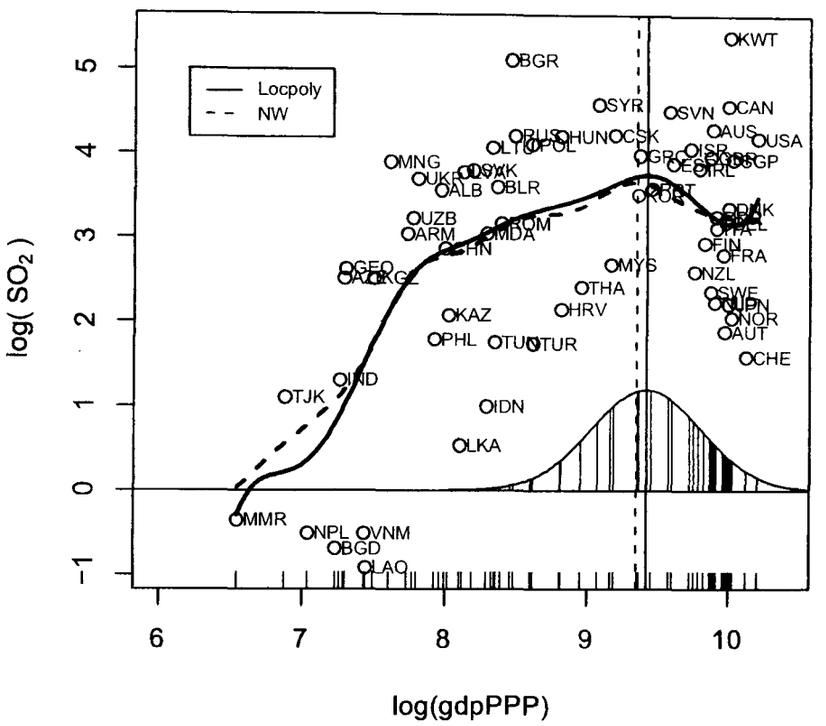



Figure 3.6: Local Polynomial and Nadaraya-Watson estimate for the $SO_2$. The two turning points data on the estimated turning point. The NW estimator assigns weights proportional to the heights of the rescaled kernel. A rugplot, which adds a mark for each observation on the x-axis, is added to aid the interpretation. The data have been jittered (a small amount of noise has been added to the data) to avoid mark's overlapping. The ISO-3166 3-letter identifications code has been used to label the countries. If the true turning point is located at high level of income the estimated turning point will be shifted to the left.

Its clear that in Figure 6.8 we judge the orientations of the short line segments that make up the estimated nonparametric environmental Kuznets curve to decode

information about the relative steepness of the curves and the amount of curvature. This decoding is greatly affected by the aspect ratio of the graph. The data rectangle of a graph is a rectangle that just encloses all of the data. The aspect ratio is the physical height of the data rectangle (measured in cm, for example) divided by the width. Figure 3.7 shows the data rectangle of an hypothetical stylized EKC, as a dashed rectangle. The aspect ratio is the height of the data rectangle in physical units divided by the width. in this case. $\frac{5.18 \text{ cm}}{9.2 \text{ cm}} \approx 0.56$. After the turning point, an increase in GDP by a thousand of US dollars results in a fall of $.75/.7 \approx 1.07$ tons of sulfur emissions.

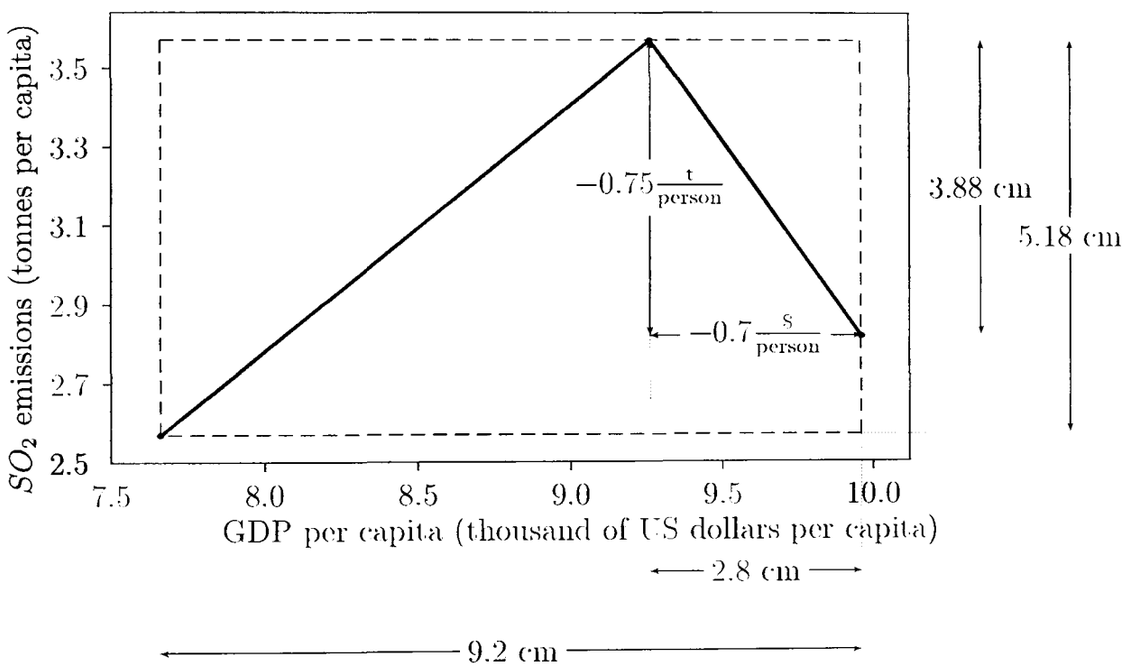



Figure 3.7: Terminology. The dashed rectangle that encloses the data is the data rectangle. The aspect ratio is the height of the data rectangle in physical units divided by the width.

To test recent models of the relationship between growth and environment we might not only be interested in determining the location of turning points but

also whether the behavior of an up swing following a down swing is symmetric. Asymmetric behavior around a turning point, besides having important consequences for the policy maker as such, might also indicate the presence of different factors affecting the downward and the upward branch of the curve. Cleveland & McGill (1987) conjectured that accuracy of comparative slope judgment is maximized when the average angle is of positively sloped line segments is set close to 45°. Cleveland refers to this averaging procedure for selecting the aspect ratio as "banking to 45°" (1994, p. 70). The conjecture is based on the maximum resolution theorem (see, Cleveland & McGill, 1987, p. 201) which states that given the orientation, in radians,

$$\alpha_i(a) = \arctan(as_i)$$

where $s_i$, for $i = 1, 2$, are the physical slopes[13] of two line segments, the *orientation resolution*, defined as the absolute difference between the orientations of two segments

$$r(a) = |\alpha_1(a) - \alpha_2(a)|$$

is maximized when the orientation of the mid-angle

$$\alpha(a) = \frac{\alpha_1(a) + \alpha_2(a)}{2}$$

is $\alpha(a^*) = \frac{\pi}{4}$. The proof of the theorem can be found in Cleveland (1994).

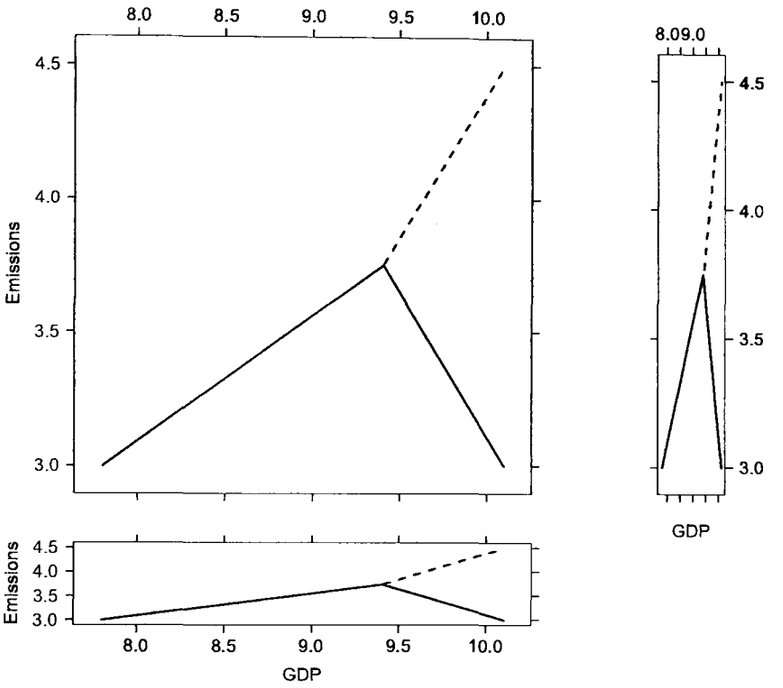
Figure 3.8

Cleveland & McGill (1987) found by experimentation that the accuracy of slope judgment increased as the slope approached 45°. Nonparametric curves are made of an entire collection of line segments. Consider a nonparametric curve consisting of $n$ line segments. Following the approach suggested in Cleveland (1994), finding the desired aspect ratio amounts to solving the following the nonlinear equation

$$\frac{\sum_{i=1}^{n} \arctan\left(a(h,v)\frac{\ddot{v}_i/\ddot{v}}{\ddot{h}_i/\ddot{h}}\right)\sqrt{\ddot{h}_i/\ddot{h} + a^2(h,v)\ddot{v}_i/\ddot{v}}}{\sum_{i=1}^{n}\sqrt{\ddot{h}_i/\ddot{h} + a^2(h,v)\ddot{v}_i/\ddot{v}}} - \frac{\pi}{4} = 0$$

[13]Physical slopes are slopes when vertical and horizontal coordinates are the physical distances from the left and bottom side of the data rectangle, where both distances are measured in the same units.

Figure 3.8: Illustration of the 45° principle. In the upper left panel the average orientation of two line segment is 45 degrees. The aspect ratios of the upper left and lower right panels are respectively larger than 6 and smaller than .2. The absolute angular separation of the latter two panels is smaller as shown with the help of the dashed line.




where $v$ and $h$ are respectively the hight and width in physical units of the data rectangle. $\ddot{v}$ and $\ddot{h}$ the hight and width respectively in scale units. $\ddot{v}_i$ and $\ddot{h}_i$ are the changes in scale units of the $i$th segment along the vertical and horizontal scale respectively. For example, for Figure 3.7, $\ddot{v} = 1$ t pc. $\ddot{h} = 2.3$ thousands of US \$ per capita. $\ddot{v}_1 = 1$ t pc. $\ddot{v}_2 = -0.75$ t pc. $\ddot{h}_1 = 1.6$ thousands of US \$ pc. $\ddot{h}_2 = 0.7$ thousands of US \$ per capita. The value of $v$ is 5.18 cm. the value of $h$, 9.2 cm. The implementation in R used for this chapter is provided in Appendix C.

The choice of aspect ratio should be dictated by the shape of the curve. Figure 3.9 shows a recursive aspect ratio plot that for each point computes the aspect ratio by banking to 45° using the 50 closest segments.

The graph clearly shows that if we want make the perception of the second mode

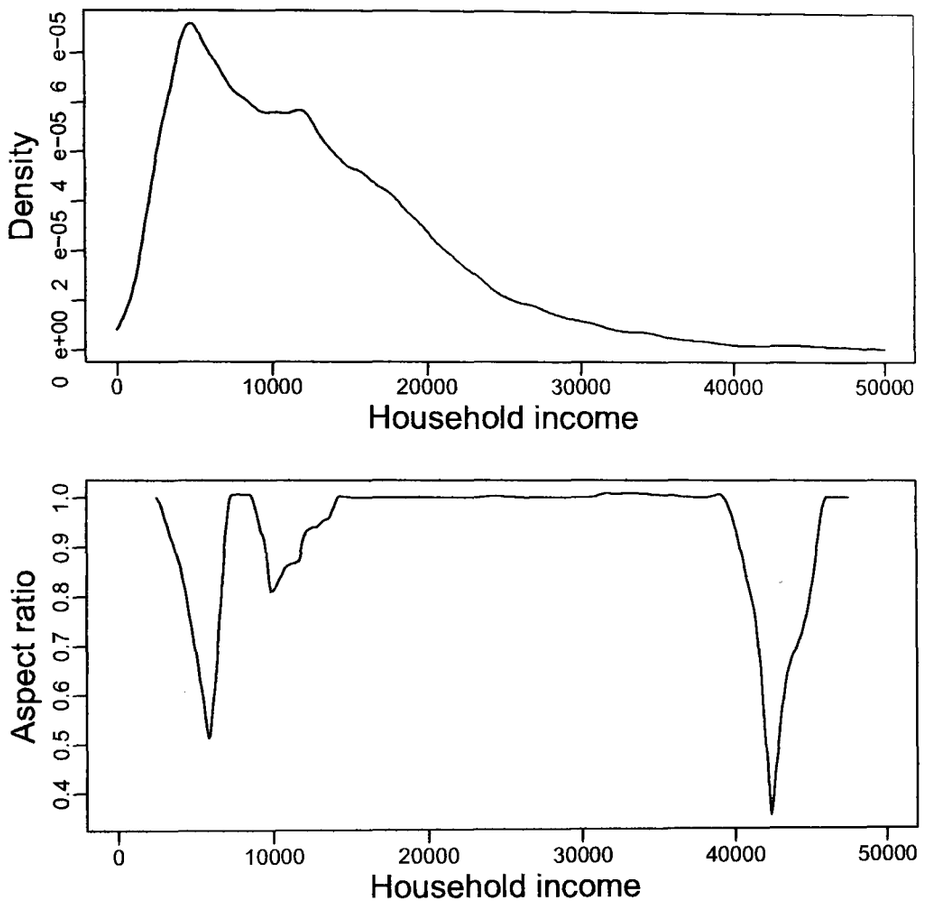



Figure 3.9: Aspect ratios

more accurate, then an aspect ratio of about 0.9 would be more appropriate. Note that the aspect ratio computed by banking using all the 500 segment constituting the curve is about 0.512. Examples of applications of this method to highlight important nonparametric results are available in other chapters.

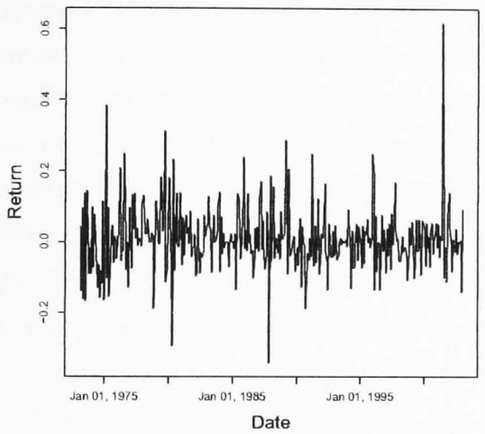
## 3.7 Example of Reporting

Consider a dataset taken from the CRSP monthly returns database produced by the Center for Research in Security Prices of the University of Chicago The Monthly Data are available for the NYSE/AMEX firms from DECEMBER of 1925 and for NASDAQ firms from December of 1972. Figure 3.10 displays a histogram and a time series plot of a typical series of returns from a randomly chosen company.[14] The sample used consists of 360 observations on 721 firms over the period from January 1973 to December 2002 for which all data was available. Panel (a) graphs the time series of returns from Jan. 1973 to Dec. 2002 using a connected plot. The series displays evidence of volatility clustering. Panel (b) displays a histogram of a typical return series.[15] A normal density with mean and standard deviation equal to their sample analogues, is superimposed on the histogram for reference. The shape of the histogram suggests that the distribution of returns appears positively skewed (sample kurtosis is 6.286), and leptokurtic (sample kurtosis is 62.918), i.e., the distribution of returns is "peaked" and "fat tailed."

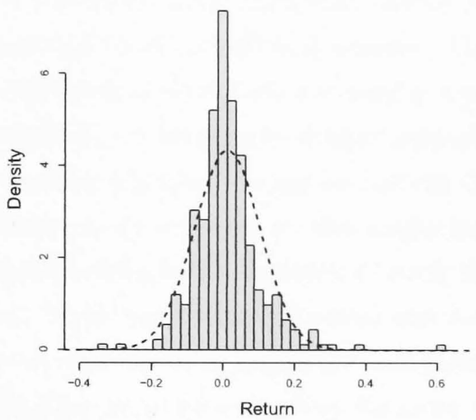
[14]The hardware used in this paper was a Dual Intel Pentium IV (Prestonia) Xeon Processors 3.06 GHz with HT Technology with 4 GB of RAM running on Microsoft Windows XP/2002 Professional (Win32 x86) 5.01.2600 (Service Pack 2).

[15]The number of bins was calculated according to the formula $\lceil n^{1/3} \cdot range/(2 \cdot iqr) \rceil$ following Freedman & Diaconis (1981) where $iqr$ is the inter-quartile range of returns, $range$ is the range of the returns, and $\lceil x \rceil$, the ceiling function, denotes the smallest integer $m$ such that $m \geq x$. Other reference rules based on the normal distribution give too few bins and an oversmoothed histogram. This rule is robust to departures from normality.

Figure 3.10: Time series plot and histogram of returns.




(a) Time series plot of monthly returns from January 1970 to December 2002 for a typical firm randomly chosen. We used R release 2.1.0, the standard Win32 release available at the time of writing the present paper, together with the routines to manipulate irregularly spaced time series provided by the ITS R package, version 1.0.9 developed by Whit Armstrong.




(b) Histogram of 360 monthly returns of panel (a) constructed over 41 equally sized bins between -0.40 and 0.65. A normal reference distribution with sample mean 0.0103 and sample standard deviation 0.0919 is shown by the dashed line.

## 3.8 Conclusion and Suggestions for Further Research

Nonparametric smoothing methods have recently become increasingly popular among economists and statisticians in recent years and have firmly established themselves as important applied tools. Their increase in popularity can be attributed in part to their flexible nature but also to the ever growing computational power, the availability of more powerful graphic devices, and their implementation many in off-the-shelf software. Many statistical and econometrics software application offer nonparametric density and regression estimators that can be accessed with few click of a mouse or with a simple function call at a prompt. This simplicity is only apparent as important implementation details are hidden from the user's point of view. Nonparametric methods are inherently computationally intensive and rely on a plethora of implementation details that can be built-in the software application, fixed as default settings, or determined by the researcher. The control available over these implementation details is a function of both the sophistication of the software and the user. More knowledgeable users and better designed software can give greater control over the nonparametric estimation procedure. Detailed control over the estimation procedure is often required to achieve more accurate results, for correct model selection strategy, for efficiency in computation, and to facilitate reproducibility and further research. Understanding many implementation details requires knowledge of computational disciplines such as numerical analysis, computer programming, and computer graphics.

In this chapter we have proposed some basic standards to improve the use and reporting of nonparametric methods in the statistics and economics literature for the purpose of accuracy and reproducibility. In particular, we made recommendations in five aspects of the process: computational practice, published reporting, numerical accuracy, reproducibility, and visualization.

Possible directions for further research include extending the benchmark from the univariate density estimator to

- bivariate density estimation with the possible choice of several popular bandwidths, and to the
- bivariate regression, again with a selected number of bandwidth selection approaches.

The best way to report the benchmarks is to have them available via the web. An obvious choice seem to make them available through the Stanford site, "Econometric Benchmarks,"[16] maintained by Clint Cummins. "Econometric Benchmarks" makes some standard benchmark datasets and models for testing the accuracy of econometrics application software available for download. So far benchmarks are available for basic statistics, linear and nonlinear regression, simultaneous equations, time series, qualitative dependent variables, panel data models, and random number generation. After having constructed the benchmarks, the next step is to test popular statistics and econometric packages that support some of these methods and to disseminate reports on how close they come to the benchmarks.

[16] Accessible at the address `http://www.stanford.edu/~clint/bench/`.

# Part II

# Economic Applications

Chapter 4

# The Determinants of Income Inequality in the UK: A Conditional Distribution Estimation Approach

## 4.1 Introduction

There seems to be a quite general consensus on the fact that Britain has experienced a dramatic increase in income inequality in the past few decades (see, e.g., Atkinson, 1997, and references therein) and that, in order to interpret the observed trend, income distribution analysis should take into account demographic and socio-economic changes in the population. In fact, at every moment in time, the heterogeneous pattern of income earning and wealth accumulation over the life-cycle of a typical individual affects the distribution's inequality. Besides life-cycle factors, other demographic and social characteristics affect the pattern of income and wealth accumulation and, therefore, the shape of the income distribution. Changes in household composition and in employment status, investment in human capital, and health issues are just a few important and recognized examples. For example, average household size has been falling in the UK over the past decades, reflecting a longer life span of individuals and an increasing preference for an independent lifestyle. Lower average fertility rates, rising average marriage age, and higher divorce rates, have also contributed to this trend.

The importance of controlling for attributes and characteristics of individuals, when investigating inequality issues, has long been recognized in the theoretical and empirical literature on income and wealth distribution. Atkinson (1971) argued that even in a egalitarian society of identical individuals in all respects apart from age, there is still likely to be considerable inequality in the distribution of current wealth as a result of age differences. In his study of U.S. family income, Paglin (1975) highlighted the importance of inter-family differences in the calculation of income inequality.

The empirical literature in this field has progressed along at least three distinct directions. One influential group of studies looks at changes in the income distribution over time and asks: "What would the distribution have been like if there had been no change in the structure of a particular determinant." Assuming that other distributional characteristics are not affected by the hypothesized "shift," the difference between the "counterfactual" and the observed distributions represents the impact of the selected determinant on the income distribution. This approach was initiated by early work by Semple (1975), Love & Wolfson (1976), and Dinwiddy & Reed (1977) and has since produced a sizable literature referred to as *shift-share* analysis or *standardization* of the income distribution. Semple

(1975), for instance, examined the effect of changes in household composition, and in the proportion of pensioner households, on the distribution of UK family income. He found that taking into account changes in household composition greatly reduces the observed increase in income inequality.

A second influential group of studies uses decomposition techniques to break up overall inequality into "within" and "between" group components. This approach was pioneered by Bourguignon (1979), Cowell (1980), Shorrocks (1982, 1984), and Mookerjee & Shorrocks (1982) examining the impact of various demographic and social factors on income inequality. For the UK, Mookerjee & Shorrocks (1982) found that the increase in inequality can be explained almost entirely by the "between" age-group component.

Although this earlier work has provided several insights into the sources of inequality, it has found it more difficult to identify the relative contribution of individual factors when several changes occur simultaneously (see, e.g., Mookerjee & Shorrocks, 1982, p. 900). Also, most of this work is descriptive in nature and lacks an adequate inferential framework.

There is a third approach, which empirically investigates the link between inequality and demographic and social factors. This branch of literature takes a different perspective and asks: "How do aggregate factors, such as the level of economic activity, inflation, and unemployment affect income inequality?" Nevertheless some of its results are relevant to our work as well. This approach typically involves regressing a measure of income inequality such as, for instance, Gini coefficients or income shares of quantiles, on a set of macroeconomic indicators and was initiated by early work by Kuznets (1955), who hypothesized that the relation between economic development and inequality follows an inverted-U shape. Work on the relationship between macroeconomic indicators, such as unemployment and inflation, and inequality include papers by Blinder & Esaki (1978) for the US, Buse (1982) for Canada, and Nolan (1988–89), for an application to the UK. Results typically show an inverse relationship between unemployment and income inequality. On the relationship between educational achievement and inequality see, e.g., Checchi (2001) and references therein. As for the question on how international trade affects inequality see, e.g., Burtless (1995).

In recent years, nonparametric methods have been applied in the study of income distribution. These methods provide visually clear and complete representation of the income distribution that is often more informative than standard

numerical measures of inequality (see Jenkins, 1995b). The use of nonparametric methods to estimate the conditional distribution of income and wealth has been pioneered by Pudney (1993) in his study on the Chinese age-income and age-wealth profiles. Trede (1998a) used a nonparametric conditional distribution estimation approach to investigate income mobility in Germany and the US.

In all these papers, the income distribution is conditioned only with respect to one determinant at a time. Though in principle the nonparametric approach is valid for the multivariate case, in practice it is fraught with the so-called curse of dimensionality problem: the rate of convergence of nonparametric estimators decreases rapidly as the number of covariates increases (Stone, 1982), thus making inference often infeasible. In order to overcome this potential limitation, we propose the use of a semiparametric method to estimate conditional measures of inequality from an estimate of a conditional distribution, in order to control for different determinants of income inequality. To estimate the conditional distribution, we resort to the semiparametric method developed by Foresi & Peracchi (1995). Conditional quantiles are obtained by inverting the estimated conditional distribution and conditional measures of income inequality are derived from the conditional quantiles. Another semiparametric approach, analogous in spirit to the shift-share approach, has been developed by DiNardo, Fortin & Lemieux (1996), and applied to the closely related field of wage inequality.

Our approach is novel in at least four respects. First, by estimating the entire conditional distribution of income over a broad set of determinants, our estimation procedure uncovers higher-order properties of the income distribution and non-linearities of its moments that cannot be captured by means of a "standard" parametric approach. For example, similar to the results obtained in the previous literature, we find that the shape of the age-income profiles agrees with the observable prediction of the life-cycle model, which assumes that resources are accumulated at a faster rate at a young age. Also, we find that income of families during the period of child rearing is higher than income in the retirement stage of the life-cycle, when economic responsibility is greatly reduced. In addition, we find that the age-income profiles peak later for the wealthier households and appear considerably non-linear, declining rapidly after the age of 60. Besides having important consequences for the policy maker as such, the asymmetry might also indicate the presence of different factors affecting the upward and downward branches of the age-income profile that have not been included in our and pre-

vious analysis. For instance, factors that determine a loss in earning capacity at retirement age of individuals, like deterioration of health and increasing aversion towards risk, could help in explaining the observed asymmetry.

Second, by estimating the whole distribution we are able to identify where in the distribution of income the various determinants exert their greatest impact. This detailed analysis can provide further insight into the determinants of inequality, of great importance to researchers as well as policy makers. For example, we find that the impact of employment status is spread over the entire income distribution. This finding seems to agree with results obtained by Nolan (1988-89) using 1977 Family Expenditure Surveys (FES) data in his analysis of the impact of UK economic conditions on income inequality. However, in addition, we find that the impact on income is substantially greater for lower income families.

Third, we devise a method for obtaining nonparametric conditional inequality measures by inverting the estimated conditional distribution. Our estimates indicate, for example, that if average household size increases from 2 to 4, households in the top 90th percentile of the income distribution move from earning 3.2 times more then households in the 10th percentile to earning about 2.5 times more. This amounts to a 20 per cent fall in inequality. This increase in inequality is obtained after controlling for other important factors, such as the age structure, the presence of a retired head and young children. Previous approaches, based on the "standardization" of inequality series, inequality decomposition by population sub-groups, or nonparametric methods, have not been to identify the contribution of individual factors on inequality, except for very simple cases.

Finally, our approach allows us to establish consistency and to estimate asymptotic variances of the proposed inequality estimators, which is useful for inference purposes. It provides a visually clear representation of both the substantive and statistical impact of each individual factor on income inequality, keeping all others constant. For instance, we find that for the UK sample, household size, number of young children, age of head, and employment status, have a large substantive and statistical impact on inequality. Factors such as years of education, marital status, and urban versus rural households, on the other hand, do not significantly impact inequality.

The paper is organized as follows. In Section 6.10 a description of the data sources and variable definitions is presented. Then in Section 6.3 the methodology used in the empirical application is outlined in detail, and conditional measures

of inequality are introduced. Section 4.4 reports the results of the estimation procedure and Section 6.11 concludes.

## 4.2 Data Description

The data used in the analysis have been taken from the database produced by the Consortium of Household panels for European socio-economic Research (CHER).[1] The CHER database for United Kingdom (UK) is based upon the results of the British Household Panel Survey (BHPS), which is carried out in the UK annually over a target sample size of 5000 households. The units specified in the data survey are adult individuals (16+ years of age), families, and households. The chosen definition of household is "One person living alone or a group of people who either share living accommodation or share one meal a day and who have the address as their only or main address".

Following previous studies and analysis in the field, the unit object of the analysis has been identified as the household, since it is believed that many economic decisions are taken at the household level (see, e.g., Jenkins, 1995b). The survey collects information about a variety of aspects of the units considered, from demographic and educational, to family (and household) structures, labour participation and main features of the job, economic, social and health status.

The response variable is the (natural log of) disposable (net) income of the household, defined in the following way

$$\begin{aligned} \text{disposable (net) income} \equiv\ & \text{total Pre-government income} \\ & + \text{total (non-pension) public transfer income} \\ & + \text{total pension income} \\ & + \text{total income from other sources} \\ & - \text{income taxes} \end{aligned}$$

[1]The aim of CHER is to create an international comparative micro database containing longitudinal datasets from many national household panels and from the European Household panel study (ECHP). This will provide the basis to facilitate comparative cross-national and longitudinal research and to study processes and dynamics of policy issues related to family structures, educational aspects, labour force participation, income distribution, poverty, etc. Access to the (beta version of the) database has been granted while visiting the Integrated Research Infrastructure in the Socio-Economic Sciences (IRISS) at CEPS/INSTEAD.

—contributions to social insurance and pension.

According to this definition, theoretically negative values for the disposable income are allowed. The BHPS does not actually ask for the disposable (net) income directly; data are recovered integrating the information available in the survey with other data sources. Once the disposable income of the household has been obtained, it has been associated with some individual characteristics of the main breadwinner inside the household (referred to as "the head" in the remainder of the paper), and with some features of the household itself.

The income measure has been adjusted for household composition according to the McClements equivalence scale (see McClements, 1977).[2]

Previous empirical studies and findings, combined with data availability, have provided the basis and guidance for the choice of predictors. The following set has been selected:

- age of the main breadwinner
- gender of the main breadwinner (male = 1, female = 0)
- marital status of the main breadwinner (married = 1, not married = 0)
- the main breadwinner is retired (retired = 1, not retired = 0)
- the main breadwinner is employed (employed = 1, not employed = 0)
- number of years of education} of the main breadwinner. This variable is not directly recorded in the Survey, which rather collects the highest level of education achieved. Therefore it has been obtained indirectly, assigning to each level of education the number of years necessary to achieve it (the variable takes the values 7, 12, 14, 17 and more)
- urban/rural indicator (urban = 1, rural = 0)
- household size
- number of people in the household with less than 16 years of age.

[2]For a discussion of how the choice of equivalence scale affect inequality measurements see Glewwe (1991), Coulter et al. (1992), and Banks & Johnson (1994).

Data are available for the period 1991–99, with an average of 4000 observations per year. We leave out of the analysis household headed by individuals older than 80 years of age because of the low sample information.[3] Summary statistics for the variables used in this study for the 1991 year appear in Table 5.1.

We will highlight the potential importance of controlling for explanatory variables when analyzing income distribution by means of the following illustrative example. Panel 4.1(c) of Figure 4.1 displays the univariate kernel density estimate of income in UK for the year 1991 (solid line) decomposed into the weighted sum of the densities of retired (dotted line) and working (dashed line) heads. The main features characterizing the income distribution are positive skewness and some degree of bimodality.[4] The figure suggests that positive skewness and the bimodal structure of the marginal distribution of income for the UK could be due to the presence of pensioners in the population. This example illustrates that the shape of the income distribution could be considerably influenced by demographic characteristics. The bivariate kernel surface estimate of the joint density of household income and age of head, displayed in panel 4.1(d), also seems to support this conclusion.[5]

The conditional distribution provides a clearer understanding on how age and income are related. Figure 4.2 displays two views of the estimated density surface of household income conditional on age of the head. Panel 4.2(a) displays a perspective view of the estimated density of household income conditional on age of the head. For any value of age, the curve resulting from slicing the surface with the vertical plane passing through that value and parallel to the income axis, gives the density of income conditional on the chosen value of age.

Panel 4.2(b) displays the contours of the estimated density of household income conditional on age of the head. The relationship between mean income and age appears to be non-linear, increasing up to the age of 50 and declining afterwards. The contours also suggest that inequality in the distribution of household income could be functions of life-cycle factors. Income inequality seems also to increase

[3]This exclusion should also mitigate the effects of a potential source of sample bias. In fact, because wealthier individuals have a higher survival probability, they might be over represented in older households (see also, Jappelli & Modigliani, 2005).

[4]The same features are described in Jenkins (1995b) and Schmitz & Marron (1992), where arguments in favor of a bimodal distribution of the density of household income in Great Britain are discussed.

[5]The marginal and joint distribution were estimated, respectively, using a univariate gaussian and a bivariate gaussian product kernel.

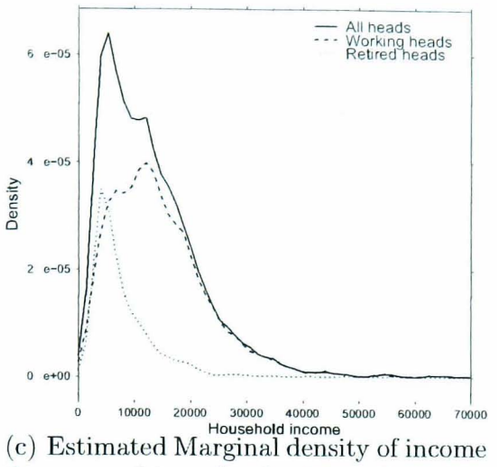
| Variable | Mean | Std. Dev. | Minimum | Maximum | Cases |
|---|---|---|---|---|---|
| Household income[a] | 12·690 | 8·941 | 0·007 | 126·4 | 4571 |
| Age of head | 48·35 | 16·598 | 17 | 80 | 4571 |
| Years of education[b] | 9·658 | 3·6413 | 7 | 17 | 4571 |
| Household size | 2·476 | 1·3080 | 1 | 8 | 4571 |
| Number of children[c] | 0·5872 | 0·9761 | 0 | 6 | 4571 |
| Employed (employed =1) | 0·5946 | 0·4910 | 0 | 1 | 4571 |
| Retired (retired=1) | 0·2260 | 0·4183 | 0 | 1 | 4571 |
| Gender (male=1) | 0·6211 | 0·4852 | 0 | 1 | 4571 |
| Urban (urban=1) | 0·7946 | 0·4041 | 0 | 1 | 4571 |
| marital (married=1) | 0·5574 | 0·4967 | 0 | 1 | 4571 |

[a] Income is expressed here in thousands of 1991 UK pounds.

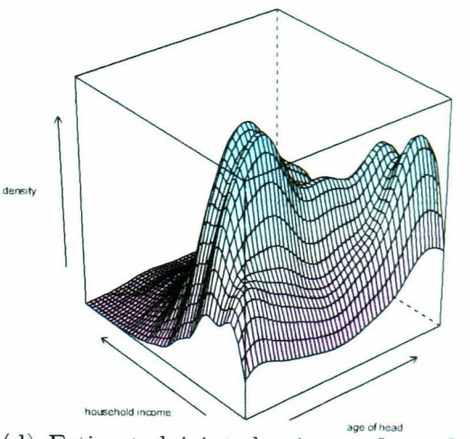
[b] Years of education is a discrete variable that takes only the values 7, 12, 14, 17. The last value represents an all inclusive category indicating the completion of 17 or more years of education.

[c] Number of members of the household with 16 or less years of age.

Table 4.1: Summary statistics




(c) Estimated Marginal density of income decomposed into the densities of working (73 per cent of the sample) and retired (27 per cent of the sample) heads. The bandwidth for income is 1000.




(d) Estimated joint density surface of income and age. The bandwidths are 3475 and 4.571 respectively for income and age.

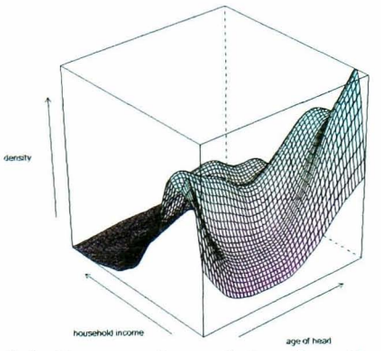
Figure 4.1: Marginal density of income and joint density of income and age.

up to the age of 50 and decline, more sharply, afterwards. Moreover, the contour view seems also to suggests that inequality is lower for older household heads than for younger ones, as the contour lines are more closely bunched together for older household head than for younger ones.

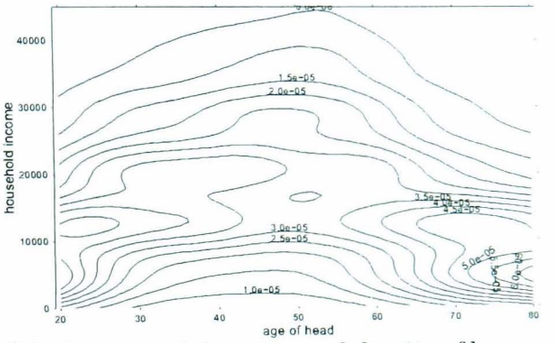
Figure 4.3 shows a perspective view and a contour plot of the estimated density of household income conditional on household size.[6] Panel 4.3(b) shows that the relationship between mean income and household size is also appreciably non-linear. Mean income seems to increase with household size up to 5, decrease afterwards, and increase again after a size of 8. The conditional inequality also seems to vary considerably suggesting, for example, that large household have a more stable income.

These pictures, though interesting, could be misleading as important determinants of income are not controlled for; therefore they represent only marginal relationships. Consider, for example, the impact of household size on income. Clearly not all the members of the household will contribute to the household income. For instance, the number of small children could affect income not only

[6]The conditional distribution was estimated using an univariate gaussian kernel and a bivariate gaussian product kernel with window width of 3475 for income and 0.3156 for household size.

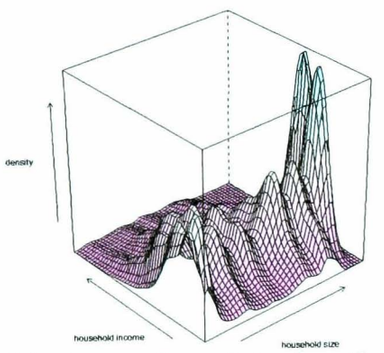



(a) Perspective plot of estimated density of household income conditional on age of head.

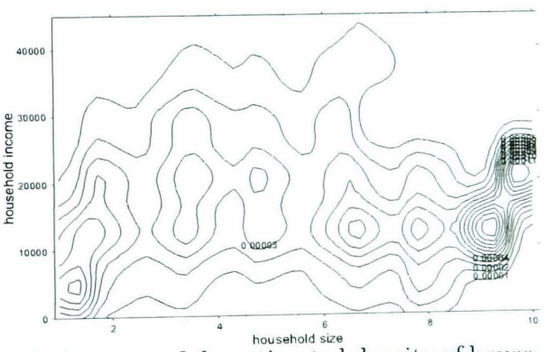



(b) Contours of the estimated density of household income conditional on age of head. The bandwidths are 3475 and 4.571 respectively for income and age.

Figure 4.2: Estimated density of household income conditional on age of head.




(a) Perspective plot of estimated density of household income conditional on household size.




(b) Contours of the estimated density of household income conditional on household size. The bandwidths are 3475 and 0.3156 respectively for income and household size.

Figure 4.3: Estimated density of household income conditional on household size.

directly because of the lack of earnings, but also indirectly, because of the time and effort needed for their care.

This example shows that a deeper insight about polarization and inequality in income distribution analysis could be gained by controlling for various determinants of income. In the next Section the methodology followed in our empirical analysis will be outlined in detail.

## 4.3 Semiparametric Estimation Method and Conditional Inequality Measures

Even though the conditional mean is an important characteristic of a distribution, it does not summarize all the information contained in it. Higher order moments can often provide a deeper understanding of the relation existing among variables. Also, the linearity assumption of the conditional mean is very restrictive and, in cases like this one, may not be appropriate.

The estimation method employed in this paper provides a detailed description of the cumulative distribution of household income, without relying on strong parametric assumptions. The conditional distribution, the "fundamental" econometric object of analysis, could uncover higher-order properties of the distribution and non-linearities of its moments that cannot be captured by means of a "standard" linear regression analysis, which focuses on just one of its moments.

The following subsections show how an estimate of the conditional cumulative distribution of income can be obtained by means of a sequence of logit models.

The data are assumed to be a realisation of a strictly stationary stochastic process $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{\infty}$, where $Y_i$ is a scalar response variable and $\boldsymbol{X}_i$ is a $k$-dimensional vector of covariates (with $k \geq 1$). This general framework includes the particular case were the pairs $(\boldsymbol{X}_i, Y_i)$ are independent and identically distributed. Let $F(y|\boldsymbol{x})$ be the conditional distribution of $Y_i$ given $\boldsymbol{X}_i = \boldsymbol{x}$, which we assume to be smooth in both $\boldsymbol{x}$ and $y$. We are interested in estimating $F(y|\boldsymbol{x})$ from a random sample $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n}$.

### 4.3.1 Estimating the Conditional Distribution Function of Income

Suppose that we are interested in estimating the conditional probability that a person's income falls below a specific threshold value. Typically, if we wish to investigate poverty, we might be interested in the individual's probability of falling below a certain poverty line $y$. In general, if we define a new random variable $Z_i = 1_{\{Y_i \leq y\}}$, then we know that $\mathrm{E}[Z_i \mid \boldsymbol{X}_i = \boldsymbol{x}] = F(y|\boldsymbol{x})$,[7] and therefore the estimation of the conditional distribution may be viewed as a regression of $Z_i$ on

[7]From basic probability theory, the expectation of an indicator function is the probability of the associated event, that is

$\boldsymbol{X}_i$. In the next subsection we review the utilized semi-parametric approach.

### 4.3.2 The Semiparametric Approach

The semiparametric method to estimate conditional distribution functions followed in this paper has been suggested by Foresi & Peracchi (1995). It consists of estimating a sequence of conditional logit models over a grid of values in the support of the dependent variable (in this case, income).

Following this method it is possible to condition upon a broad set of predictors, which can have an influence on determining the behaviour of income. This constitutes an advantage relative to fully nonparametric methods, which can be not feasible to employ when the number of predictors becomes moderate to large (usually, greater than 3). Furthermore, the method enjoys the feature of economic interpretability; in fact, using the linear logit specification, one can think of the effects of the different predictors on income in terms of "derivatives".

As previously described, the semiparametric approach consists of running $J$ distinct logistic regressions on the binary variables $Y_{j,i} \equiv 1_{\{-\infty < Y_i \leq y_j\}}$, where $y_1 < \cdots < y_J$ are distinct points in the support of $Y$, $j = 1, \ldots, J$ and $i = 1, \ldots, n$. By estimating $J$ distinct functions $F(y_j|\boldsymbol{x})$, it is then possible to approximate the conditional distribution $F(y|\boldsymbol{x})$, defined as

$$\begin{pmatrix} F(y_1|\boldsymbol{x}) \\ F(y_2|\boldsymbol{x}) \\ \vdots \\ F(y_J|\boldsymbol{x}) \end{pmatrix}.$$

Following Foresi & Peracchi (1995), it is possible to impose that the sequence of conditional distributions is bounded between 0 and 1, by modeling the log-odds ratios, defined as $\eta(y_j|\boldsymbol{x}) = \ln(F(y_j|\boldsymbol{x}) / (1 - F(y_j|\boldsymbol{x})))$. Given an estimate of

---

$$\mathrm{E}[1_{\{A\}}] = 1 \cdot \Pr(A) + 0 \cdot \Pr(A^C) = \Pr(A).$$

So if $A = [Y_i \leq y | \boldsymbol{X}_i = \boldsymbol{x}]$

$$\mathrm{E}[1_{\{Y_i \leq y | \boldsymbol{X}_i = \boldsymbol{x}\}}] = \Pr(Y_i \leq y | \boldsymbol{X}_i = \boldsymbol{x}),$$

that is $E[A] = F(y|\boldsymbol{x})$.

$\eta(y_j|\boldsymbol{x})$ it is possible to recover an estimate of $F(y_j|\boldsymbol{x})$ through the relationship

$$F(y_j|\boldsymbol{x}) = \frac{\exp(\eta(y_j|\boldsymbol{x}))}{1+\exp(\eta(y_j|\boldsymbol{x}))}.$$

A convenient and easily interpretable (from an economic point of view) way of dealing with the log-odds ratios is to impose linearity, i.e.

$$\eta(y_j|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}_j,$$

where $\boldsymbol{\beta}_j$ is a vector of coefficients. In this way each component of $\boldsymbol{\beta}_j$ can be interpreted as the constant partial derivative of the log-odds ratio of $F(y_j|\boldsymbol{x})$ with respect to the relevant predictor variable. Notice that this way of modeling the log-odds ratios is equivalent to running ordinary logit regressions on $F(y_j|\boldsymbol{x})$.

One of the potential limitations of the method outlined above is that it does not guarantee the monotonicity property of the conditional distribution function.[8] The potential violation the monotonicity property could create difficulties when inverting the conditional distribution estimate to obtain the estimates of the conditional quantiles. Particularly problematic is the possibility of multiple solutions. We have decided to retain the ordinary logit specification, because of its direct interpretability and also because in the empirical application monotonicity is rarely violated.

Under mild regularity conditions, the logit estimators are consistent and asymptotically normally distributed,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{j,n} - \boldsymbol{\beta}_j\right) \overset{a}{\sim} \mathrm{N}\left(\mathbf{0}, \mathscr{I}_j^{-1}\right),$$

where $\mathscr{I}_j$ can be consistently estimated by

$$\widehat{\mathscr{I}}_j = \sum_{i=1}^{n} \widehat{F}_n(y_j|\boldsymbol{x}_i)\left(1-\widehat{F}_n(y_j|\boldsymbol{x}_i)\right)\boldsymbol{x}_i\boldsymbol{x}_i'.$$

Using the delta method it is immediate to derive the limit theory for the estimator

[8]For more details, see Foresi & Peracchi (1995). Peracchi (2001) suggests a method of modeling the log-odds ratios which implies monotonicity of the conditional distribution function.

of $F(y_j|\boldsymbol{x})$:

$$\sqrt{n}\left(\widehat{F}_n\left(y_j|\boldsymbol{x}\right) - F\left(y_j|\boldsymbol{x}\right)\right) \overset{a}{\sim} \mathrm{N}\left(0, F\left(y_j|\boldsymbol{x}\right)^2\left(1 - F\left(y_j|\boldsymbol{x}\right)\right)^2 \boldsymbol{x}'.\mathscr{J}_j^{-1}\boldsymbol{x}\right).$$

It is possible to generalize the results considering the stacked vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_J')'$. Then

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) \overset{a}{\sim} \mathrm{N}\left(\mathbf{0}, \mathscr{J}^{-1}\right),$$

where

$$\mathscr{J} = \sum_{i=1}^{n}\left[\boldsymbol{V}\left(\boldsymbol{x}_i\right) \otimes \boldsymbol{x}_i\boldsymbol{x}_i'\right]$$

and $\boldsymbol{V}\left(\boldsymbol{x}_i\right)$ is a $J \times J$ matrix with generic element

$$V\left(\boldsymbol{x}_i\right)_{ms} = \min\left(F\left(y_m|\boldsymbol{x}_i\right), F\left(y_s|\boldsymbol{x}_i\right)\right) - F\left(y_m|\boldsymbol{x}_i\right)F\left(y_s|\boldsymbol{x}_i\right),$$

with $m, s = 1, \ldots, J$ and

$$\min\left(F\left(y_m|\boldsymbol{x}_i\right), F\left(y_s|\boldsymbol{x}_i\right)\right) = \begin{cases} F\left(y_m|\boldsymbol{x}_i\right), \text{ if } m < s \\ F\left(y_s|\boldsymbol{x}_i\right), \text{ if } s < m \end{cases}.$$

Therefore, letting $\boldsymbol{A}\left(\boldsymbol{x}\right) = \left(\boldsymbol{I}_J \otimes \boldsymbol{x}'\right).\mathscr{J}^{-1}\left(\boldsymbol{I}_J \otimes \boldsymbol{x}\right)$, the limit theory for $F\left(y|\boldsymbol{x}\right)$ is given by

$$\sqrt{n}\left(\widehat{F}_n\left(y|\boldsymbol{x}\right) - F\left(y|\boldsymbol{x}\right)\right) \overset{a}{\sim} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}\left(\boldsymbol{x}\right)\boldsymbol{A}\left(\boldsymbol{x}\right)\boldsymbol{V}\left(\boldsymbol{x}\right)\right). \tag{4.1}$$

Notice that, since $F\left(y|\boldsymbol{x}\right)$ belongs to a class of uniformly bounded functions satisfying the $L^2$ continuity condition, then the convergence established in (4.1) holds as a process and not just pointwise.

### 4.3.3 Conditional Income Inequality Measures

In this study we construct two conditional quantile-based measures of inequality. Our procedure is in the spirit of Pudney (1993), who defined an age-specific wealth inequality measure based on the unconditional interquartile range coefficient (IQRC). Denoting by $F(w|a)$, the (strictly increasing) conditional distribution of wealth, $w$, given age, $a$, the age-specific inequality measure was defined by Pudney as

$$IQRC(w|a) = \frac{Q^{3/4}(a) - Q^{1/4}(a)}{Q^{1/2}(a)},$$

where $Q^p(a) = \{w \in \mathbb{R} : F(w|a) = p\}$. Pudney used a nonparametric kernel approach to estimate the conditional distribution of wages, $w$, given age, $a$, $F(w|a)$ and then invert the distribution to obtain the conditional quantiles. The main advantage of using a kernel-based approach is that the resulting $\widehat{F}(w|a)$ is always between 0 and 1 and monotonically increasing in $a$. This is particularly advantageous when inverting it to obtain the conditional quantiles estimates. However, the curse of dimensionality greatly limits the number of conditioning variables and thereby the extensibility of this measure.

We propose to define analogous measures of income inequality conditioning for a large set of income determinants. In general, if $F(y|\boldsymbol{x})$ is strictly increasing in $y$ given $\boldsymbol{X} = \boldsymbol{x}$ then the $p$th conditional quantile of $Y$ is the inverse of $F(y|\boldsymbol{x})$ and is defined as

$$Q^p(\boldsymbol{x}) = \{y \in \mathbb{R} : F(y|\boldsymbol{x}) = p\}.$$

Besides being easy to interpret and readily available in our framework, the advantage of using conditional based measures of income inequality lies in the robustness of the quantiles as they are not affected by extreme values in the tail of the distribution. Moreover, this particular choice of measures would allow, if examined jointly, to capture an important case of income polarization, usually referred to as the disappearing middle class (see, e.g., Jenkins, 1995b). The main disadvantage is that the chosen unconditional measure of inequality has no axiomatic base.

Based on the analogous unconditional quantile-based measures of inequality, we introduce one measure of conditional inequality in the central part of the income distribution and one of conditional inequality in the tail of the income distribution. We define the *Conditional Relative InterQuartile Range* (CRIQR), a measure of the dispersion in the central portion of the distribution of $Y$ given $\boldsymbol{X} = \boldsymbol{x}$ relative to the median, as

$$CRIQR(y|\boldsymbol{x}) = \frac{Q^{3/4}(\boldsymbol{x}) - Q^{1/4}(\boldsymbol{x})}{Q^{1/2}(\boldsymbol{x})}. \tag{4.2}$$

High figures of CRIQR indicate greater relative inequality. CRIQR can be estimated as follows. Notice that

$$F(y_j|\boldsymbol{x}) = \frac{\exp \boldsymbol{x}'\boldsymbol{\beta}_j}{1 + \exp \boldsymbol{x}'\boldsymbol{\beta}_j} = u_j,\ u_j \in [0, 1].$$

Then the conditional quantile function is defined by inverting the conditional dis-

tribution function as

$$F^{-1}\left(u_j|\boldsymbol{x}\right) = \ln\frac{u_j}{1-u_j} = Q^{u_j}\left(\boldsymbol{x}\right) = \boldsymbol{x}'\boldsymbol{\beta}_j.$$

Given the uniform convergence established in (4.1), conditional quantiles can be estimated by inverting the estimated conditional distribution function. It follows that

$$\sqrt{n}\left(\widehat{Q}_n^{u_j}\left(\boldsymbol{x}\right) - Q^{u_j}\left(\boldsymbol{x}\right)\right) \overset{a}{\sim} \mathrm{N}\left(0, \boldsymbol{x}'\mathscr{J}_j^{-1}\boldsymbol{x}\right). \tag{4.3}$$

By the delta method, it is possible to find the asymptotic variance of the estimator of the proposed inequality measure. In fact, for each $j, j', j'' = 1, \ldots, J$ and for $u_j, u_{j'}, u_{j''} \in [0,1]$, from (4.3) we have that

$$\begin{aligned}
&\operatorname{avar}\left(\frac{\widehat{Q}_n^{u_j}\left(\boldsymbol{x}\right) - \widehat{Q}_n^{u_{j'}}\left(\boldsymbol{x}\right)}{\widehat{Q}_n^{u_{j''}}\left(\boldsymbol{x}\right)}\right) \\
&= \frac{\left[Q^{u_{j''}}\left(\boldsymbol{x}\right)\right]^2 \operatorname{avar}\left(\widehat{Q}_n^{u_{j'}}\left(\boldsymbol{x}\right)\right) + \left[Q^{u_j}\left(\boldsymbol{x}\right) - Q^{u_{j'}}\left(\boldsymbol{x}\right)\right]^2 \operatorname{avar}\left(\widehat{Q}_n^{u_{j''}}\left(\boldsymbol{x}\right)\right)}{\left[Q^{u_{j''}}\left(\boldsymbol{x}\right)\right]^4} \\
&+ \frac{\left[Q^{u_{j''}}\left(\boldsymbol{x}\right)\right]^2 \operatorname{avar}\left(\widehat{Q}_n^{u_j}\left(\boldsymbol{x}\right)\right)}{\left[Q^{u_{j''}}\left(\boldsymbol{x}\right)\right]^4},
\end{aligned}$$

where $\operatorname{avar}(\cdot)$ signifies asymptotic variance.

We also define the *Conditional Decile Dispersion Ratio* (CDDR) as

$$CDDR(y|\boldsymbol{x}) = \frac{Q^{9/10}(\boldsymbol{x})}{Q^{1/10}(\boldsymbol{x})}.$$

The CDDR expresses the income of the top decile of the income distribution (the "rich") as a multiple of that of those in the bottom decile (the "poor"), given $\boldsymbol{X} = \boldsymbol{x}$. High figures indicate greater inequality in the tail of the distribution of income. Similarly to before

$$\operatorname{avar}\left(\frac{\widehat{Q}_n^{u_j}\left(\boldsymbol{x}\right)}{\widehat{Q}_n^{u_{j''}}\left(\boldsymbol{x}\right)}\right) = \frac{\left[Q^{u_j}\left(\boldsymbol{x}\right)\right]^2 \operatorname{avar}\left(\widehat{Q}_n^{u_{j''}}\left(\boldsymbol{x}\right)\right) + \left[Q^{u_{j''}}\left(\boldsymbol{x}\right)\right]^2 \operatorname{avar}\left(\widehat{Q}_n^{u_j}\left(\boldsymbol{x}\right)\right)}{\left[Q^{u_{j''}}\left(\boldsymbol{x}\right)\right]^4}.$$

Consistency of the estimators of the proposed measures follows directly by unbiasedness and the variance tending asymptotically to zero.

In the next section the results of the estimation procedure will be presented

and analyzed.

## 4.4 Estimation Results

In the following subsection we report the results of the estimation procedures outlined in the previous section and the estimated conditional inequality measures.

The results shown in this section refer to the year 1991. Data from the Central Statistical Office show that the year 1991 represents a highwater mark of income inequality in the UK (Atkinson, 1997).[9] Results for the following years display a qualitatively very close behaviour and therefore are omitted for space reasons.

A parametric implementation based on quantile regression that broadly supports our findings is presented in Appendix D on page 236.

### 4.4.1 Conditional Distribution Estimates

Each panel in Figure 4.4 graphs the estimated conditional distribution of income against each predictor (keeping the others constant).[10] To aid interpretation income is graphed on a log scale. Also, the range of the income axis has been chosen so as to aid comparison of quantitative impacts across graphs. Each panel in the Figures is constructed by first evaluating the estimated functions $\widehat{F}(y_j|\boldsymbol{x})$, over a grid of 200 equally spaced points between the 0th and 100th percentile of each explanatory variable (keeping the other constant at their mean value) and then plotting the iso-probability contours. With the aid of these contours it is possible to clearly appreciate non-linearities and higher order relations, such as inequality changes, in the conditional distribution of income. In fact, the iso-probability contours can be viewed as a powerful generalization of the conditional mean and median that are conventionally employed in econometric inference.[11] Each iso-

[9]Brewer et al. (2005) using more recent data point out that changes in income distribution before 1991 were very different from changes that occurred in later years. They show that over the period 1979 to 1990, the increase in inequality was determined by higher income growth of wealthier households.

[10]The estimated logit coefficients used to construct the conditional distribution, for the chosen evaluation points, are shown in Appendix E.

[11]Though conditional mean and median are both measures of central tendency, they do not in general agree. As Manski (1988) points out, one might be a linear function of the covariates and the other not, both might be linear but with different, even of opposite sign, coefficients.

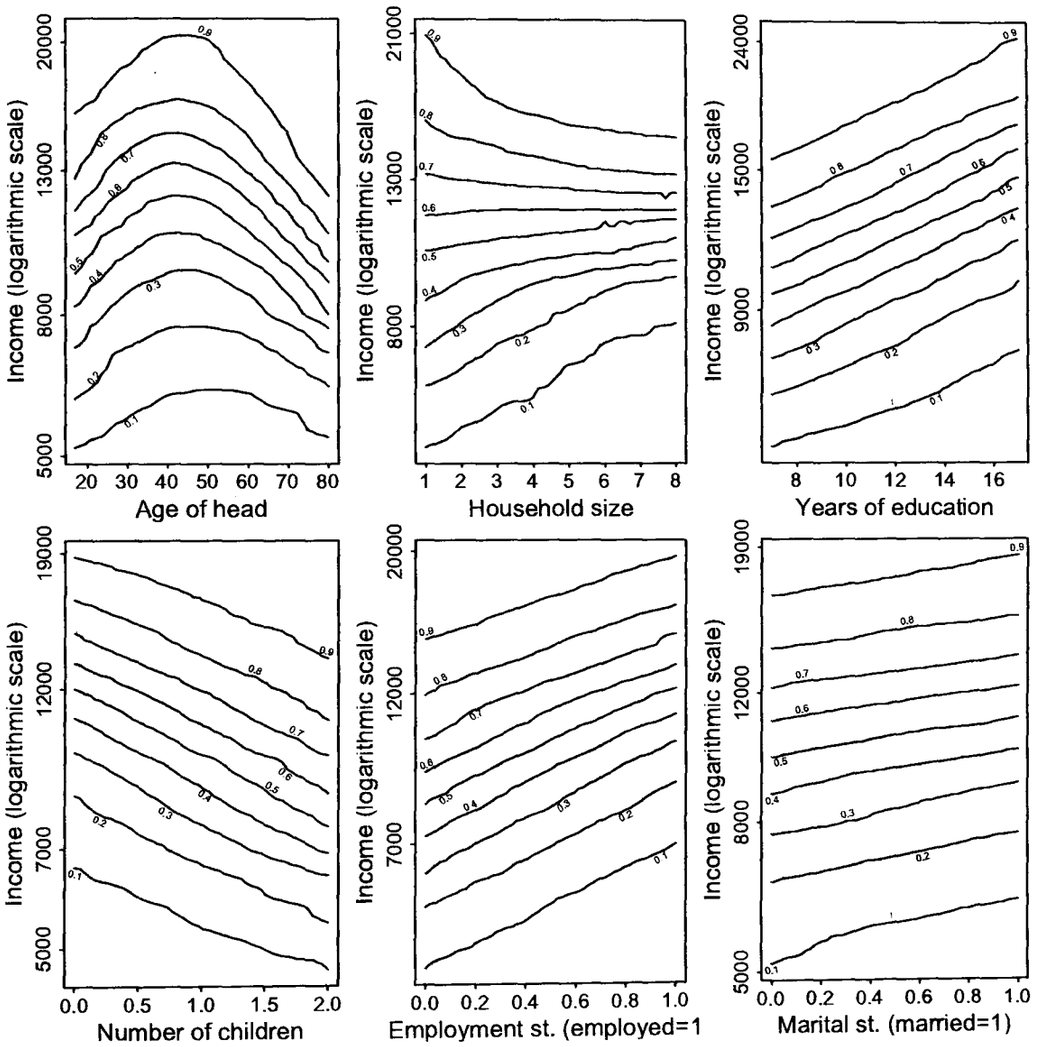
probability contour can be interpreted as a regression curve, corresponding to a particular percentage points of the conditional distribution of income. For example the 0.5 iso-probability curve represents the more "traditional" conditional median regression, i.e. it describes the behavior of the conditional median of income as one explanatory variable changes while the others remain constant. Positively sloped iso-probability curves are indicative of a positive relationship between the explanatory variable and the corresponding conditional quantile. A horizontal contour signifies that the explanatory variable does not appreciably influence any shape characteristics of the conditional distribution. The percentile points for 0-1 dummy variables, such as employment an marital status, are shown for convenience.

Similarly to Foresi & Peracchi (1995), a violation of the monotonicity constraint on the conditional distribution function can be readily spotted when any vertical line crosses the conditional quantile one in more than one point. In the present case it does not seem to be a problem In general, all conditional relationships are poorly determined at the extreme of the range, so that care has to be taken when interpreting those values.

For the sake of parsimony, the results for the categorical variables Gender, Urban/Rural, and Retired/Working are not discussed in the following paragraphs, as they were found to be not statistically significant.[12]

[12]See the coefficient estimates in Appendix E.

Figure 4.4: Iso-probability contours of the estimated conditional distribution function of income



**Age-income profiles** The top-left panel of Figure 4.4 graphs the estimated quantiles of log income conditional on the age of the head. The results are consistent with the observable implications of the life-cycle hypothesis. The age-income profile, keeping all other factors constant, is clearly hump-shaped. Also, income of families during the period of child rearing is higher than income in the retirement stage of the life-cycle, when economic responsibility is greatly reduced. At the early stages of the life-cycle, household's median income is just above £9,200. The profile for the conditional median peaks at around 43 years of age, with an income value of about £11,950, and declines afterwards, eventually reaching an income of about £8,000 in the later stages of the life-cycle. These results agree with previous empirical findings (see, e.g., Jappelli & Modigliani, 2005).

Our approach reveals also nonlinearities in the age-income profiles not realized in previous studies. We find that extreme order quantiles peak later. The conditional lower-decile peaks at the age of about 52 and appears relatively flat, and the upper-decile peaks earlier, at the age of 45, and appears considerably non-linear. The age-income profile for the richest families in the sample, represented by the conditional upper-decile, has a value of about £15,700 for younger households, reaches about £20,500 at its peak and declines sharply after retirement, reaching a value of slightly more than £11,900 at the age of 80. This non-linearity is unlikely to be captured adequately by parametric methods, unless some ad-hoc assumptions are made. Since we control for retirement, this dramatic fall is most likely induced by the decreased earning capacity of older heads due to, among other things, worsening health conditions and changing attitude towards risk. The age-income profile for the poorest groups in the sample, represented by the conditional lower-decile, appears relatively flat, peaking at around the age of 52, where it reaches an income of about £6,300, just £1,000 more than at the early stages of the life-cycle, and declines to just above £5,300. Besides non-linearities, the conditional quantiles show that age also provides information about higher properties of the conditional distribution of income. The spread of the conditional distribution seems to increase at first, till about 40–50 years of age, and declines dramatically afterwards. This pattern of inequality is due to the greater arching of the age-income profiles of the richer families. This implies that the impact on inequality is much greater for the high income households.

**Education-income profiles** The top-right panel of Figure 4.4 graphs the estimated quantiles of log income conditional on years of education. Education seems to convey information mostly about location. In fact, parallel conditional quantile lines imply just a location shift, while the shape of the distribution remains the same. A more careful look reveals that the middle of the distribution seems to become less spread whereas the tails seem to diverge, as the years of education increase. Also, the impact of education seems slightly greater for lower income families.

**Household size-income profiles** The top-center panel of Figure 4.6 graphs the estimated quantiles of log income conditional on household size. The profiles for the conditional quantiles appear to be non-linear in household size. In particular, the conditional quantiles increase at decreasing rates. Household size conveys considerable information about the spread of the distribution as well. Conditional quantiles appear to be getting tighter as the numbers in the household increase. Because of the changes in slope, the decrease in spread, though substantial, is difficult to assess visually. Figure 4.5 shows the difference between the 0.9 and the 0.1 conditional quantiles.[13] The difference is graphed on percentage change scale.[14] The graph clearly shows that the change is economically substantial. The top 10 per cent of the households earn almost 300 per cent more than the bottom 10 per cent with a family size of two, and about 120 per cent more with a family of four.

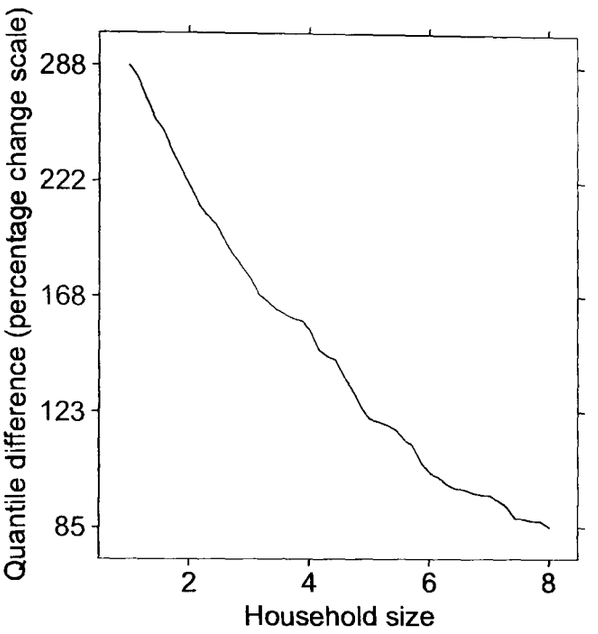
**Number of young children-income profiles** The bottom-left panel of Figure 4.6 graphs the estimated quantiles of income conditional on the number of young children present in the household.

The conditional quantiles of log income are decreasing with the number of young children. The conditional upper-decile has an income of about 18,750 for childless households and decreases by about 30 per cent (about 5,445), to 13,300 for

[13]The line segments of the graph are banked to 15°, i.e., the aspect ratio of the display is chosen such that the absolute values of the orientations of the segments constituting the curve are centered on 15°, which allows a much clearer decoding of visual information.

[14]This involves just a minor adjustment, as log differences can already be interpreted approximately in terms of percentage changes.

Figure 4.5: Difference between the 0.9 and the 0.1 conditional quantiles




households with two children.[15] The fall for the lower-decile over the same range is very similar. Income for childless households is about 6,560, and decreases by about 30 per cent (about 1,920), to reach 4,650.

The spread of the conditional distribution seems to vary with the number of children in a non-linear fashion. The upper quantiles are convex, implying that decreases in income are increasing with the number of young children, whereas the lower deciles are convex, so that decreases slacken with the number of young children.[16]

**Employment status-income profiles** The bottom-center panel of Figure 4.6 graphs the estimated quantiles of log income conditional on the employment status of the head. Values between zero and one are computed and displayed as a continuous curve to simplify comparisons and the interpretation. Households with

[15]Less than 6 per cent of the households in the sample have more than 2 children and only 1 per cent more than three.

[16]A positively sloped convex line on a logarithmic scale shows that the rates of increase are increasing, while a concave one shows decreasing changes. The logarithmic scale is computed on base $e$ here.

employed heads have higher conditional quantiles. The impact of employment status is spread over the entire income distribution. This finding agrees with the results obtained by Nolan (1988-89) using 1977 FES data in his analysis of the impact of UK economic conditions on income inequality. We find that the impact is almost two times larger for the lower income families. The conditional upper-decile has an income of about £14,600 for unemployed households and increases by about 34 per cent, to £19,600 for employed households. The increase for the lower-decile is much sharper. Income for households with unemployed heads is about £4,500, and increases by more than 55 per cent, to reach £6,980 for households with employed heads. This differential impact implies that inequality at the extremes of the distribution is higher for households with unemployed heads.

**Marital status-income profiles** The bottom-right panel of Figure 4.6 graphs the estimated quantiles of log income conditional on the marital status of the household's head. Values between zero and one are computed and displayed as a continuous curve to simplify comparisons and the interpretation. Households with married heads have higher conditional quantiles. The impact of marriage is spread over the entire income distribution. We find that the impact is almost two times larger for the higher income families. The conditional upper-decile has an income of about £16,340 for unmarried household heads and increases by about 22 per cent, to £18,500 for married household heads. The increase for the lower-decile is much lower. Income for households with unmarried heads is about £5,125, and increases by more than 13 per cent, to reach £6,250 for households with married heads. This differential impact implies that inequality at the extremes of the distribution is higher for households with married heads.

### 4.4.2 Conditional Inequality Measures Estimates

Estimates of the CRIQR and the CDDR inequality measures are presented in Figure 4.7. Once the estimate of the conditional distribution is obtained, $\widehat{F}_n(y_j|\boldsymbol{x})$, the $p$th conditional quantile can be obtained numerically as the root of the equation

$$\widehat{F}_n(y_j|\boldsymbol{x}) - p = 0 \qquad \text{with } 0 < p < 1.^{17} \tag{4.4}$$

[17]We used Brent's method, which combines an interpolation strategy with the bisection algorithm, to obtain the conditional inverses. This root finding method has the disadvantage that it can only search for one root at a time. Multiple roots, or very close roots,

Figure 4.6 shows the estimated conditional quantiles and asymptotic confidence intervals for a set of relevant determinants used to derive the inequality measures.

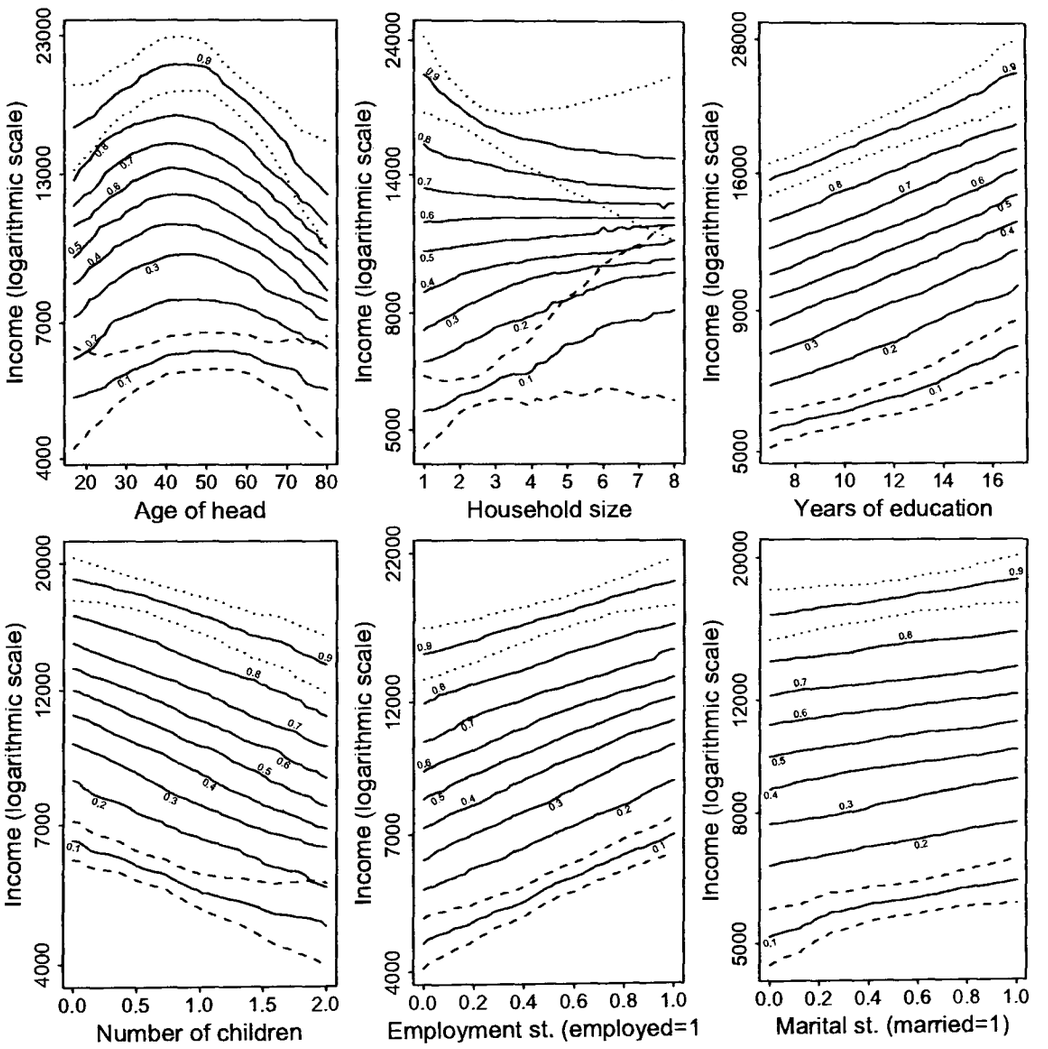
We used Brent's method, which combines an interpolation strategy with the bisection algorithm, to obtain the conditionalinverses.[18]

Each panel in Figures 4.7 and 4.8 graphs the estimated conditional inequality measures against each predictor (keeping the others constant). Some interesting features are highlighted in the following paragraphs.

are a problem, not only from a theoretical point of view, since they represent violation of the monotonicity assumption of the conditional quantiles, but also from a numerical point of view. This is true especially with roots of order 2 ("turns" of the conditional quantiles). In that case, there will be no readily apparent sign change in the function, so that bracketing a root becomes impossible. In the case of more than one root, only the first root to be found will be returned. Obviously this could make computing and interpretation of the conditional quantiles and derived measures more problematic. From the analysis of the isoprobability curves we already know that this does not appear to be a conspicuous problem.

[18]The method approximates the function using an interpolating quadratic curve. Whenever the zero of the interpolating curve falls outside the bracketing interval (the starting interval containing the zero), the algorithm falls back to an ordinary bisection step.

Figure 4.6: Estimated conditional deciles of the conditional distribution function of log income with one standard deviation confidence interval shown for the lower (dashed lines) and upper (dotted lines) deciles



**Age-inequality profiles** The two panels in the first column of Figure 4.7 display the estimated age-inequality profiles. The CDDR conditional age-inequality profile, holding all other variables constant at their mean value, is hump-shaped with the declining branch of the profile much longer and steeper than the ascending one.

The profile for the inequality in the center of the income distribution is overall increasing and has a slightly convex shape. The Age-inequality profile, for changes in the tail of the income distribution (CDDR), increases by more than 9 cent over the 20–40 range, flattens out, and falls dramatically for households with a head aged more than 50, decreasing by about 30 per cent over the 50–80 age range. Our estimates indicate that, after the age of 60, households in the top 90th percentile of the income distribution move from earning about 3.2 times more than households in the 10th percentile to earning less than 2.3 more. Inequality in the middle of the distribution (CRIQR) decreases by about 32 per cent over the 20–80 age range.

In his study on inequality trends in the UK, Jenkins (1995a), using indices of inequality decomposed by population sub-groups with FES data, found an analogous pattern of declining inequality for elderly households.

These changes in inequality, as we are controlling for many factors, are most likely induced by the decreased earning capacity of older heads probably due to, among other things, worsening health conditions and lower attitude towards risk.

**Household size-inequality profiles** The two panels in the middle column of Figure 4.7 display the estimated household size-inequality profiles. As expected, household size has a stabilizing effect on income. Over the household size range inequality in the tails of the distribution decreases substantially. The top 90th percentile of the income distribution move from earning about 3.9 times more then households in the 10th percentile to earning about 2 times more. This amount to a fall of about 50 per cent.

In particular, the results for inequality in the tails of the distribution show that if, for instance, household size increases from 2 to 4, households in the top 90th percentile of the income distribution move from earning 3.2 times more then households in the 10th percentile to earning about 2.5 times more. This amounts to a 20 per cent fall in inequality. For the same variation in household size, inequality in the middle of the distribution increases by about 30 per cent. Over the whole range, 1 to 6, the fall is about 54 per cent.

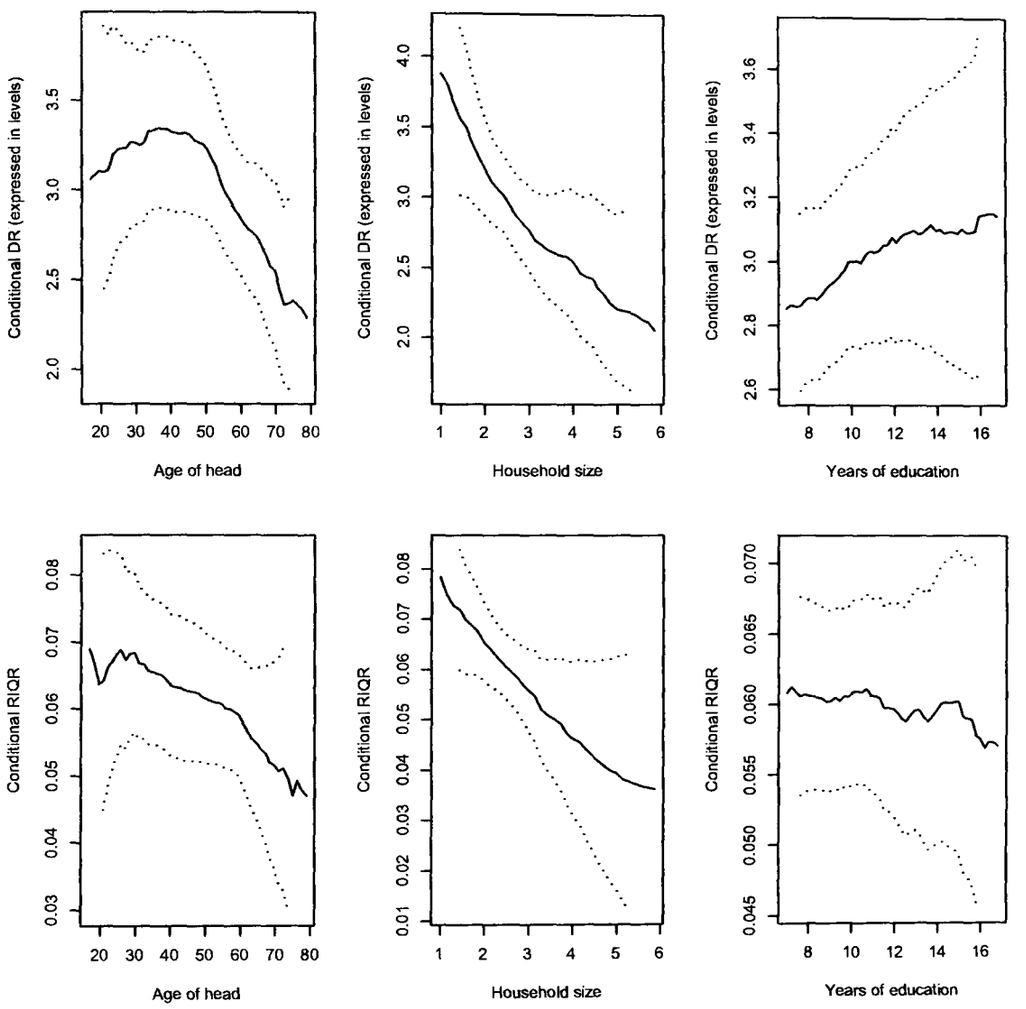
As average household size has been falling in the past decades, these results seem to be able to explain the increase in inequality in the UK income distribution. These results are consistent with previous empirical literature. Semple (1975), for instance, using FES data, found that inequality, expressed in terms of the ratio of highest and lowest quintile income, respectively, to median income, is reduced if the effect of falling household size is controlled for.

**Years of education-inequality profiles** The two panels in the last column of Figure 4.7 display the estimated years of education-inequality profiles. The inequality measures for education are more difficult to interpret as the effects are both economically and statistically small. The CDDR appears to be increasing non-linearly with education. Though inequality in the tail of the distribution increases with education it increases at increasing rates over the compulsory education range, where public education is virtually free, and increases at decreasing rates afterwards. Our estimates indicate that in the compulsory education range, households in the top 90th percentile of the income distribution move from earning about 2.85 times more than households in the 10th percentile to earning about 3.08 more, an increase of 8 per cent. After that, richer households go from earning 3.08 more to earning 3.14 more than poorer ones, an increase of only 2 per cent. This is consistent with a liquidity constraint explanation: access to education is impeded by the lack of financial resources. This interpretation is corroborated by the pattern of inequality in the center of the distribution, where we would expect liquidity constraints to be less binding. The CRIQR on years of education has a downward trend. Inequality in the middle of the distribution decreases by about 6 per cent over the education range. An inverted-U shaped relationship between income inequality and educational achievements is found by Checchi (2001) (see also references therein) using a cross-country panel data approach.

**Number of children-inequality profiles** The two panels in the first column of Figure 4.8 display the estimated number of young children-inequality profiles. Both the conditional DR and RIQR number of children-inequality profile, holding all other variables constant at their mean value, appear to be hump-shaped.

Both graphs show an initial large increase of inequality if the household composition changes from no children to one child. Our estimates indicate that house-

Figure 4.7: Conditional measures of income inequality on age of head, household size, and years of education with one standard deviation confidence intervals




holds with one young child in the top 90th percentile of the income distribution earn about 3 times more than households in the 10th percentile while households with no children earn about 2.85 times more, an increase of about 15 per cent. The increase is about 6 per cent for the central part of the income distribution.

Because of low sample information for households with more than 2 children, both the economic and statistical impact, respectively because of numerical instabilities and the large standard deviation, cannot be reliably determined for larger values of the predictor.

**Employment status-inequality profiles** The two panels in the middle column of Figure 4.8 display the estimated employment status-inequality profiles. As expected, employment has a negative impact on inequality.

Though inequality in the tail of the distribution increases with education it increases at increasing rates over the compulsory education range, where public education is virtually free, and increases at decreasing rates afterwards. Our estimates indicate that households with unemployed head in the top 90th percentile of the income distribution earn about 3.25 times more than households in the 10th percentile while households with employed heads earn about 2.8 times more, an decrease of about 13 per cent.

The decrease is about 23 per cent for the central part of the income distribution.

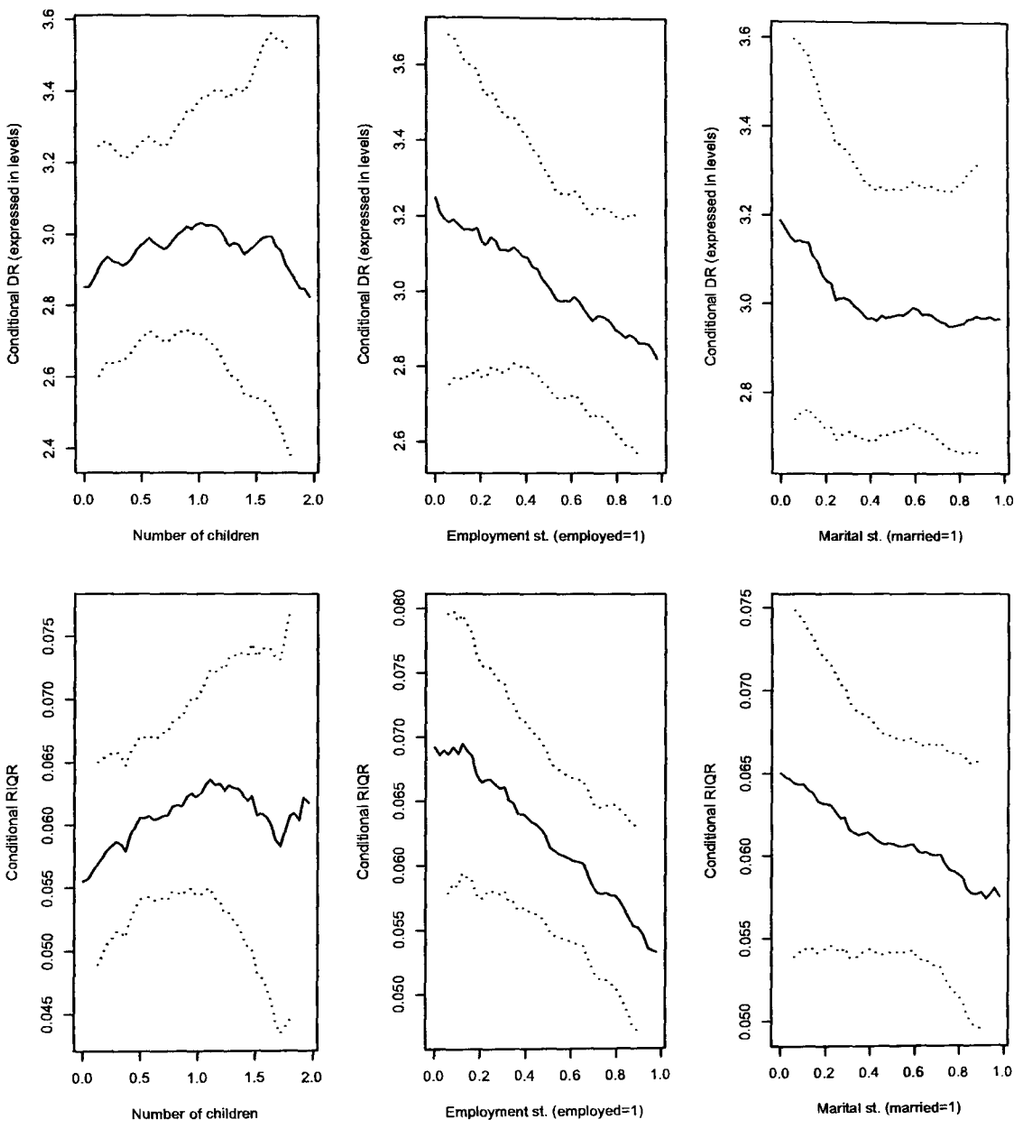
**Marital status-inequality profiles** The two panels in the last column of Figure 4.8 display the estimated marital status-inequality profiles. The inequality measures for marital status shown in the two panels of the last column of Figure 4.8 are difficult to interpret as the effects are both economically and statistically small. The extent of marriage seems to have a slightly stabilizing effect, particularly in the middle of the distribution.

The decrease for the central part of the income distribution is about 11.6 per cent. This conclusion has precedents in the empirical literature. For instance, Dinwiddy & Reed (1977), examining four factors separately, also reached the same conclusions about the impact of marital status on income inequality in the UK.

## 4.5 Conclusion and Future Research Directions

The purpose of this paper has been to analyze the impact of demographic and social factors on the conditional distribution of household income for the UK, and in particular on their impact on income inequality. We started by estimating the conditional distribution of income over a broad set of determinants. We then devised a method for obtaining conditional inequality measures by inverting the estimated conditional distribution. Our results provide a visually clear representation of both

Figure 4.8: Conditional measures of income inequality on number of children, employment and marital status, keeping all other determinants fixed at their respective mean values




the substantive and statistical impact of each factor on income inequality, keeping all others constant.

For instance, we find that for the UK sample, household size has a large substantive and statistical impact on inequality. Combined with the recent trend of declining household size in the UK, this result can help explain the trend of

increasing income inequality observed in the past decades in the UK.

The following research directions would seem appropriate to improve and extend the chapter.

- Extend the approach to make use of the panel nature of the data. Though preliminary analysis did not show any significant change in the results, a panel approach would allow to track households over time and to model age and cohort effects.

  Ignoring cohort effects produces age-income profiles that could be biased. age-income profiles can vary across cohorts, particularly for cohorts that are distant in time.
- The estimation method assumes that regressors are exogenous. This can be certainly argued for age and possibly education. However, household income is an important determinant of the decision to have children, household formation, marriage, household dissolution, retirement to some extent, and so on. Some other econometric approach, such as instrumental variables, could be explored to obtain improved estimates.
- There are several interesting hypothesis that emerge from this study such as the possible effect of liquidity constraints on education and the possibility that the impact of worsening health condition or and changing attitudes toward risk. It would be interesting to extend the paper to formally test these hypotheses.
- Based on the parametric conditional quantile regression approach presented in Appendix D on page 236, it would seem more appropriate to use a non-linear quantile regression approach (see, e.g., Busovaca, 1985, and references therein) for a more fruitful comparison.

Chapter 5

# Economic Growth, Trade, and the Environment: An Endogenous Determination of Multiple Cross-Country Regimes

## 5.1 Introduction

While developed countries are responsible for most of the increase in greenhouse gas emissions to date, greenhouse gas emissions from developing countries are expected to expand significantly (World Bank, 1992). There is a growing concern that should developing countries follow the same development path of currently developed countries there could be catastrophic consequences for the environment.

The relationship between economic development and the environment has been explored by many authors, starting with Grossman & Krueger (1993a), with their study into the environmental implications of the North American Free Trade Agreement (NAFTA). The Environmental Kuznets Curve (EKC) hypothesis envisages an inverted-U-shaped relationship between income and environmental degradation. According to this hypothesis, pollution rises with income as long as income is relatively low and starts declining once income has exceeded a threshold level, known as income turning point (ITP).

To test this hypothesis, typically the natural logarithm of an indicator of environmental quality is assumed to depend on two sets of variables. One set of variables consists of a polynomial in the natural logarithm of *per capita* income. The second set of variables consists of control variables that correspond to additional determinants of environmental quality proposed by researchers.

Dozens of additional variables have appeared in the literature, despite the fact that fewer than 100 countries are available for analysis in a typical data set (for a survey see, e.g., Panayotou, 2000). Data limitations relatively to the abundance of theories has resulted in a large number of non-nested relationships that seem to support various and alternative theories. The list of control variables used in the literature on the EKC includes, industrial composition of output (see, e.g., Grossman & Krueger, 1995), population density (see e.g., Cropper & Griffiths, 1994; Selden & Song, 1994), openness to trade (see, e.g., Antweiler et al., 2001; Hettige et al., 1992; Grossman & Krueger, 1993b; Suri & Chapman, 1998), environmental regulation and control (see, e.g., Shafik, 1994a; Baldwin, 1995), democracy (see, e.g., Torras & J.K., 1998; Harbaugh et al., 2002), corruption (see, e.g., Lopez & Mitra, 2000), civil and political liberties (see, e.g., Barrett & Graddy, 2000; Torras & J.K., 1998), power inequality (see, e.g., Boyce, 1994), literacy (see, e.g., Torras & J.K., 1998), geographical factors (see, e.g., Neumayer, 2002), income inequality (see, e.g., Torras & J.K., 1998; Magnani, 2000; Ravallion et al., 2000),

and so on. It appears that, given the vast number of proposed variables affecting environmental quality, any parsimonious regression will necessarily leave out many factors that would be likely to bias the estimated parameters of the included regressors.

Regression analysis on cross-section data has shown that some pollutant increase until they reach an income *per capita* approximately between $5,000 and $8,000. However, it has been noted in the literature that the shape of the estimated EKCs differs widely according to the sample of countries included, the time span of the sample, the pollutant, the data used, etc. For instance, Cavlovic et al. (2001), meta-analysis to investigate systematic variation across Environmental Kuznets Curve studies, showed that EKC relationships and their associated income turning points depend on the scale of analysis and the type of pollutants. Harbaugh et al. (2002) re-examined the empirical evidence for the EKC for three local pollutants, i.e., sulfur dioxide, smoke, and total suspended particles (TSP) using a more representative data set. Harbaugh et al. (2002) are unable to find support for an EKC using Grossman & Krueger's specification with an updated version of the data. They also found that the estimates are extremely sensitive to the sample chosen and the econometric specification. Extensive literature reviews by Barbier (1997), Panayotou (2000), and Stern (2004), found considerable variability in the estimated results across types of environmental quality indicators and samples. Li et al. (2007), in a more recent meta-analysis that included about three times more studies than Cavlovic et al. (2001), found that data characteristics, econometric methods, and the chosen measure of environmental degradation, all considerably affect the existence of an EKC and the location of a predicted turning point.

One important assumption underlying the majority of cross-country pollution studies is that all countries obey a common linear model specification. However, there is increasing evidence coming from theoretical and empirical literature of heterogeneity problems. Studies such as Brock & Taylor (2004) and Dijkgraaf & Vollebergh (2005) have illustrated, albeit in very different ways, that the constant coefficient linear model assumptions made in standard EKC analyses are not supported by the data. Dijkgraaf & Vollebergh (2005) contrasted time-series against panel estimates for $CO_2$ emission in a sample of OECD countries. They found that combining different emissions-income relationship into a panel distorts estimates. Brock & Taylor (2004) demonstrated that the relationship between income and the environment can be exceedingly complex. They argued that income-emissions

profiles are likely to differ across countries depending on initial conditions or the level of other structural parameters such as savings, technological change, and population growth rates.

Regional studies, such as Carson et al. (1997b), Vincent (1997), and de Bruyn et al. (1998) also provide evidence of the importance of heterogeneity. Vincent (1997) exposed the limitations of previous cross-section studies by comparing Malaysia's actual pollution trends with those that would be predicted by Selden & Song's (1994) estimates. It was found that their forecasts overestimated emission levels for particulate, $NO_X$ and $CO_2$ whilst even the direction of $SO_2$ emissions was incorrect. Moreover, it was illustrated that if one took the "one size fits all" argument, underlying the EKC approach, through with Selden & Song's findings, one would expect Malaysia to be on the upward portion of the EKC, given the turning point found by Selden & Song of \$8,079 *per capita* (PPP) and Malaysia's GDP *per capita* in 1987 was only \$4,727 *per capita* (PPP). Nevertheless, from 1987–1991 Malaysia witnessed a drop in $SO_2$ emissions in contrast to the EKC predictions due to an unobserved shock, "geology and a desire for energy interdependence, not rising income ... were responsible for the decline in $SO_2$ emissions" (Vincent, 1997). Vincent therefore concluded that although his study did not refute the existence of the EKC in some nations, "policymakers in developing countries should not assume that economic growth will automatically solve air and water pollution problems" (Vincent, 1997).

Carson et al. (1997b) using US time-series spanning the years 1988–1994, found a negative relationship between seven types of pollutants and income. Their general findings were consistent with the EKC hypothesis since a negative relationship between emissions *per capita* and income per capita for the seven pollutants examined was found. However, surprisingly, high income states had low *per capita* emissions and *vice versa* for low-income states. They suggested that it is more difficult to forecast emission levels for countries, which are about to approach the apparent turning point thus more research should be directed into this "greater variability in *per capita* emissions in lower income jurisdictions than in higher income political jurisdictions" and that this may lead to a "better understanding of what factors lie behind the cross-sectional EKC" (Carson et al., 1997b). This interpretation is strengthened by Vincent's analysis on Malaysia.

de Bruyn et al. (1998) found, by estimating regional time series model individually for the Netherlands, Germany, the United Kingdom, and the USA, that

economic growth increases emissions of carbon dioxide, nitrogen oxides, and sulfur dioxide. They also highlighted the importance of structural changes within countries and argued that conventional cross-section based techniques have produced spurious results by neglecting important dynamic processes.

In this Chapter we re–examine the relationship between economic activity and the environment, in order to identify the presence of multiple regimes using a threshold estimation approach based on Breiman, Friedman, Olshen & Stone's regression-tree (1984). The EKC has led in some cases to unwarranted and misleading interpretations that countries can overcome their environmental problems in the long run without consciously adopting environmental policies (see, e.g., Beckerman, 1992). However, increasingly, it has been recognized that the effect of such changes on environmentincome links are not exogenous processes but influenced by policy choices (see, e.g., Panayotou, 1995; Stern, 1996; World Bank, 1992). In particular, the World Bank's World Development Report 1992, focusing on environmental issues, observed that for most air and water pollution, environmental problems "initially worsen but then improve as incomes rise," and stated that "There is nothing automatic about this improvement; it occurs only when countries deliberately introduce policies to ensure that additional resources are devoted to dealing with environmental problems (World Bank, 1992, p. 10). Our approach permits in principle to specify better econometric models and to avoid the dangers of misinterpretation by acknowledging that the relationship between economic development and the environment is affected by structural differences across heterogeneous countries. Understanding regime differences in the relationship between economic growth and the environment, is the first step in bringing about more desirable outcomes through active policy interventions.

In this chapter, we first identify the presence of multiple regimes by using specification tests which entertain a single regime model as the null hypothesis. Then we develop an easily interpretable measure, based on an application of the Blinder-Oaxaca decomposition (Oaxaca, 1973; Blinder, 1973), of the impact on the environment due to differences in regimes. Finally we apply a recursive partitioning algorithm (regression tree) to endogenously identify the separate regimes.

Our conclusions are threefold. First, we reject the null hypothesis that all countries obey a common linear model. Second, we find that quantitatively regime differences can have a significant quantitative impact. Thirdly, by using regression tree analysis we find subsets of countries which appear to possess very different

environmental/economic relationships.

The chapter is organized as follows. Section 5.2 introduces the econometric and theoretical arguments that support the threshold approach. Section 5.4 reviews the links between trade and environment. Section 5.5 derives the parameter heterogeneity implications of the theoretical models. Section 5.6 surveys the ways parameter heterogeneity has been accounted for in empirical models. Section 5.7 describes the data used in this study. In Section 5.8 we attempt to identify the existence of multiple regimes in the data by means of specification tests. Section 5.9 presents an easily interpretable measure, based on an application of the Blinder-Oaxaca decomposition, (Oaxaca, 1973; Blinder, 1973) of the impact on the environment due to differences in regimes. In Section 5.11 the threshold estimation methodology based on tree regressions is presented. Section 5.12 uses regression tree techniques to identify groups of countries obeying common linear model. Section 5.13 concludes.

## 5.2 Environmental-Economic Regimes

The processes of economic growth and environmental change are clearly complex and evolving over time. Identifying all the complex interactions and feedback relationships that are expected to play a significant role in the evolution of these processes may be an impossible task at this point in time. One important assumption underlying the majority of cross-country pollution studies is that all countries obey a common linear model specification. Because of the inherent complexity of the environment-economy interaction, our limited knowledge of it, and the often poor quality of data, this assumption appears at best as a crude approximation. Limits in our econometric models can reveal themselves as apparent structural change. Identifying these structural changes could further our understanding of the links between the economy and the environment.

Besides econometric arguments, recent theoretical developments in modeling the relationship between income and the environment also imply the existence of different regimes. A simple and frequently used explanation for the EKC is based on a traditional demand-and-supply analysis. A possible way to obtain an inverted-U shaped EKC consistent with a demand-and-supply framework is to suggest that the EKC reflects a demand for environmental quality. Assuming that environmental quality is a normal good, pollution may at first rise with income, but

will eventually fall as income continues to rise. More formal developments can be found in Lopez (1994) and Copeland & Taylor (2003). The resulting smooth EKC from these models is graphed in Figure 6.1. Other models based on traditional economic theory, such as the one by Andreoni & Levinson (2001), also predicts a smooth EKC curve for a technology with increasing returns to scale.

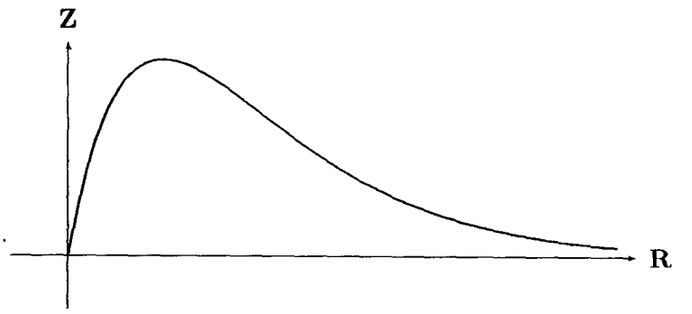



Figure 5.1: EKC generated by income effects

Several recent papers have attempted to explain the EKC relationship by introducing threshold effects in modeling either pollution abatement. (see. e.g.. Jones & Manuelli, 1995), or environmental policy regulation (see, e.g., Stokey, 2001). Threshold effects lead to a very different relationship between environmental quality and income during early stages of economic development as opposed to later stages. For instance the abatement-threshold model predicts a kink in the relationship between pollution and income, as shown in Figure 6.2. The policy threshold model predicts an even more drastic change in regimes, and produces a discontinuous EKC with a discrete drop in pollution and income once the threshold is reached.

The policy threshold models assume that governments do not adopt environmental policy regulations until income surpasses a threshold level. In this eventuality, regime differences could manifest themselves in parameter changes in the estimated basic EKC regression model. In the basic EKC equation, income, and powers of it, could serve as proxies for different sets of variables for different subsets of countries subject to different regimes. To estimate and test this class of models a simple linear specification is obviously not appropriate. Methods that take into account parameter heterogeneity have to be employed instead.

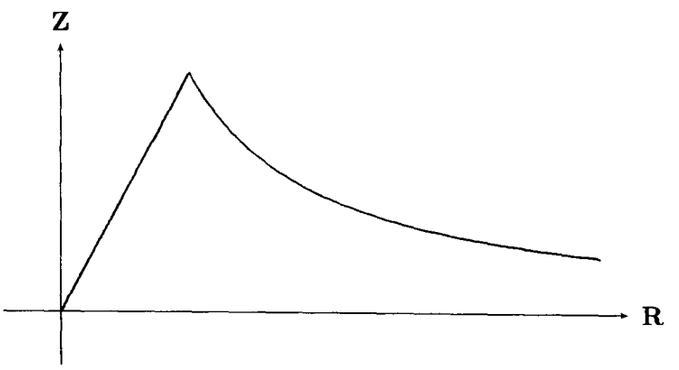



Figure 5.2: EKC generated by threshold effects model

## 5.3 The Impact of Trade on the Environment

The heterogeneity issue seems particularly important in the study of the relationship between trade and the environment. For instance, with regard to the relationship between, economic growth, trade policies, and the environment economic theory suggests that an increased openness to foreign markets might have a different impact on the environment in developed and developing countries. Grossman & Krueger (1991), identify three possible mechanisms by which trade and foreign investment policies can impact pollution.

(i) Scale effect. Trade and foreign investment liberalization determine an expansion of economic activity and therefore increase pollution. For instance, if economic growth is fueled by an increase in the demand of energy, which if satisfied using the pre-existing methods determines an expansion in the emission of harmful pollutants.

(ii) Composition effect. Trade liberalization should encourage countries to specialize in the production of goods in which they enjoy a competitive advantage.

(iii) Technique effect. Freer, trade and foreign investment might also impact production methods. Pollution intensity of production might fall because of the transfer environmentally friendlier technologies of production. This effect has become also known as the gain from trade hypothesis.

This composition effect, for instance, gives rise to several competitive hypothesis with regard to the impact of trade and foreign investment on the environment.

If the comparative advantage enjoyed by a country in the production of goods originates from differences in environmental regulation, then the composition effect of engaging in more international trade could result in a deterioration of the environment.

According to the pollution haven hypothesis (PHH), income differences between countries translate into differences in the strictness of environmental regulations. The premise of this hypothesis is that high-income countries tend to demand cleaner environments. To satisfy this demand, governments attempt to enforce more stringent regulation regimes over the domestic industry, and allow importing pollution-intensive goods from less regulated countries. Assuming that production costs positively related to the level of regulation, low-income countries have a comparative advantage in the production of pollution-intensive goods compared to high-income countries. There will be the tendency for dirty industries to relocate to low-income countries as a result of international trade.

The race to the bottom hypothesis (RTB) argues that, given a level of income *per capita*, the more a country opens to international trade, the laxer regulation on the environment it will adopt in order to gain international competitiveness. This hypothesis presupposes that pollution abatement costs are an important component of an enterprise's investment decision, so the countries will compete to lower the environmental standards in order to gain its comparative advantage.

On the other hand, if the sources of the comparative advantage stems from the more traditional differences in factor abundance and technology, then the impact of freer trade will depend on the degree of pollution-intensity of production. This classical argument gives rise to the factor endowment hypothesis (FEH) concerning the impact of trade on the environment.

## 5.4 A survey of empirical evidence from the literature

Empirical analyses of the impact of trade on pollution generally follow a common strategy. A cross-section or panel of countries is employed in which an indicator of environmental degradation is assumed to depend on a polynomial of degree up to 3 in *per capita* income, an indicator of trade activity, and a set of control variables that correspond to additional determinants of environmental degradation.

Cavlovic et al. (2000) in their meta-analysis study, which can be seen as a summary of most empirical work don in the area, found that including trade tend to yield higher turning points. An important result reported in this study is that carbon dioxide is predicted to have an extremely high turning point.

In a more recent Meta-analysis to investigate systematic variation across Environmental Kuznets Curve studies, Li et al. (2007) found also that controlling for the impact of trade lowers the probability of finding an EKC relationship.

Suri and Chapman (1998) analyse the impact of international trade on commercial energy consumption and find that most exports by industrialising countries are consumed in industrialised countries, allowing these countries to benefit from avoided pollution. They find empirical evidence that incorporating trade effects would tend to increase the turning point for pollutant emissions related to energy use, a result echoed by other studies[1]. They find that reductions in environmental degradation that follow a rise in income is not a result of a positive net improvement in environmental quality, but purely a displacement of pollution from rich countries to poor. International trade allows this displacement to occur. Several theories have been established to explain why this may arise. For instance, Frankel & Rose (2005) and Antweiler et al. (2001) all provide empirical models based on the Pollution Haven Hypothesis. The assumption is that low income countries have less stringent environmental regulations and hence have a comparative advantage in dirty industries. These studies nonetheless fail to find strong pollution haven effects. Hettige et al. (1992), on the other hand find empirical evidence which is consistent with the hypothesis that stricter environmental regulation in OECD countries has led to a locational displacement of dirty industries towards poorer countries. The only paper to directly test for a factor endowment effect is Antweiler et al. (2001) who examine the impact of trade liberalisation on sulphur dioxide concentrations. They found some evidence for factor endowment effects.

Frankel and Rose (2002) provide empirical evidence that trade may indeed have a beneficial effect on some measures of environmental quality . Thus, it seems from this perspective that trade at the very minimum will not certainly result in environmental damage, in fact in many cases it yields environmental benefits. The findings of the paper generally support the EKC and the proposition that openness

[1]Suri and Chapman (1998) report that imports of manufactured goods by developed countries play a role in the EKC downturn and they suggested that with increasing world trade it is likely that this trend will intensify.

to trade accelerates the growth process. However, as expected, the criticisms of this approach are numerous in response to this standard neo-classical type argument. Harbaugh et al. (2000) provides recent empirical evidence against the existence of an EKC for certain pollutants . This evidence is inconsistent with other studies emphasising the controversy surrounding the existing substantiation. Nevertheless where EKCs are found, trade is seen to be a contributor for the high ITPs which is in accordance with empirical evidence from other studies.

Clearly more work needs to be done to fully understand the role of international trade in mediating the relationship with the environment. On the one hand, there appears to be little evidence in support of the pollution haven hypothesis; to the contrary, there is increasing evidence open economies tend to be cleaner than closed economies. However, a growing body of empirical literature has showed that the existence of EKC's has profound effects on the environment.

## 5.5 Parameter Heterogeneity Implied by Trade Models

In the previous Sections we have have shown that, with regard to the relationship between, economic growth, trade policies, and the environment

(i) there is substantial empirical evidence that the impact of an increase in income on the environment depends on the stage of development, and

(ii) economic theory suggests that an increased openness to foreign markets might have a different impact on the environment in developed and developing countries.

The standard approach in investigating the relationship between trade, growth and the environment, assumes that parameters do not vary across countries, i.e.,

$$E = \beta_0 + \sum_{i=1}^{p} \beta_i \, INC^i + \beta_{p+1} OPEN + \sum_{j=p+2}^{k} \beta_j \, Z_j, \tag{5.1}$$

where $E$ is a measure of environmental degradation. $INC$ is a measure of economic activity. $OPEN$ a measure of openness to trade, and $Z_j$ are other determinants of environmental degradation. In particular, the impact of trade on the environment is a constant

$$\frac{\partial E}{\partial OPEN} = \beta_{p+1}$$

The theoretical arguments illustrated in the previous section imply heterogeneity. For instance, consider the the pollution haven hypothesis: since high-income countries demand a cleaner environment, their governments enforce stricter regulations over the domestic industry and allow importing pollution-intensive goods from less regulated countries, so that

$$\frac{\partial E}{\partial OPEN} = \beta_3(INC) \quad \begin{cases} < 0, & \text{for large } INC, \\ > 0, & \text{for small } INC, \end{cases}$$

i.e., the impact of openness on the environment is a function of income. Also, for the factor endowment hypothesis: the impact of trade liberalization will depend on the relative availability of the different factors of production (KAPW) in a

country,

$$\frac{\partial E}{\partial OPEN} = \beta_3(KAPW) \quad \begin{cases} > 0, & \text{for large } KAPW, \\ < 0, & \text{for small } KAPW, \end{cases}$$

i.e., the impact of openness on the environment is a function of capital abundance.

## 5.6 Accounting for Heterogeneity in Empirical Work

It seems reasonable to assume that the marginal impact on environmental degradation of a variable such as the GDP *per capita* and openness to trade depend on several factors, such as the level of economic development, factor endowments, trade policies, etc. This has been explicitly and implicitly recognized in the empirical literature on the subject in several ways.

Empirical analyses of the impact of trade on pollution generally follow a common strategy. A cross-section or panel of countries is employed in which an indicator of environmental degradation is assumed to depend on a polynomial of degree up to 3 in *per capita* income, an indicator of trade activity, and a set of control variables that correspond to additional determinants of environmental degradation.

In the EKC literature, a GDP squared term is added to capture those aspects in the relationship between growth and environment that do not remain the same as countries develop. These include structural changes in the composition of GDP and environmental awareness and regulation.

Another approach is to add interaction terms (cross-products) to the basic regression. For instance, Frankel & Rose (2005), to test the pollution haven hypothesis add to the equation linking pollution with growth and trade, the product of openness to trade and income *per capita*. For instance,

$$E = \beta_0 + \beta_1\, INC + \beta_2\, INC^2 + \beta_3 OPEN + \beta_4(OPEN \cdot INC) \tag{5.2}$$

The partial effect of OPEN is given by

$$\frac{\partial P}{\partial OPEN} = \beta_3 + 2\beta_4 INC$$

which is a linear function of income. If rich countries export pollution-intensive activities to poor countries through trade activities, then the interaction term is expected to have a negative impact on a country's environmental quality.

Antweiler et al. (2001), in order to test the alternative factor endowment hypothesis, include also interaction terms between openness and capital per worker levels and squares. Polluting capital-intensive activities should relocate through to the more capital-rich developed countries. The estimated coefficient for the cross-product is expected to have a positive sign. intensive The impact of an increase in income on pollution depends on the composition of output, and therefore on capital per worker. Adding interaction terms has several problems. Firstly, the inclusion of an interaction term makes the model nonadditive, in the sense that the effect of one independent variable on the dependent variable varies according the value of a second independent variable. If, for instance the cross-product between income *per capita* and openness is added, the partial effect of openness depends now on the level of income. The coefficient for openness measures the effect when income *per capita* is zero. which makes little sense. Also its statistical significance of the partial effect of income on pollution will not be a constant. For example in (5.2) the variance of the impact of trade on the environment, is given by the expression

$$\text{var}\left(\frac{\partial E}{\partial OPEN}\right) = \text{var}(\beta_3) + 4\,INC^2\,\text{var}(\beta_4) + 2\,INC\,\text{cov}(\beta_3, \beta_4)$$

The $t$-statistic can then be derived by dividing the partial effect of openness given a particular value of income, by the standard error for the partial effect computed at a particular value of income *per capita*. It is possible that the impact of openness on pollution is significant at some levels of income, while non-significant at other values. For example, openness could impact significantly pollution only in low-income countries.

Another approach consists of fitting separate regressions based on a *threshold variable* ($INC$), and a *threshold* ($\tau$)

$$\begin{aligned} E &= \beta_0 + \beta_1\, INC + \beta_2\, INC^2 + \beta_3\, OPEN, \qquad INC \leq \tau \\ E &= \delta_0 + \delta_1\, INC + \delta_2\, INC^2 + \delta_3\, OPEN, \qquad INC > \tau \end{aligned}$$

Stern & Common (2001) fit an EKC model to OECD and non-OECD samples. The problem with this approach is the arbitrary choice of splitting variable and, for continuous variables, the threshold.

Nonparametric methods can also be used in this context. I am aware of only one applications of this kind. Azomahou & Phu (2006) employs a smooth coefficient models to estimate the model

$$y_{jt} = \boldsymbol{X}_{jt}^{T}\beta(z_{jt}) + \mu_i + \epsilon_{jt}, \tag{5.3}$$

where $y_{jt}$ is the response variable (deforestation rate) of country $j$ with $j = 1,\ldots,N$ in year $t$. with $t = 1,\ldots,T$. $\boldsymbol{X}_{it} = \left(1, \boldsymbol{x}_{it}^{T}\right)$. with $\boldsymbol{x}_{it}$ being a $p \times 1$ vector and $\mu_i$ represents the fixed effect specific to country $i$. The coefficient $\psi(z_{it}) = \left(\alpha(z_{it}), \beta(z_{it})^{T}\right)^{T}$. The model is fitted using Robinson's (1988) approach

$$\beta(z) = \left[\frac{1}{nh^{q}}\sum_{j=1}^{N}\boldsymbol{X}_j\boldsymbol{X}_j'K\left(\frac{z_j - z}{h}\right)\right]^{-1}\left[\frac{1}{nh^{q}}\sum_{j=1}^{N}\boldsymbol{X}_j y_j' K\left(\frac{z_j - z}{h}\right)\right]$$

$$y_{it} = \boldsymbol{X}_{it}^{T}\psi(z_{it}) + \mu_i + \epsilon_{it}, \tag{5.4}$$

where $y_{it}$ is the response variable (deforestation rate) of country $i$ with $i = 1,\ldots,N$ in year $t$. with $t = 1,\ldots,T$. $\boldsymbol{X}_{it} = \left(1, \boldsymbol{x}_{it}^{T}\right)$. with $\boldsymbol{x}_{it}$ being a $p \times 1$ vector and $\mu_i$ represents the fixed effect specific to country $i$. The coefficient $\psi(z_{it}) = \left(\alpha(z_{it}), \beta(z_{it})^{T}\right)^{T}$ is a vector of smooth functionsThe application suffers from the curse of dimensionality. Azomahou & Phu (2006) coefficients depend only on on variable, GDP.

## 5.7 Data

Our data consists of 2,294 observations representing 74 countries, 23 OECD and 51 non-OECD members, spanning the years 1960-1990. The dataset was constructed using data from various sources.

- For the sulfur emissions, we took the data from the ***Historical Global Sulfur Emissions*** data set of A.S.L and Associates (1997), which includes the sulfur dioxide emissions from burning hard coal, brown coal, and petroleum,

and sulfur emissions from mining and related activities for most of the countries of the world during the period 1850-1990 (Allen S. Lefohn 1999).

- The carbon dioxide emissions data come from the 1998 World Bank ***World Development Indicators*** CD-ROM.

- Most macroeconomic data is derived from the ***Penn World Tables*** (PWT) Mark 5.6 which compiles data for 152 countries on 29 subjects for the period 1950-1992.[2]

- Foreign Direct Investment data are taken from the UN ***World Trade Data Base*** discussed in Feenstra, Lipsey, and Bowen (1997).

The sample of countries used in this analysis together with their associated PWT numeric code, are contained in in Appendix F. The variables are:

$SO2_{i,t}$ = Sulfur Emissions measured in tons of sulfur *per capita*. country $i$. year $t$.

$CO2_{i,t}$ = Carbon Emissions measured in tons of carbon *per capita*. country $i$. year $t$.

$RGDPL_{i,t}$ = Real GDP *per capita* (1985 intl. prices). country $i$. year $t$.

$KAPW_{i,t}$ = Non-residential Capital Stock per Worker (1985 intl. prices). country $i$. year $t$.

$OPEN_{i,t}$ = Openness, $\frac{\text{Exports+Imports}}{\text{Nominal GDP}}$, country $i$. year $t$.

$FDI_{i,t}$ = Gross Foreign Direct Investment. in % of GDP. country $i$. year $t$.

Summary statistics for the variables used in this study appear in Table 5.1.

Figure 5.3 shows scatterplots of emissions against real income *per capita*. The variables on the x-axes are graphed on a natural log scale. Panel 5.4(a) displays a scatterplot of the log of $SO_2$ emissions against the log of *per capita* income. The graph suggests a non-linear relationship between mean $SO_2$ emissions and income *per capita*, consistent with an inverted-U shape. Panel 5.4(b) displays a scatterplot of the log of $CO_2$ emissions against the log of *per capita* income. The

[2]The PWT are described in Alan Heston and Robert Summers The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950-1988, Quarterly Journal of Economics, May 1991, pp.327–368. Though PWT Mark 6.1 was available at the moment of writing, the years spanned by the pollution data, and the fact that updated capital stock estimates were not yet available, made us keep the older version. The PWT are available at the Computing in the Humanities and Social Sciences (CHASS) website at http://datacentre2.chass.utoronto.ca/pwt56/ at the University of Toronto.

| Variable | Name | Dimension | Mean | Std. Dev. | Minimum | Maximum | Cases |
|---|---|---|---|---|---|---|---|
| Log of $SO_2$ p.c. | LSO2 | $\log_e$ (t/person) | -5.034 | 1.8803 | -13.93 | -0.7645 | 2294 |
| Log of $CO_2$ p.c. | LCO2 | $\log_e$ (t/person) | 0.7256 | 1.4699 | -2.751 | 4.332 | 2099 |
| GDP p.c. | GDP | \$/person | 5360 | 6244.2 | 303 | 80830 | 2294 |
| Openness to trade | OPEN | – | 58.99 | 46.16 | 4.99 | 423.4 | 2204 |
| Capital intensity | KAPW | \$/person | 15580 | 12993.8 | 261 | 73460 | 1274 |
| FDI | FDI | – | 1.374 | 1.406 | 0.003127 | 17.70 | 1956 |

*Notes:* All monetary figures are in 1985 US dollars. The natural log transformed variables are denoted with the corresponding name prefixed by the capital letter L.

Table 5.1: Summary statistics

graph suggests a linear relationship between $CO_2$ emissions and income *per capita*. Figure 5.3 shows scatterplots of emissions against real income *per capita*.

Figure 5.3 shows scatterplots of the log of emissions against the log of capital stock per worker. The horizontal scale are logarithmic. Figure 5.4(c) is a scatterplot of the log of $SO_2$ emissions against the log of capital stock per worker. Figure 5.4(d) is a scatterplot of the log of $CO_2$ emissions against the log of capital stock per worker. Figure 5.3 shows scatterplots of emissions against real income *per capita*. The horizontal The Figures' patterns are similar to the ones in Figure 5.3.

Figure 5.4 shows scatterplots of the log of emissions against the log of openness. The horizontal scale are logarithmic. Figure 5.5(a) is a scatterplot of the log of $SO_2$ emissions against the log of openness. Figure 5.5(b) is a scatterplot of the log of $CO_2$ emissions against the log of openness. Figure 5.4 shows scatterplots of emissions against the log of openness.

## 5.8 Statistical Significance of Multiple Regimes

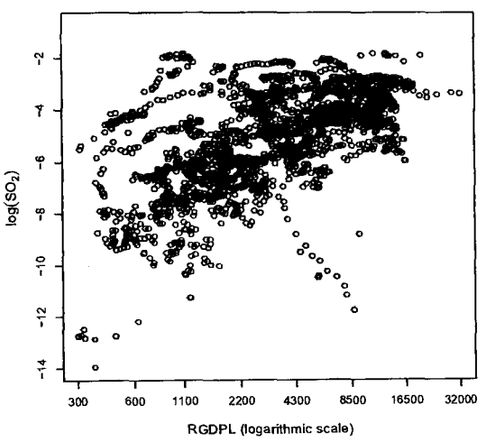
In this section we attempt to identify the existence of several regimes using specification tests which entertain a single regime model as the null hypothesis. We split the sample into sub-groups based upon various determinants of pollution to test whether the regression functions differ across the sub-groups.

We start by fitting, for each sub-group, the model

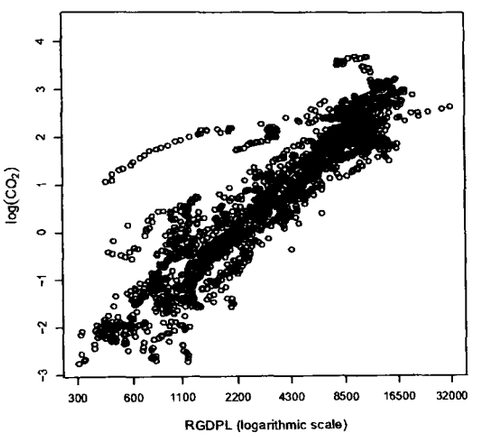
$$\ln(E)_{i,t} = \beta_0 + \beta_1 \ln(GDP)_{i,t} + \beta_2 \ln^2(GDP)_{i,t} + \beta_3 \ln(OPEN)_{i,t} + \\ + \beta_4 \ln(KAPW)_{i,t} + \lambda_t + \alpha_i + \epsilon_{it} \quad (5.5)$$

with $i = 1,\ldots,N$, $t = 1,\ldots,T$. where $E$ is either $SO2_{i,t}$ or $CO2_{i,t}$. $\alpha_i$ and $\lambda_t$ are respectively individual and time specific effects, and $\epsilon_{it} \sim IID(0, \sigma_\epsilon^2)$. The estimated regression represents the unconstrained version of the model. We then fit several constrained versions of the model by imposing cross-coefficient restrictions. We examine sample splits based upon $GDP$. $KAPW$. and $OPEN$. Table 5.2 reports the results of several data splits. Each entry in the table represents the F statistic of the null hypothesis that all parameters are equal across the subsamples under investigation. The first panel of the table divides the countries into two groups

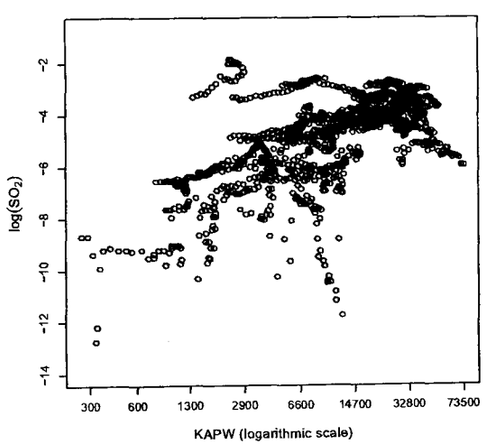
Figure 5.3: Scatterplots of emissions against *per capita* income.



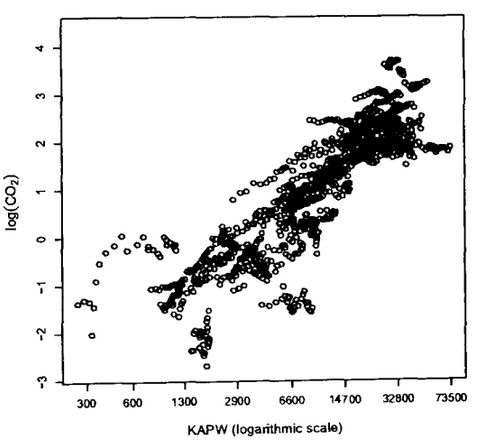

(a) Scatterplot of $\log(SO_2)$ against $\log(RGDPL)$.




(b) Scatterplot of $\log(CO_2)$ against $\log(RGDPL)$.

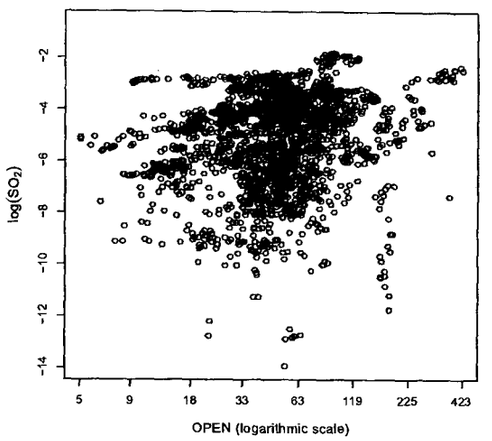



(c) Scatterplot of $\log(SO_2)$ against $\log(KAPW)$.

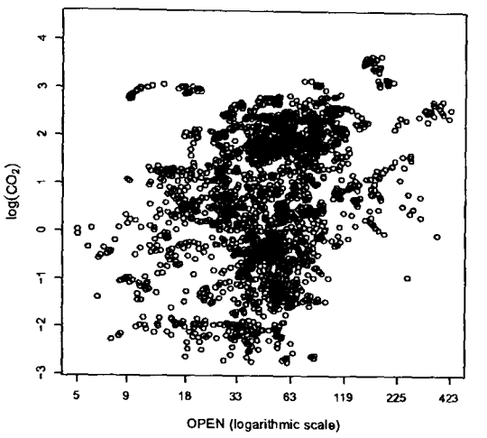



(d) Scatterplot of $\log(CO_2)$ against $\log(KAPW)$.

Figure 5.4: Scatterplots of emissions against openness.




(a) Scatterplot of $\log(SO_2)$ against $\log(OPEN)$.




(b) Scatterplot of $\log(CO_2)$ against $\log(OPEN)$.

Table 5.2: Specification tests for different regimes

| Samples defined by | Wald statistic[a] |
|---|---|
| Two-way split based on | |
| $GDP_{i,t}$ | 35.426 |
| $KAPW_{i,t}$ | 38.483 |
| $OPEN_{i,t}$ | 33.938 |
| Eight-way split based on | |
| $GDP_{i,t}$, $KAPW_{i,t}$, and $OPEN_{i,t}$ | 45.677 |

[a] The Wald statistic is a test of parameter constancy across subsamples asymptotically distributed $\chi^2_k$ under the null of constant parameters, where $k$ is the number of coefficient estimated (excluding country dummies).

Two- and eight-way output splits are based on $GDP < \$3.500$ and $GDP \geq$ \$3,500. For capital stock per worker two- and eight-way splits are based on $KAPW <$ \$11.500 and $KAPW \geq$ \$11,500. For openness to trade two- and eight-way splits are based on $OPEN < \%\ 50$ and $OPEN \geq \%\ 50$.

## 5.9 Economic Significance of Regimes

### 5.9.1 Introduction

Using a large and representative of both high, OECD, and low-income, non-OECD, countries, sample, we estimate a reduced-form relationship between the natural logarithm of *per capita* income and an environmental indicator. Using the estimates from the high and the low income countries samples, we decompose the mean log difference in emissions *per capita* between rich and poor countries into the effects of differences in their average economic activity and the effect due to differences in regimes. We find a significant positive effect for the first component and a large significant negative for the second. The latter part of the decomposition can be interpreted as the excess emissions occurring in developing countries that cannot be explained by income related effects. We argue that the second term in the decomposition can then be interpreted as the part of log of emissions difference due regime differences between rich countries and poor countries. We proceed to assess weather the "regime differences" component of the "emission gap" is significantly reduced if openness and foreign direct investments are included as explanatory variables.

### 5.9.2 Data and $SO_2$ Emission Gap

We define the "Emission Gap" between rich and poor countries as the difference of their respective log of emissions, i.e.,

$$\text{Emission Gap} = \overline{\ln(\boldsymbol{E}_O)} - \overline{\ln(\boldsymbol{E}_N)},$$

where $\ln(\boldsymbol{E}_O)$ and $\ln(\boldsymbol{E}_N)$ be the natural logs of OECD (O) and non-OECD (N) *per capita* sulfur emissions.

The variables used in this study are,

$$SO_{i,t} = \text{Sulfur Emissions measured in tons of sulfur per capita. in year } t.$$
$$GDP_{i,t} = \text{Real GDP per capita (1985 intl. prices). in year } t.$$
$$TRADE_{i,t} = \frac{\text{Exports+Imports}}{\text{Nominal GDP}}, \text{in year } t.$$
$$FDI_{i,t} = \text{Gross Foreign Direct Investment. in \% of GDP. in year } t.$$

Sulfur emissions were taken from the data from the *Historical Global Sulfur Emissions* data set of A.S.L and Associates. The real GDP per capita came from the *Penn World Tables* (PWT) Mark 5.6. The values are all measured in 1985 international US dollars. [3] Foreign Direct Investment data were obtained from the UN *World Trade Data Base.*

The sample consists of 11 annual observations, from 1980 to 1990, the last year available for emissions, on each of 95 countries. The descriptive statistics of the data for the OECD and non-OECD subsamples are reported in Tables 5.3 and 5.4. All monetary figures are in 1985 US dollars. The natural log transformed variables are denoted with the corresponding name prefixed by the capital letter L.

Table 5.3: Descriptive statistics for non-OECD countries

| Variable | Mean | Std.Dev. | Minimum | Maximum | Cases |
|---|---|---|---|---|---|
| LSO | 1.40194684 | 1.15732856 | -1.96134011 | 4.15271265 | 814 |
| LGDP | 3.46441876 | .401948778 | 2.48432661 | 4.63970573 | 814 |
| LGDP2 | 12.1635617 | 2.77844739 | 6.17187869 | 21.5268693 | 814 |
| LTRADE | 1.30102595 | .426146389 | .000000000 | 2.49011346 | 814 |
| LFDI | .361835099 | .557601851 | .000000000 | 2.94443898 | 814 |

Using our data the $SO_2$ emission gap is then,

$$SO_2 \text{ Emission Gap} = \overline{\text{LSO}}_O - \overline{\text{LSO}}_N =$$

$$= 2.26470704 - 1.40194684 = 0.8627602.$$

[3]The PWT are described in Alan Heston and Robert Summers The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950-1988, Quarterly Journal of Economics, May 1991, pp. 327–368.

Table 5.4: Descriptive statistics for OECD countries

| Variable | Mean | Std.Dev. | Minimum | Maximum | Cases |
|---|---|---|---|---|---|
| LSO | 2.26470704 | 1.06196150 | -1.89688075 | 4.11972491 | 231 |
| LGDP | 4.06674365 | .166822076 | 3.47654372 | 4.32092077 | 231 |
| LGDP2 | 16.5661131 | 1.31290362 | 12.0863562 | 18.6703563 | 231 |
| LTRADE | 1.62284191 | .255359757 | 1.01105803 | 2.07993370 | 231 |
| LFDI | .963775374 | .643684250 | .000000000 | 2.56494936 | 231 |

Since natural logarithmic differences are approximately equal to percentage differences, we can state that OECD countries emit 86 per cent more sulfur dioxide then non-OECD countries. In the next Section we provide a useful decomposition of this gap by an application of the Blinder-Oaxaca decomposition.

### 5.9.3 Decomposition of the Emission Gap

With the coefficients from separate models for OECD and non-OECD countries, the emission gap between rich and poor countries can be decomposed into the differences in pollution emissions due to difference in the level of economic activity and differences due to differences in the reduced form relationship between environmental quality and economic activity.

The method employed was developed by Blinder (1973) and Oaxaca (1973) and has been used traditionally to investigate discrimination in wages. Let $\ln(\boldsymbol{E}_O)$ and $\ln(\boldsymbol{E}_N)$ denote the natural logs *per capita* sulfur emissions of OECD (O) and non-OECD (N) countries. The decomposition presupposes the estimation by OLS of the standard emission/income model for the two samples separately

$$\ln(\boldsymbol{E}_O) = \boldsymbol{X}_O'\boldsymbol{\beta}_O + u_O \tag{5.6}$$

$$\ln(\boldsymbol{E}_N) = \boldsymbol{X}_N'\boldsymbol{\beta}_N + u_N. \tag{5.7}$$

where $\boldsymbol{X}$ is a vector of characteristics and $\boldsymbol{\beta}$ is a conforming vector of regression coefficients. A numerical consequence of using ordinary least square is that the residuals sum to zero. This implies that the regression hyperplane includes the

points of means of the data. So for the two samples we have

$$\ln(\widetilde{E}_O) = \overline{X}'_O \widehat{\beta}_O \tag{5.8}$$

$$\ln(\widetilde{E}_N) = \overline{X}'_N \widehat{\beta}_N, \tag{5.9}$$

where $\widetilde{E}$ denotes the geometric mean of emissions *per capita*, $\overline{X}$ is a vector of mean values of regressors, and $\widehat{\beta}$ is a conforming vector of estimated coefficients. The observed difference in the mean log of *per capita* emissions must equal

$$\ln(\widetilde{E}_O) - \ln(\widetilde{E}_N) = \overline{X}'_O \widehat{\beta}_O - \overline{X}'_N \widehat{\beta}_N, \tag{5.10}$$

i.e., either

$$\ln(\widetilde{E}_O) - \ln(\widetilde{E}_N) = \left(\overline{X}_O - \overline{X}_N\right)' \widehat{\beta}_O + \overline{X}'_N \left(\widehat{\beta}_O - \widehat{\beta}_N\right) \tag{5.11}$$

or

$$\ln(\widetilde{E}_O) - \ln(\widetilde{E}_N) = \left(\overline{X}_O - \overline{X}_N\right)' \widehat{\beta}_N + \overline{X}'_O \left(\widehat{\beta}_O - \widehat{\beta}_N\right) \tag{5.12}$$

where (5.11) and (5.12) are obtained by adding $\left(\overline{X}'_N \widehat{\beta}_O - \overline{X}'_N \widehat{\beta}_O\right)$ to (5.10) and $\left(\overline{X}'_O \widehat{\beta}_N - \overline{X}'_O \widehat{\beta}_N\right)$ to (5.10), respectively. The first term of the decomposition is the factor endowment component of the *per capita* emissions gap, the and the second term the structural change component.

Neumark (1988) has pointed out, in the context of wage discrimination, that considerable variation may exist in the estimate of the components obtained using (5.11) as opposed to (5.12). If (5.11) is selected as the model, it is assumed the richer countries' environment/economic regime becomes the one that would exist in the absence of differences in the technologies adopted, environmental regulations, and displacement effects, among other factors. In (5.12), the poorer countries' regime would be the prevailing one. In principle, a weighted average approach as suggested by Cotton (1988) might be more suitable. As this is a preliminary study, this choice is not critical and a more detailed analysis can be the subject of further research.

The structural change component of the "emission gap" between rich and poor countries, can be interpreted as the difference in emissions occurring in developing countries that cannot be explained by income related effects. In the absence of regime differences, OECD and non-OECD countries would have identical emissions

with the same level of economic activity.

## 5.10 Decomposition Results

The specification used, using the variable defined in Section 5.9.2, is

$$LSO_{i,t} = \beta_0 + \beta_1\, LGDP_{i,t} + \beta_2\, LGDP2_{i,t} + \beta_3\, LTRADE_{i,t} + \\ + \beta_4\, LFDI_{i,t} + \alpha_i + \lambda_t + \epsilon_{it} \quad (5.13)$$

with $i = 1, \ldots, N$, $t = 1, \ldots, T$, where $\alpha_i$ and $\lambda_t$ are respectively individual and time specific effects, and $\epsilon_{it} \sim IID(0, \sigma_\epsilon^2)$. Tables 5.5, 5.6, and 5.7 display the results of the fixed and random effects estimation of equation (5.13) with only GDP variables, with openness to trade, and with both openness to trade and FDI, respectively.

Table 5.5: Panel regression results

| Region | World | | OECD | | Non-OECD | |
|---|---|---|---|---|---|---|
| Model | FE | RE | FE | RE | FE | RE |
| LGDP | 2.918 (2.866) | 2.875 (3.013) | 33.974 (5.182) | 34.056 (5.302) | 2.231 (2.017) | 2.387 (2.324) |
| LGDP2 | −0.301 (−2.240) | −0.293 (−2.285) | −3.811 (−4.522) | −4.047 (−5.028) | −0.217 (−1.470) | −0.242 (−1.733) |
| Constant | −4.955 (−2.562) | −4.902 (−2.757) | −72.763 (−5.560) | −69.189 (−5.360) | −3.688 (−1.783) | −3.924 (−2.076) |
| LTRADE | | | | | | |
| FDI | | | | | | |
| TP | 4.847176 | 4.906143 | 4.45736 | 4.207561 | 5.140553 | 4.931818 |
| TP | 127.3802 | 135.1173 | 86.25948 | 67.19246 | 170.8102 | 138.6313 |
| LM Test | 4842.41 (0.000000) | | 1045.50 (0.000000) | | 3763.40 (0.000000) | |
| Hausman | .18 (0.912921) | | 5.51 (0.063662) | | 1.49 (0.474632) | |
| DW | 0.975858 | 0.975858 | 0.86596 | 0.86596 | 1.045988 | 1.045988 |
| Wald (joint) | 3953 (10) | 17080 (10) | 3953. (10) | 3953. (10) | 3953. (10) | 3953. (10) |
| Wald (time) | 526.3 (27) | 639.7 (27) | 526.3 (27) | 526.3 (27) | 526.3 (27) | 526.3 (27) |

*Notes:* The dependent variable is the natural log of sulfur emissions sulphur dioxide ($SO_2$) *per capita*. $t$ statistics are reported in parenthesis. H is the Fixed vs. Random Effects Hausman test statistic; LM is the Lagrange Multiplier Test for the significance of individual effects; TP are the income turning point expressed in thousands of US dollars.

Table 5.6: Panel regression results with trade

| Region | **World** | | **OECD** | | **Non-OECD** | |
|---|---|---|---|---|---|---|
| Model | FE | RE | FE | RE | FE | RE |
| Constant | −5.526 (−2.848) | −5.129 (−2.852) | −72.452 (−5.550) | −68.932 (−5.367) | −3.850 (−1.872) | −3.967 (−2.075) |
| LGDP | 3.158 (3.097) | 2.972 (3.097) | 34.240 (5.234) | 34.542 (5.398) | 2.191 (1.993) | 2.347 (2.272) |
| LGDP2 | −0.349 (−2.580) | −0.310 (−2.402) | −3.833 (−4.559) | −4.083 (−5.096) | −0.227 (−1.549) | −0.243 (−1.733) |
| LTRADE | 0.248 (2.568) | 0.078 (0.923) | −0.638 (−1.419) | −1.011 (−2.705) | 0.328 (3.109) | 0.147 (1.600) |
| TP (logs) | 4.524355 | 4.793548 | 4.466475 | 4.229978 | 4.825991 | 4.829218 |
| TP (levels) | 92.23641 | 120.7290 | 87.04933 | 68.71572 | 124.71 | 125.1131 |
| LM Test | 4694.35 (0.000000) | | 956.47 (0.000000) | | 3701.42 (0.000000) | |
| Hausman | 13.32 (0.003994) | | 6.13 (0.105530) | | 12.24 (0.006616) | |
| DW | 0.992356 | 0.992356 | 0.874958 | 0.874958 | 1.061302 | 1.061302 |

*Notes:* The dependent variable is the natural log of sulfur emissions sulphur dioxide ($SO_2$) *per capita*. $t$ statistics are reported in parenthesis. H is the Fixed vs. Random Effects Hausman test statistic; LM is the Lagrange Multiplier Test for the significance of individual effects; TP are the income turning point expressed in thousands of US dollars.

Table 5.7: Panel regression results with trade and FDI

| Region | World | | OECD | | Non-OECD | |
|---|---|---|---|---|---|---|
| Model | FE | RE | FE | RE | FE | RE |
| LGDP | 2.485 (2.434) | 3.003 (2.977) | 27.444 (4.172) | 27.758 (4.309) | 2.196 (1.996) | 2.350 (2.273) |
| LGDP2 | −0.253 (−1.865) | −0.320 (−2.390) | −2.975 (−3.524) | −3.195 (−3.952) | −0.229 (−1.556) | −0.244 (−1.738) |
| LTRADE | 0.306 (3.173) | 0.179 (2.015) | −0.637 (−1.465) | −0.959 (−2.702) | 0.320 (2.995) | 0.141 (1.510) |
| LFDI | −0.162 (−4.437) | −0.026 (−1.709) | −0.146 (−3.801) | −0.144 (−3.823) | 0.012 (0.429) | 0.009 (0.348) |
| Constant | −130.526 (−4.623) | −25.384 (−1.851) | −58.882 (−4.486) | −55.995 (−4.343) | −3.847 (−1.869) | −3.963 (−2.071) |
| TP | 4.915 | 4.690 | 4.612 | 4.344 | 4.805 | 4.821 |
| TP | 136.325 | 108.899 | 100.706 | 76.997 | 122.065 | 124.105 |
| LM Test | 4842.41 (0.000000) | | 1045.50 (0.000000) | | 3763.40 (0.000000) | |
| Hausman | 14.23 (0.006586) | | 6.43 (0.169351) | | 13.30 (0.009920) | |
| DW | 0.975858 | 0.975858 | 0.86596 | 0.86596 | 1.045988 | 1.045988 |
| Wald (joint) | 3953 (10) | 17080 (10) | 3953. (10) | 3953. (10) | 3953. (10) | 3953. (10) |
| Wald (time) | 526.3 (27) | 639.7 (27) | 526.3 (27) | 526.3 (27) | 526.3 (27) | 526.3 (27) |

*Notes:* The dependent variable is the natural log of sulfur emissions sulphur dioxide ($SO_2$) *per capita*. $t$ statistics are reported in parenthesis. H is the Fixed vs. Random Effects Hausman test statistic; LM is the Lagrange Multiplier Test for the significance of individual effects; TP are the income turning point expressed in thousands of US dollars.

Table 5.8 displays the results of the Blinder-Oaxaca decomposition of the contribution to the gap of each explanatory variables, the total, and the proportion over the total, for the income and regime differences components. The penultimate line of Table 5.8 reports the $t$ statistics associated with the two components of the "emission gap" for the without trade variables, with openness to trade, and with all trade related variables.[4]

For instance, when only income related variables are included, we find that the that the pollution emission differential due to differences in the level of economic activity is 2.696, whereas the emission differential due to differences in the reduced form relationship between environmental quality and economic activity is -1.833. This implies that richer countries would emit 270 per cent more sulfur dioxide than poorer ones, i.e., the gap would be much wider, if poorer countries' emissions were obtained by evaluating their level of economic using the OECD estimated relationship. The difference between this "counterfactual emission gap" and the existing one of 180 per cent can be interpreted as the excess pollution emissions that poorer countries are currently causing because of differences in regimes.This suggests that rapid growth of developing countries that is not accompanied by significant structural changes, in the shape of technologies adopted, environmental regulations, and so on, could have detrimental consequences on the environment. Including trade and FDI, reduces the regime differences component to 130 per cent.

[1]The $t$ statistics were computed using standard econometric results. If $\boldsymbol{w}$ is a $k \times 1$ vector of constants, then $\boldsymbol{w}'\boldsymbol{\beta} \sim N\left(\boldsymbol{w}'\boldsymbol{\beta}, \boldsymbol{w}'\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w}\right)$. The variance can be estimated with $\boldsymbol{w}'\boldsymbol{s}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w}$. To calculate the $t$ statistics of the unexplained component we can apply the above result by treating the vector of sample means for the non-OECD countries as constant. Assuming that the two sets of observations are independent than $\widehat{\boldsymbol{\beta}}_O$ and $\widehat{\boldsymbol{\beta}}_N$ will be independent with means $\boldsymbol{\beta}_O$ and $\boldsymbol{\beta}_N$, and covariance matrices $\sigma_O^2(\boldsymbol{X}_O'\boldsymbol{X}_O)^{-1}$ and $\sigma_N^2(\boldsymbol{X}_N'\boldsymbol{X}_N)^{-1}$. The estimated covariance matrix for $\boldsymbol{d} = \widehat{\boldsymbol{\beta}}_O - \boldsymbol{\beta}_N$ is given by $\text{var}(\boldsymbol{d}) = \sigma_O^2(\boldsymbol{X}_O'\boldsymbol{X}_O)^{-1} + \sigma_N^2(\boldsymbol{X}_N'\boldsymbol{X}_N)^{-1}$. Applying the above result, for $\boldsymbol{w} = \overline{\boldsymbol{X}}_N$, the variance of $\overline{\boldsymbol{X}}_N'\boldsymbol{d}$ is then $\overline{\boldsymbol{X}}_N'\,\text{var}(\boldsymbol{d})\overline{\boldsymbol{X}}_N$, so that $t = \frac{\overline{\boldsymbol{X}}_N'\boldsymbol{d}}{\sqrt{\overline{\boldsymbol{X}}_N'\,\text{var}(\boldsymbol{d})\overline{\boldsymbol{X}}_N}}$. For the explained component, the variance of $\left(\overline{\boldsymbol{X}}_O - \overline{\boldsymbol{X}}_N\right)'\widehat{\boldsymbol{\beta}}_O$, applying the above result with $\boldsymbol{w} = \left(\overline{\boldsymbol{X}}_O - \overline{\boldsymbol{X}}_N\right)$, is then $\left(\overline{\boldsymbol{X}}_O - \overline{\boldsymbol{X}}_N\right)'\text{var}(\widehat{\boldsymbol{\beta}}_O)\left(\overline{\boldsymbol{X}}_O - \overline{\boldsymbol{X}}_N\right)$, so that

$$t = \frac{\left(\overline{\boldsymbol{X}}_O - \overline{\boldsymbol{X}}_N\right)'\widehat{\boldsymbol{\beta}}_O}{\sqrt{\left(\overline{\boldsymbol{X}}_O - \overline{\boldsymbol{X}}_N\right)'\sigma_O^2(\boldsymbol{X}_O'\boldsymbol{X}_O)^{-1}\left(\overline{\boldsymbol{X}}_O - \overline{\boldsymbol{X}}_N\right)}}.$$

Table 5.8: Blinder-Oaxaca decomposition of sulfur emissions of OECD and non-OECD countries

| variable | **Income** | | **with Trade** | | **with Trade/FDI** | |
|---|---|---|---|---|---|---|
| Model | $\Delta\hat{\beta}\cdot\overline{X}$ | $\hat{\beta}\cdot\Delta\overline{X}$ | $\Delta\hat{\beta}\cdot\overline{X}$ | $\hat{\beta}\cdot\Delta\overline{X}$ | $\Delta\hat{\beta}\cdot\overline{X}$ | $\hat{\beta}\cdot\Delta\overline{X}$ |
| Constant | -65.265 | | -64.965 | | -52.033 | |
| LGDP | 109.713 | 20.513 | 111.534 | 20.805 | 88.024 | 16.720 |
| LGDP2 | -46.281 | -17.817 | -46.706 | -17.974 | -35.900 | -14.067 |
| LTRADE | | | -1.506 | -0.325 | -1.430 | -0.309 |
| LFDI | | | | | -0.055 | -0.086 |
| Total Gap | −1.833 (−5.998) | 2.696 (5.707) | −1.643 (−5.688) | 2.506 (5.368) | −1.395 (−4.896) | 2.258 (5.060) |
| Proportion | 40.47 | 59.53 | 39.60 | 60.40 | 38.19 | 61.81 |

*Notes:* Calculations are based on mean values of all variables in Tables 5.3 on page 147 and 5.4 on page 148 and the estimation results in Tables 5.5 on page 151, 5.6 on page 152, and 5.7 on page 153. $t$ statistics associated with each component are given in parenthesis.

We found that the Blinder-Oaxaca decomposition is a promising technique that can be used decompose the emission gap between rich and poor countries. Using a large sample of panel data representative of both high and low-income countries, we find that structural differences between developed and developing countries can be quite substantial. We also find support for the hypothesis of pollution displacement from rich to poor countries via international trade and foreign direct investment. More general decompositions could provide information on the evolution over time of the various effects. Also focussing on the entire distribution rather than the mean only could provide interesting insight into the relationship between economic activity and the environment. In the next sections we will employ a tree regression approach to explore the existence of different environment/economic regimes.

## 5.11 Tree Regression Methodology

Though the exogenous splits introduced in Sections 5.8 and 5.9 allow simple specification testing and to assess the economic environmental significance of regime differences, they do not permit the identification of economies with a common relationships between the environment and economic factors.

In this section we are going to describe the regression-tree approach introduced by Breiman et al. (1984). This approach is particularly well suited when there is significant interaction structure between the explanatory variables. The method was applied by Durlauf & Johnson (1995) to investigate the existence of multiple regimes in cross-country growth behavior. If we rewrite the support of each $x_{i,j}$ as the union of $M$ intervals, $a_{i,0} \leq x_{i,j} < a_{i,1}, \ldots, a_{i,M-1} \leq x_{i,j} < a_{i,M}$, the support $S$ of $X$, can then be expressed as $S = \cup_{m=1}^{M^r} S_m$. The function $f(X)$ can then be approximated by a piecewise linear function of the form

$$f(X) \approx \sum_{m=1}^{M^r} \delta_m(X) X \beta_{S_m},$$

where

$$\delta_m(X) = \begin{cases} 1, & \text{if } X \in S_m; \\ 0, & \text{otherwise.} \end{cases}$$

If for each variable $x_i$, $i = 1, \ldots, r$ we split the data into two subgroups following the decision rule: assign observation $j$ to $S_{a,i}$ if $x_{i,j} < a$, otherwise assign the observation to $S_{\overline{a},i}$. Letting $a$ take on values across the support of $x_i$ and repeating this operation for all variables included in the model, we can identify all possible binary data splits. Let $\widehat{\beta}_{a,i}$ denote all OLS estimate of $y_j$ onto $X_j$ using the observation assigned to $S_{a,i}$; Define $\widehat{\beta}_{\overline{a},i}$ in an analogous way. Some split variable $x_i$ and some value $a$ will minimize the sum of squared residuals (SSR) over all possible splits

$$\sum_{j \in S_{a,i}} \left(y_j - X_j \widehat{\beta}_{a,i}\right)^2 + \sum_{j \in S_{\overline{a},i}} \left(y_j - X_j \widehat{\beta}_{\overline{a},i}\right)^2.$$

One crucial limitation of this approach is that the estimated thresholds have

no known distribution theory. Hansen (2000) developed an threshold estimation testing procedure procedure with accompanying distribution theory. This methodology will be applied at the end of Chapter 6, in Section 6.9, starting on 189.

We adapt the method to work with panel data by applying the within (time demeaning) transformation to the dependent, $y$ and the independent, $x$, variables, i.e.,

$$\tilde{y}_{it} - \bar{y}_{i.}$$

$$\tilde{x}_{it} - \bar{x}_{i.}$$

where $\bar{y}_{i.} = \frac{1}{T}\sum_t y_{it}$ and $\bar{x}_{i.} = \frac{1}{T}\sum_t x_{it}$ are the within averages.

By running an OLS on the transformed data we are estimating a panel fixed effect model and are able to control for individual heterogeneity. Since the regression has been performed with an ordinary least squares program, a degrees of freedom correction has to be applied to standard errors and $t$-statistics to obtain the corresponding correct values,

$$se_a(\widehat{\beta}) = \sqrt{\frac{v_a}{v_u}}se_u(\widehat{\beta})$$

$$t_a(\widehat{\beta}) = \sqrt{\frac{v_a}{v_u}}t_u(\widehat{\beta})$$

where $v_a = v_u - N$, "a" denotes adjusted and, "u" unadjusted.

## 5.12 Tree Estimation results

In this section we present the fixed effects tree regression result from the $SO_2$ and the $CO_2$ equations. Regression tree estimates where obtained using GUIDE, developed by Loh (Loh, 2005).[5]

[5]GUIDE stands for Generalized, Unbiased, Interaction Detection and Estimation. It is freely available from the Internet address `www.stat.wisc.edu/~loh/` as compiled executables for Linux and Windows on Intel and compatible processors, and for Mac OS X. The hardware used was a Dual Intel Pentium IV (Prestonia) Xeon Processors 3.06 GHz with HT Technology with 4 GB of RAM running on Microsoft Windows XP/2002 Professional (Win32 x86) 5.01.2600 (Service Pack 2). We used GUIDE Regression Tree version 3.1, the standard Win32 release available at the time of writing the present Chapter.

**Sulfur Emissions** The specification used to estimate a tree is the log-log functional form,

$$\ln(SO_2)_{i,t} = \beta_0 + \beta_1 \ln(GDP)_{i,t} + \beta_2 \ln^2(GDP)_{i,t} + \beta_3 \ln(OPEN)_{i,t} + \\ + \beta_4 \ln(FDI)_{i,t} + \beta_5 \ln(KAPW)_{i,t} + \alpha_i + \epsilon_{it} \quad (5.14)$$

with $i = 1,\ldots,N$, $t = 1,\ldots,T$, using the variable defined in section 5.7 on page 140. The results of the tree regression procedure applied to the sulfur emissions equation are shown as a binary tree in Figure 5.5. Diamonds in this figure indicate the splitting criteria for the sample expressed in terms of splitting variable and threshold value; circles represent terminal nodes which contain the estimated subsamples. Number in italics beneath a leaf is the sample mean of LSO. The regression tree for sulfur emissions partitions the sample into low-, intermediate- and high-income countries, groups (4), (5) and (3), respectively, and then partitions the low output countries into low- and high-capital intensive countries, groups (6) and (7) respectively. The fact that, given the opportunity to split the sample by either income, capital intensity, openness to trade, and foreign direct investment, the regression tree shows preference for income splits suggests that income dominates trade and endowment variables as a variable useful in identifying multiple regimes in the $so_2$ data. The estimated terminal subsamples are: (3) $GDP >$ \$9,400,

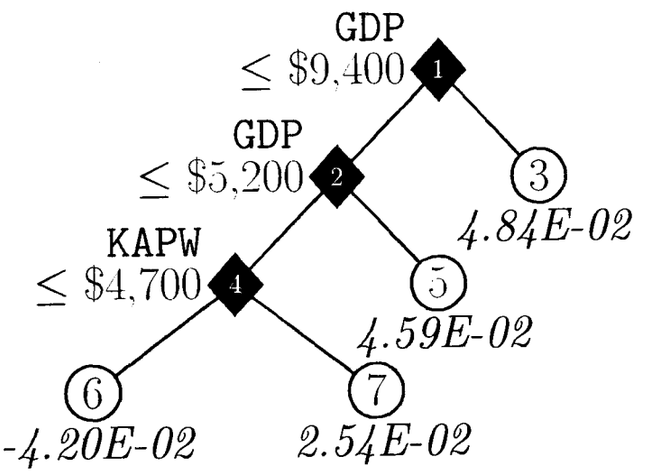



Figure 5.5: Regression binary tree for sulfur emissions

(5) \$5,200 $< GDP \leq$ \$9,400. (6) $KAPW \leq$ \$4,700 and $GDP \leq$ \$5,200, and (7) $KAPW >$ \$4,700 and $GDP \leq$ \$5,200. The list of countries belonging to each

Table 5.9: Regression tree sample break for $SO_2$

| Terminal node number | | | |
|---|---|---|---|
| 5 | 3 | 6 | 7 |
| Barbados | Australia | Guatemala | Argentina |
| Cyprus | Austria | Honduras | Bolivia |
| Spain | Belgium | India | Chile |
| Greece | Canada | Kenya | Colombia |
| Ireland | Switzerland | Morocco | Romania |
| Israel | Germany, West | Madagascar | Czechoslovakia |
| S. Korea | Denmark | Nigeria | Peru |
| Mexico | Finland | Philippines | Iran |
| Taiwan | France | Thailand | Sri Lanka |
| Portugal | United Kingdom | Zambia | Syria |
| U.S.S.R. | Hong Kong | Zimbabwe | Turkey |
| Trinidad and Tobago | Italy | | Yugoslavia |
| Venezuela | Japan | | |
| | Kuwait | | |
| | Luxembourg | | |
| | Netherlands | | |
| | Norway | | |
| | New Zealand | | |
| | United Arab E. | | |
| | Singapore | | |
| | Sweden | | |
| | U.S.A. | | |

subsample are presented in Table 5.9.

The list indicates that there is substantial geographic homogeneity within each group, giving some support to findings by geographical factors (see, e.g., Neumayer, 2002). The low income high capital intensity is composed almost exclusively by Latin American and Eastern European countries. The low income and low capital intensity group is composed almost exclusively by developing African countries. North American and European countries dominate the high-income group. This classification also suggests the importance of democracy (see, e.g., Torras & J.K., 1998; Harbaugh et al., 2002), corruption (see, e.g., Lopez & Mitra, 2000), and civil and political liberties (see, e.g., Barrett & Graddy, 2000; Torras & J.K.,

Table 5.10: Fixed Effects coefficient estimates at each node for sulfur

| Node[a] | 2 | 4 | 6 | 7 | 5 | 3 |
|---|---|---|---|---|---|---|
| Constant | −0.0008179 (−0.04) | 0.04756 (2.433) | 0.1166 (3.852) | −0.02178 (−0.844) | −0.07181 (−1.324) | −0.245 (−9.242) |
| LGDP | 2.611 (2.757) | 0.4866 (0.453) | −1.066 (−0.598) | 1.614 (0.905) | 0.1182 (0.032) | 50.35 (14.3) |
| LGDP2 | −0.1217 (−2.074) | 0.01661 (0.238) | 0.1365 (1.112) | −0.07242 (−0.649) | 0.03534 (0.16) | −2.601 (−13.25) |
| LKAPW | 0.01842 (0.196) | 0.1824 (2.03) | 0.04662 (0.355) | 0.513 (3.866) | −0.5246 (−2.328) | −2.16 (−13.11) |
| LOPEN | 0.2544 (2.413) | 0.1207 (1.25) | 0.09293 (0.527) | 0.1011 (0.932) | 0.6897 (2.467) | 0.8793 (5.307) |
| LFDI | 0.07533 (3.316) | 0.07777 (3.795) | 0.1029 (3.286) | 0.06401 (2.457) | 0.07759 (1.206) | 0.01739 (0.529) |
| Cases | 1893 | 1447 | 1136 | 311 | 446 | 401 |
| $R^2$ | 0.1628 | 0.2832 | 0.2876 | 0.3464 | 0.1011 | 0.6770 |
| TP[b] | 10.73 (45,610) | | 3.905[c] (49.64) | 11.14 (68,970) | | 9.679 (15,980) |

*Notes:* The dependent variable is the natural log of sulfur emissions *per capita*. For the coefficient $t$ statistics are reported in parenthesis. Country-specific dummies are included in all equations.

[a] Node numbers correspond to the node numbers in Figure 5.5.

[b] Turning points values in US dollars are reported in parenthesis.

[c] Implied curve is U-shaped, monotone increasing over the observed sample.

1998). This is particularly striking if we consider the country composition of group 7, namely: Argentina, Bolivia, Chile, Colombia, Romania, Czechoslovakia, Peru, Iran, Sri Lanka, Syria, Turkey, and Yugoslavia. Our findings support studies that based on the poor environmental performance of Soviet economies and dictatorships established in Latin America, Asia and Africa, have been advocating democratic reforms as a way to promote both economic and environmental welfare (see, e.g., McCloskey, 1983; Payne, 1995). For instance, McCloskey argues that "Many of the important ecological measures that are being implemented are being implemented in democracies. ... [omissis] By contrast, if we consider actual totalitarian states, China, Chile, the USSR, Argentina, the dictatorships of Africa and the Arab world, we find that they are far from ecologically minded. ... [omissis] China and the USSR are among the worst ecological offenders" (McCloskey, 1983, p. 157).

Table 5.10 presents the fixed effects panel regression estimates for each subgroup. The $R^2$ reported are weighted values for fitted cases. We find that for the

low-income countries with low capital intensity, the model explains about 29 per cent, for the low-income with high-capital intensity 35 per cent, for the middle-income countries 10 per cent, and for the high-income countries almost 68 per cent of the total variation in emissions. The estimates differ considerably both in their economic and statistical significance across subsamples. This agrees with Panayotou (2000) which concludes, after examining the evidence form Vincent (1997) and Carson et al. (1997b) concerning the existence of a Kuznets curve within individual countries, that: "whereby rising incomes result in a more effective regulatory structure by changing public preferences and making resources available to regulatory agencies. States with low-income levels have a far greater variability in emissions per capita than high-income states suggesting more divergent development paths. This has the implication that it may be more difficult to predict emission levels for low-income countries approaching the turning point."

Consistently with most of the literature on sulfur emissions, only for thefor the high-income countries belonging to group (3).[6] defined here for $GDP >$ \$9,400, we find evidence of a statistically significant within-sample-range turning point located above the sample mean per capita income of \$5,360, at \$16,000 *per capita*. This could be interpreted as implying that many of the rich nations have crossed the turning point and lie on the downward sloping branch of an environment-economy relationship. For medium and low income countries, the turning point is either non-existent or so high that the curve is monotone increasing over the observed sample range. For the poorest countries of group (4), the income variables are not statistically significant.For this subset of countries sulfur emission are monotone increasing.[7]

Turning point estimates agree with recent empirical studies on similar local impact pollutants. Though the turning point we find is higher than the those typically found in earlier published studies, such as Selden & Song (1994) with a turning point of \$8,700 *per capita*, it is still much lower than some recent much higher estimates. For instance, Harbaugh et al. (2002) found a turning point of

[6]For sulfur emissions group definition and membership, see Figure 5.5 on page 158 and Table 5.9 on page 159.

[7]In particular, for the poorer countries with lower capital intensity of group (6), emissions are very low and the estimated curve is U-shaped, but statistically non significant, whereas for the countries with higher capital-per-worker belonging to group (7), the turning point, at about \$69,000 *per capita*, is well outside the sample range, so that the curve is *de facto* monotone increasing.

$39,700 *per capita* for sulfur dioxide emissions. Harbaugh et al. (2002) in their work re-examined the empirical evidence for the EKC for three local pollutants, i.e., sulfur dioxide, smoke, and total suspended particles (TSP) using a more representative data set. They found that turning point estimates are extremely sensitive to the sample chosen and to econometric specifications. For instance, they also found that sulfur dioxide emissions increase with income with no evidence of a turning point for countries with $GDP > \$8{,}000$. In general, they found that for most specifications, using cleaner data makes the EKC disappear alltogether for the local pollutant included in their study. Evidence from recent literature surveys also supposrt our findings. Cavlovic et al. (2001) in their meta-analysis study, which can be seen as a summary of most empirical work don in the area, found that including trade tend to yield higher turning points. In a more recent Meta-analysis to investigate systematic variation across Environmental Kuznets Curve studies, Li et al. (2007) found also that controlling for the impact of trade lowers the probability of finding an EKC relationship.

Our results for sulfur emissions seem also to give some support to the pollution haven hypothesis. The impact of openness to trade on pollution is almost 4 times higher than it is for rich countries then for poor countries. Frankel & Rose (2005) and Antweiler et al. (2001) fail to find strong pollution haven effects. Hettige et al. (1992), on the other hand found empirical evidence supporting the hypothesis that stricter environmental regulation in OECD countries has led to a relocation of dirty industries towards poorer countries. However, all provide empirical models based on the Pollution Haven Hypothesis. The assumption is that low income countries have less stringent environmental regulations and hence have a comparative advantage in dirty industries. These studies nonetheless fail to find strong pollution haven effects. Hettige et al. (1992), on the other hand find empirical evidence which is consistent with the hypothesis that stricter environmental regulation in OECD countries has led to a locational displacement of dirty industries towards poorer countries.

We find no evidence supporting the factor endowment hypothesis. This finding is somewhat in contrast with the work of Antweiler et al. (2001) who examined the impact of trade liberalisation on sulphur dioxide concentrations and found some evidence for factor endowment effects. The discrepancy could be caused by the way capital abundance was defined in their study. Productivity of workers in different countries is adjusted for differences in their average human capital levels.

Note that European countries follow different regimes. Poorer European countries, and accession countries belong to the poorest economies with higher capital intensity. Our finding suggest that just joining the European Union will not by itself be accompanied by improvements in the environment. Structural changes will need to occur for this to happen.

**Carbon emissions** The specification used to estimate a tree is the log-log functional form,

$$\ln(CO_2)_{i,t} = \beta_0 + \beta_1 \ln(GDP)_{i,t} + \beta_2 \ln^2(GDP)_{i,t} + \beta_3 \ln(OPEN)_{i,t} + \\ + \beta_4 \ln(FDI)_{i,t} + \beta_5 \ln(KAPW)_{i,t} + \alpha_i + \epsilon_{it} \quad (5.15)$$

using the variable defined in section 5.7. The results of the tree regression procedure applied to the carbon dioxide emissions equation, are shown in Figure 5.6.

The regression tree for carbon dioxide emissions partitions the sample into low-, and high-capital intensity countries and then partitions the low capital intensity countries into low- and high-income *per capita* countries. The fact that, given the opportunity to split the sample by either by either income, capital intensity, openness to trade, and foreign direct investment, the regression tree shows preference for income splits suggests that income dominates trade and endowment variables as a variable useful in identifying multiple regimes in the data. The estimated subsamples are: (3) $KAPW > \$22.500$. (4) $KAPW \leq \$22.500$ and $GDP < \$5{,}600$, and (5) $KAPW < \$22.500$ and $GDP \geq \$5{,}600$.

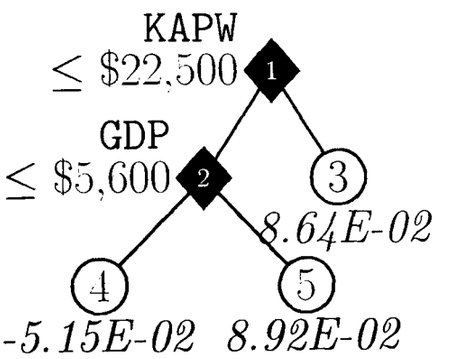



Figure 5.6: Regression tree for carbon dioxide emissions

We find that for the low-income countries with low capital intensity, we explain

about 29 per cent, for the low-income with high-capital intensity 35 per cent, for the middle-income countries 10 per cent, and for the high-income countries almost 68 per cent of the total variation in emissions.

Table 5.12 presents panel regression estimates for each subgroup. The $R^2$ reported are weighted values for fitted cases. All variables are significant except for the trade related variables in group 5 with high income per capita and low capital intensity. The impact of trade for countries with high capital intensity and high income is negative.

Even for carbon emissions the impact of openness to trade and FDI is negative for rich, high capital intensive countries. Though only is for the high capital intensity sample is statistically significant. For poorer, low capital intensity countries openness to trade and FDI tend to increase emissions. The results for carbon emissions also give some support to the pollution haven hypothesis, according to which there will be the tendency for the production of dirty to be moved to low-income countries as a result of international trade.

The estimates of income variables are all statistically significant. We find evidence of an EKC for carbon dioxide emission for all groups of countries.The estimated turning points for high capital intensity and high income countries, groups (3) and (5) respectively,[8] presented in Table 5.12 in Section 5.12 on page 163, are all well above the sample mean income but within the sample range. These turning points are much higher than the the turning point for $SO_2$ emissions. The turning point for the low capital intensity, low income group, is well out of the range of the sample ($391,400 *per capita*) so that emissions are *de facto* monotonically increasing. Non-existent or higher that $SO_2$ emissions turning points for $CO_2$ are consistent with previously published literature suggesting that EKC relationships are more likely to be found for certain types of environmental indicators, particularly those with a more short-term and local impact rather than those with a more long-term and global impacts (see, e.e, Arrow et al., 1995; Cole et al., 1997; Selden & Song, 1994).

These findings also agrees with, for instance, Schmalensee et al. (1998) and Dijkgraaf & Melenberg (2005). Schmalensee et al. (1998) found clear evidence of an inverted-U relationship, with a within-sample turning point between carbon dioxide emissions and per capita income. More recently Dijkgraaf & Melenberg

[8]For carbon dioxide emissions group definition and membership, see Figure 5.6 on page 163 and Table 5.11 on page 165.

Table 5.11: Regression tree sample break for carbon

| Terminal node number | | |
|---|---|---|
| 4 | 5 | 3 |
| Argentina | United Kingdom | Australia |
| Bolivia | Hong Kong | Austria |
| Chile | Ireland | Belgium |
| Colombia | Israel | Canada |
| Guatemala | Korea, Dem. People's Rep. | Switzerland |
| Honduras | Portugal | Germany, West |
| India | Venezuela | Denmark |
| Iran, Islamic Republic of | | Spain |
| Kenya | | Finland |
| Sri Lanka | | France |
| Morocco | | Greece |
| Madagascar | | Italy |
| Mexico | | Japan |
| Nigeria | | Luxembourg |
| Peru | | Netherlands |
| Philippines | | Norway |
| Syrian Arab Republic | | New Zealand |
| Thailand | | Taiwan |
| Turkey | | Sweden |
| Yugoslavia | | U.S.A. |
| Zambia | | |
| Zimbabwe | | |
| Romania | | |
| Czechoslovakia | | |

*Notes:* The dependent variable is the natural log of carbon emissions *per capita*. For the coefficient $t$ statistics are reported in parenthesis. Country-specific dummies are included in all equations.

[a] Node numbers correspond to the node numbers in Figure 5.6.

[b] Turning points values in US dollars are reported in parenthesis.

Table 5.12: Fixed Effects coefficient estimates at each node for carbon dioxide emissions

| Node[a] | 2 | 4 | 5 | 3 |
|---|---|---|---|---|
| Constant | 0.04315 (6.562) | 0.04372 (5.273) | 0.02146 (2.171) | −0.02138 (−1.39) |
| LGDP | 3.056 (10.67) | 1.842 (4.101) | 5.591 (8.413) | 16.63 (8.959) |
| $LGDP^2$ | −0.1427 (−8.214) | −0.07151 (−2.465) | −0.2725 (−7.115) | −0.8266 (−8.24) |
| LKAPW | 0.14 (4.668) | 0.2574 (6.958) | −0.1 (−2.291) | −0.841 (−7.623) |
| LOPEN | 0.09736 (2.899) | 0.1292 (3.267) | −0.005713 (−0.104) | −0.295 (−3.147) |
| LFDI | 0.02442 (3.387) | 0.03922 (4.578) | −0.01982 (−1.723) | −0.0002541 (−0.014) |
| cases | 1949 | 1489 | 460 | 345 |
| $R^2$ | 0.6999 | 0.7308 | 0.7407 | 0.3974 |
| TP[b] | 10.71 (44,710) | 12.88 (391,400) | 10.26 (28,510) | 10.06 (23,420) |

*Notes:* The dependent variable is the natural log of carbon emissions *per capita*. For the coefficient $t$ statistics are reported in parenthesis. Country-specific dummies are included in all equations.

[a] Node numbers correspond to the node numbers in Figure 5.6.

[b] Turning points values in US dollars are reported in parenthesis.

(2005) also found that an inverted-U for $CO_2$ is likely to exist for several, but not all, countries.

## 5.13 Conclusion and Further Studies

In this Chapter using a combination of parametric of nonparametric techniques, we reject the linear model commonly used in the previous empirical literature in favor of a multiple regime alternative in which different countries obey different models. We demonstrated that regime differences explain an environmentally and statistically significant proportion of environmental degradation. We find fundamental differences in the relationship between growth and environment between developing and developed countries. Using the Kuznets curve metaphor, we find that some rich countries may already have passed a turning point and begun to see improvements in the environment with additional growth while for most others, while most others are becoming increasingly polluted.

We find that the impact of openness to foreign markets varies according to the level of development, trade policies, and the productive structure of an economy.

This approach suggests that rapid growth of developing countries that is not accompanied by significant structural changes could have devastating consequences on the environment. In addition, evidence to support pollution displacement from rich to poor countries via international trade and foreign direct investment is reported. This suggests that in the absence of coordination across countries in environmental policy, overall world environmental quality will fall with trade. There is some evidence that China has mitigated some of the negative environmental consequences by adopting new technology from developed countries through FDI. In particular, Gallagher (2003) finds that China is adopting cleaner vehicle technology from the United States. Zhang (2000) finds that the decline in energy intensity in China almost halved the increase in emissions that would otherwise have occurred. It is the responsibility of developed countries to assist developing countries by sharing and facilitate the use of new and cleaner technologies through investment and trading and in promoting better environmental standards.

There are some important caveats. Parameter heterogeneity might reflect the impact of omitted pollution determinants. Nonlinearities in the relationship that cannot be easily captured by parametric models can also produce heterogeneous parameters. The first problem is partially addressed by adapting the tree regression to perform a panel data estimation. The second is mitigated by splitting the sample. Another limitation of these methods based on Breimans' (Breiman et al., 1984) tree regression is that they have no known distribution theory. In Chapter 6,

we will apply to the EKC the threshold estimation and testing procedure developed in Hansen (2000) that overcomes this limitation. It would be of interest for further study, to include other pollutants.

It's clear that the variable used in this study for capital intensity behaves anomalously, as for richer countries we find that more capital intensity significantly reduces emissions. It is often assumed that capital intensity translates into pollution-intensity. This seems too simplistic. There appears to be the need to control for the level of "dirtiness" of the industry to improve our analysis. This variable seems highly correlated with income, as Panel 5.3, showing scatterplots of the log of emissions against the log of capital stock per worker, clearly illustrates. They patterns are very similar to the ones in Figure 5.3 for the GDP variable. Also, in order to improve the comparability of this study with others, it would be beneficial to re-estimate the models using capital abundance adjusted for differences in worker's productivity, as done in Antweiler et al. (2001).

Clearly more work needs to be done to fully understand the role of international trade and foreign direct investment in mediating the relationship with the environment.We believe that the next important empirical step for this line of work, after having identified those pollutants and countries having similar economic/environment relationships using a nonparametric approach, is to formally test through parametric methods, the importance of the factors identified in this study. Such level of detail will allow to test more appropriately alternative theoretical specifications, investigate dynamic relations over time, and enable researchers to draw more specific and useful policy implications.

Chapter 6

# The Relationship Between Growth and Environment: Should we be Looking for Turning or Break Points?

## 6.1 Introduction

This Chapter purports to explore the existence and nature of an empirical "law" of development and environmental economics by means of nonparametric techniques. The empirical law features two variables of considerable interests to economists and policy makers, namely an indicator of environmental quality and the level of per capita income. The link between these variables takes the form of an "inverted-U" shaped curve in the pollutant/income space and is referred to by the literature as the Environmental Kuznets Curve (EKC, hereafter). Environmental degradation will increase with income at low levels of income, reach a peak and then decrease with income at high levels of income. After the seminal papers by Grossman & Krueger (1993a, 1995), and by Shafik (1994b), this relationship has attracted considerable interest and today is one of the most lively research lines in development and environmental economics.

Several *ad hoc* explanations have been proposed to justify this empirical law. Some economists have stressed the impact of structural changes in the economy, others the link between demand for environmental quality and income, international trade, technologies improvement, and policies. For a comprehensive review of this literature see Panayotou (2000).

If testing for the possible determinants of the EKC has been a quite popular exercise in the literature, surprisingly, less attention has been devoted to the econometric and methodological problems arising from the quantity and quality of data. Stern & Common (2001) pointed out that environmental data are "patchy in coverage, and poor in quality." Also, most of the empirical work is based on the parametric approach. Only recently, Taskin & Zaim (2000, 2001), have suggested the use of non parametric methods to test the existence of an EKC. The nonparametric approach should be more suitable than a parametric one because of its flexibility, as there is no need to specify an a priori functional form but by letting the data reveal the shape of the relationship. Adding non-linear terms in a parametric framework, a popular solution for this problem, may also not be appropriate. This standard approach based on the linear model suffers from a few drawbacks, especially when studying the EKC relationship.

(i) Polynomial function have all orders of derivatives everywhere. This property might smooth out important features such as an asymmetric behavior around the turning points. We think that we should not only be inter-

ested in determining the location of turning points but also whether the behavior of an up swing following a down swing is symmetric. Asymmetric behavior around a turning point, besides having important consequences for the policy maker as such, might also indicate the presence of different factors affecting the downward and the upward branch of the curve. Stern & Common (2001) have pointed out that trade might play an important role in explaining the downward part of the EKC for developed countries. Panayotou (2000) after examining the evidence form Vincent (1997) and Carson et al. (1997b) concerning the existence of a Kuznets curve within individual countries concludes that "whereby rising incomes result in a more effective regulatory structure by changing public preferences and making resources available to regulatory agencies. States with low-income levels have a far greater variability in emissions per capita than high-income states suggesting more divergent development paths. This has the implication that it may be more difficult to predict emission levels for low-income countries approaching the turning point."

(ii) The polynomial degree cannot be finely controlled. Regression concerning the EKC are basically polynomials of second or third order. Usually we are interested in discriminating between an inverted U and an N shape. Nonparametric regressions do not have this built in constraint. We will exploit this particular feature to device a procedure to test nonparametrically the inverted-U versus the N shaped EKC hypothesis. The test is in the spirit of the bootstrap based on Silverman's (1981) test of multimodality of a probability density function, and its adaptation to testing monotonicity in a nonparametric regression by Bowman et al. (1998).

(iii) Harbaugh et al. (2002) after re-examining the empirical evidence for the EKC for three local pollutants, i.e., sulfur dioxide, smoke, and total suspended particles (TSP) using a more representative data set, found that estimates are extremely sensitive to the sample chosen and the econometric specification. In particular, they found that for cubic polynomials very small changes in estimated coefficients, translate into large changes in the shape of the estimated relationship. They pointed out that the problem is aggravated by highly correlated independent variables and suggested that nonparametric can allow for nonlinearities without making use of functional

forms with correlated polynomial terms.

There are two main parts in this Chapter. The first part uses nonparametric regression to explore the relationship between economic growth and the environment. We device a procedure to test nonparametrically the inverted-U versus the N shaped EKC hypothesis. The test is in the spirit of the bootstrap based Silverman's (1981) test of multimodality of a probability density function, and of Bowman's et. al. (1998) adaptation of this to testing monotonicity in a nonparametric regression. The second part uses a threshold model as a more parsimonious nonparametric function estimation strategy. This approach will allow to formally test hypothesis concerning the relationship between economic growth and the environment that have emerged from the nonparametric regression approach and that have been referred to frequently in the EKC literature, but that have never been formally tested.

The first part of the Chapter is organized as follows. Section 6.2 introduces the econometric and theoretical arguments that justify the nonparametric approach. Section 6.3 addresses methodological issues specific to the application of nonparametric regression to the estimation of an EKC curve. Successful empirical modeling and the choice of appropriate statistical techniques come from careful consideration of the economic theory behind the problem and the quality of the measured data. Section 6.7 discusses the test. Section 6.8 presents and discusses the econometric results. The second part of the chapter is organized as follows. Section 6.9 the models threshold estimation and testing methodology are introduced. Section 6.10 presents the estimation results. Section 6.11 concludes.

## 6.2 Environmental-Economic Regimes

The processes of economic growth and environmental change are clearly complex and evolving over time. Identifying all the complex interactions and feedback relationships that are expected to play a significant role in the evolution of these processes may be an impossible task at this point in time. One important assumption underlying the majority of cross-country pollution studies is that all countries obey a common linear model specification. Because of the inherent complexity of the environment-economy interaction, our limited knowledge of it, and the often poor quality of data, this assumption appears at best as a crude approximation.

Limits in our econometric models can reveal themselves as apparent structural change. Identifying these structural changes could further our understanding of the links between the economy and the environment.

Besides econometric arguments, recent theoretical developments in modeling the relationship between income and the environment also imply the existence of different regimes. A simple and frequently used explanation for the EKC is based on a traditional demand-and-supply analysis. A possible way to obtain an inverted-U shaped EKC consistent with a demand-and-supply framework is to suggest that the EKC reflects a demand for environmental quality. Assuming that environmental quality is a normal good, pollution may at first rise with income, but eventually fall as income continues to rise. More formal developments can be found in Lopez (1994) and Copeland & Taylor (2003). The resulting EKC from these models is graphed in Figure 6.1. Other models based on traditional economic theory, such as the one by also predicts a smooth EKC curve for a technology with increasing returns to scale.

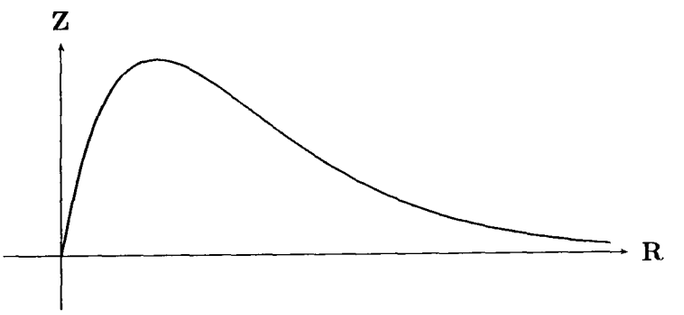



Figure 6.1: EKC generated by income effects

Several recent papers have attempted to explain the EKC relationship by introducing threshold effects in modeling either pollution abatement. (see. e.g.. Jones & Manuelli, 1995), or environmental policy regulation (see, e.g., Stokey, 2001). Threshold effects lead to a very different relationship between environmental quality and income during early stages of economic development as opposed to later stages. For instance the abatement-threshold model predicts a kink in the relationship between pollution and income, as shown in Figure 6.2. The policy threshold model predicts an even more drastic changes in regime, and produce discontinuous

EKC with a discrete drop in pollution and income once the threshold is reached. The main advantage of the policy model over the abatement model, is that the policy model generalizes to the multiple good case. It is noteworthy that both models rely on strong policy responses to increases in income in the development process.

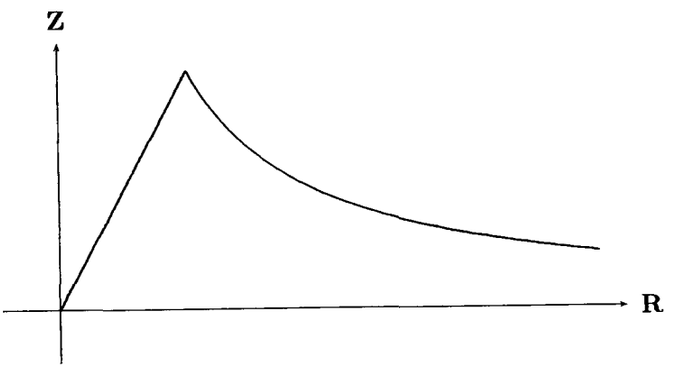



Figure 6.2: EKC generated by threshold effects model

## 6.3 Nonparametric Regression

Let $(X_i, Y_i)$. $i = 1, \ldots, n$, be a random sample from an unknown bivariate population distribution $f(x, y)$. Econometrics frequently focuses on the conditional expectation function $m(x) = \mathrm{E}(Y|X = x)$, where x is some fixed value of X. We can write

$$Y_i = m(X_i) + u_i, \qquad i = 1, \ldots, n,$$

where $u_i$ is an independent random error satisfying $\mathrm{E}(u_i|X_i = x) = 0$. It is not necessary that the conditional variance is a constant function. Typically one assumes

$$\mathrm{Var}(u_i|X_i = x) = \sigma^2(x).$$

The standard assumption made in econometrics that $m(x) = \alpha + \beta x$ implies certain strong assumptions about the data generating process. If f is a bivariate normal density than it can be shown that the mean of the conditional density of $Y$ given $X$ is linear

$$\mathrm{E}(Y|X = x) = \alpha + \beta x.$$

There are many ways to obtain a nonparametric regression estimate of $m$ (see Wand & Jones (1995) and for a few examples). In this study we consider the two important families of estimators and their suitability to estimate the EKC.

The most popular estimator, proposed independently by Nadaraya (1964) and Watson (1964) (denoted NW thereafter), can be derived from the definition conditional expectation

$$m(x) = \mathbf{E}(Y|\ X = x) = \int y f(y|x)\, dy = \int y \frac{f(x,y)}{f_X(x)}\, dy, \tag{6.1}$$

where $f_X(x)$, $f(x,y)$, and $f(y|x)$ are the marginal density of $X$, the joint density of $X$ and $Y$, and the conditional density of Y give X, respectively. An intuitive approach for estimating $m(x)$ is to substitute the unknown joint and marginal densities in eq. 6.1 with appropriate kernel estimators.

The NW estimator obtained this way is

$$\hat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}\right)}$$

The alternative estimator considered in this investigation is the local linear estimator, whose better properties have been established only more recently (Fan, 1992, 1993; Hastie & Loader, 1993). To find the estimate of $m$ at a particular point $x$ it fits a regression line by weighted least squares, using weights coming from the height of a kernel function centered at $x$. Observation closer to $x$ are accorded greater weight. This method belongs to the more general class of estimators known as *local polynomial* regressions. Another popular member of this class is the Cleveland's LOESS estimator (see, Cleveland, 1979). Formally the local linear regression estimate of m(x) at point $x$ solves the least squares minimization problem

$$\min_{a,b} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) (Y_i - a - b(x - X_i))^2 .$$

Note that the NW estimator can be seen as solving the following minimization problem

$$\min_{a} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) (Y_i - a)^2 .$$

Econometrics is the application of mathematical statistical techniques to investigate an economic problem using economic data. Successful empirical modeling and the choice of appropriate statistical techniques come from careful consideration of the economic theory behind the problem and the quality of the measured data. In fact, we will show how the nature of the economic relationship and the quality of environmental data can considerably impact estimates and therefore the implied policy recommendations. In particular, we will concentrate on the concave nature of the EKC and on the problem of environmental data quality and their impact on the nonparametric estimates. Environmental data availability and quality, though improving with time, remains an important problem in investigating the existence of the EKC. The use of nonparametric regression techniques insures that missing or less accurately measured observations do not affect distant parts of the estimated curve as much as the parametric estimator would. However, even nonparametric methods are not immune to problems. We will see how the asymmetric nature of the data, in the sense that most environmental data come from the most industrialized countries, can affect a nonparametric estimator. In particular, we will see how bias problems resulting from data asymmetry affects more seriously the Nadaraya-Watson estimator, the standard nonparametric regression estimator (for an application of this estimator to the ECK see, Taskin & Zaim, 2000).

## 6.4 Bias in Nonparametric Regression

Bias in estimating the EKC, whether originating from the nature of the EKC relationship or the environmental data quality, by NW has two important effect. The first makes the identification of the curve more difficult. The bias has the effect of "attenuating" the estimated EKC. The bias also affects location and height of the turning point where a EKC relationship is found. Table 6.1 reports the pointwise asymptotic bias and variances for the NW and the Local linear estimator.

One first important observation is that given that the variances of the two estimators are the same, the local linear estimator is expected to perform better. If we compare the bias of the Nadaraya-Watson estimator with the local linear estimator we note that both depend on $m''$ whereas only the NW because of the local constant fit depends on $m'$ and $f'/f$. When $|m'|$ or when $f'/f$ are large, i.e. when the slope of the curve is high or when data are highly grouped, then the

Table 6.1: Bias and Variance of Kernel and Local linear smoothers (Fan, 1992)

| Est. | Bias | Variance |
|---|---|---|
| N-W | $\left(\frac{1}{2}m''(x)+\frac{m'(x)f'(x)}{f(x)}\right)\int_{-\infty}^{\infty}u^2K(u)duh_n^2$ | $\frac{\sigma^2(x)}{f(x)nh_n}\int_{-\infty}^{\infty}K^2(u)du$ |
| Loc. lin. | $\frac{1}{2}m''(x)\int_{-\infty}^{\infty}u^2K(u)duh_n^2$ | $\frac{\sigma^2(x)}{f(x)nh_n}\int_{-\infty}^{\infty}K^2(u)du$ |

bias of NW is also large. Because the Kuznets curve would be a concave function of GDP. the negative $m''$ term implies that the the curve is biased downward no matter which of the two estimators we use. The $m'$ bias component of the NW estimator being positive and then negative respectively in the ascending and descending part of the curve would tend to attenuate the estimated curve. These are asymptotic results. The bias would tend vanish as the sample size grows and the bandwidth smaller. Unfortunately large datasets are usually not easy to come by.

Figure 6.3 illustrates the bias caused by the asymmetry of observations and the slope of m of the NW estimator. Since most observations are on the right of the point we are trying to estimate (0.3). the estimate is biased upward. This problem is aggravated at the boundary regions. Suppose that the observations are confined to the [0,1] interval and that we are trying to estimate m(0).The figure also shows how at 0, where the slope is positive, the local average is considerably biased upward. Therefore another source of bias that "attenuates" the EKC stems from the fact that in practice we have a bounded support. When estimating the regression at the leftmost observation, only points that are on the right can be included, so that if the regression function is positively sloped, as we expect for the EKC, there will be an upward bias at that point.

Figure 6.4 shows how with equally space observation these biases are visibly reduced. Economic data being of a non experimental nature depart considerably from this ideal design. Economic data tend to be clusterd. This will also be illustrated in the practical application.

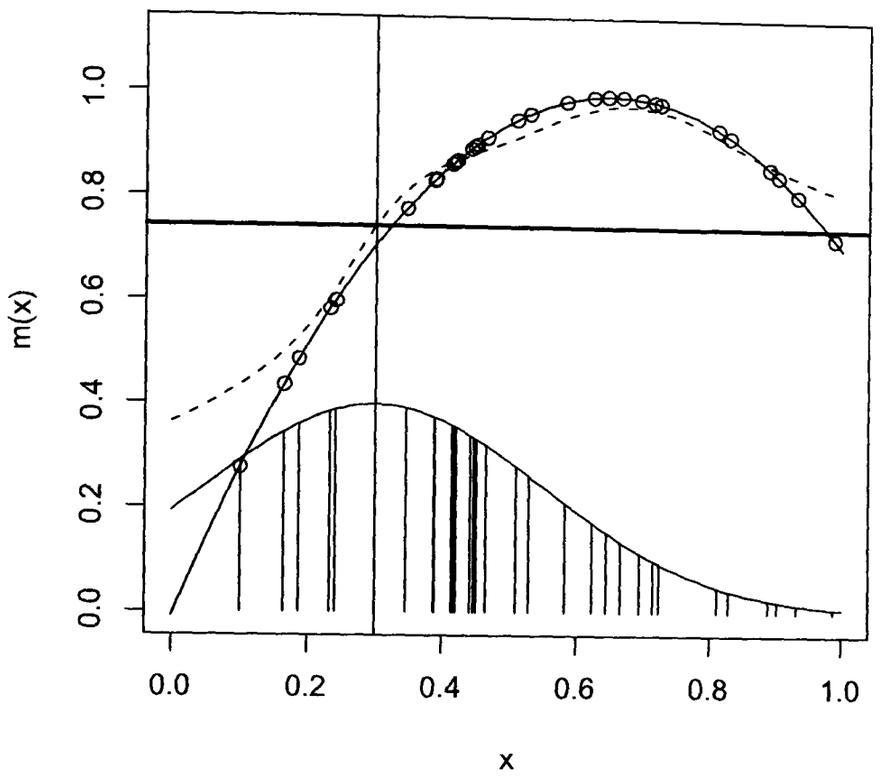



Figure 6.3: Combined effect of the slope of the mean function and the asymmetry of the observations on the Nadaraya-Watson estimator. Suppose we observe the data indicated by the circles on a quadratic $m(x)$. The data are shown with no noise to simplify the illustration. We estimate $m(0.3)$ using the locally constant NW fit (represented by the horizontal thick line) using the normal kernel shown at the bottom of the picture.

## 6.5 Potential Impact of Bias on Turning Point and 'Environmental Price'

In this section we illustrate the consequences of the NW bias induced by the combination of slope of the mean function and the boundary effect on the location and height of the EKC turning point. We will follow the convention established by the existing literature on the EKC which is to compute the turning points from the estimated functional relationship. In the existent EKC studies, the estimation of the turning point has been widely proposed. The reason is twofold: "Estimated of per capita income associated with the turning point can be compared with the actual income levels of the observed dataset, thus indicating whether the turning

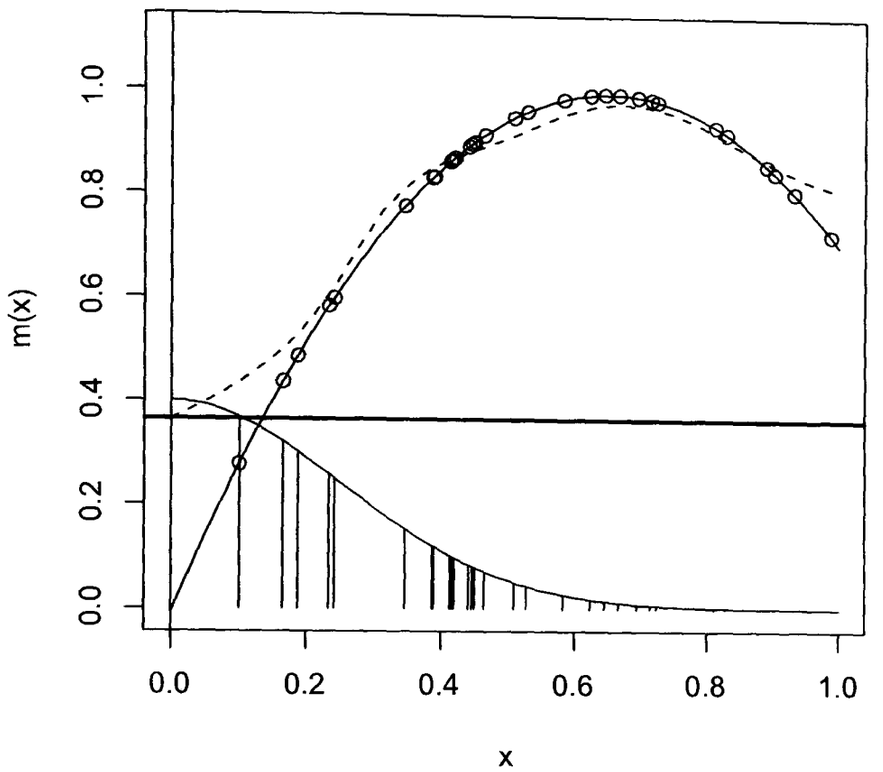



Figure 6.4: Effect of boundary bias on the Nadaraya-Watson estimator. We estimate $m(0)$ using the locally constant NW fit when all the data are within the $[0, 1]$ interval.

point income falls within or outside the observed income range. Analysis of stability of the turning point can also shed light on the reliability of the EKC estimates" Barbier (1997).

Furthermore, if there exists a threshold level of per capita income after which economic growth "sow the seeds" for the improvement of the environmental quality is important to know it. If the estimation, and the consequent considerations, of the turning point has been a popular practice in the EKC literature, surprisingly not the same can be said for the height of the curve. Of course, the implications of estimation of the height of the EKC, are not trivial issue. Following Panayotou (1997), it specifies the 'environmental price' of economic growth. So that it represents the maximum stress that must be carried out by the environment before experiencing an environmental improvement path. So underestimating the height may have serious consequences to some ecological threshold (see, Arrow & others, 1995, and Munasinghe, 1999). Following the above definitions, if $\widehat{m}(x)$ is an es-

timator of the EKC, the nonparametric estimators of the turning point and the environmental price can be defined respectively as the interior global maximum,

$$\widehat{TP} = \arg \max_{x \in (x_1, x_n)} \widehat{m}(x),$$

and

$$\widehat{EP} = \max_{x \in (x_1, x_n)} \widehat{m}(x).$$

We assume that the curve has bounded support and is defined on $[x_1, x_n]$. Figure 6.5 illustrates the effect of the boundary bias on the estimated turning point and environmental price. The turning point in the example is close to the right boundary. In this case, the combined effect of downward bias caused by the curvature and the upward boundary bias shifts the turning point to the right. The estimated environmental price is lower then the true one.

The consequences of this bias can be quite serious. Suppose, for example, the curve was estimated by using a cross section sample containing mostly rich countries, not a particularly contrived situation since more reliable data are available for these countries. These countries might be situated mostly on the downward part of the curve. Under these circumstances the turning point and the associated level of pollution, the environmental price, could be seriously underestimated. Figure 6.6 exemplifies this scenario. If these findings where employed for policy implication for poorer countries the consequences could be serious. Learning from the experience of the most industrialized countries when using an inappropriate estimator could be seriously misleading. We will employ an applied example to see whether these problems could significantly affect estimates.

## 6.6 Nonparametric Estimation of the Kuznets Curve Example

We will employ an applied example using 1990 cross-section data from World Resources Institute (around 160 countries) to see whether the aforementioned problems could significantly affect estimates. As environmental quality indicators we

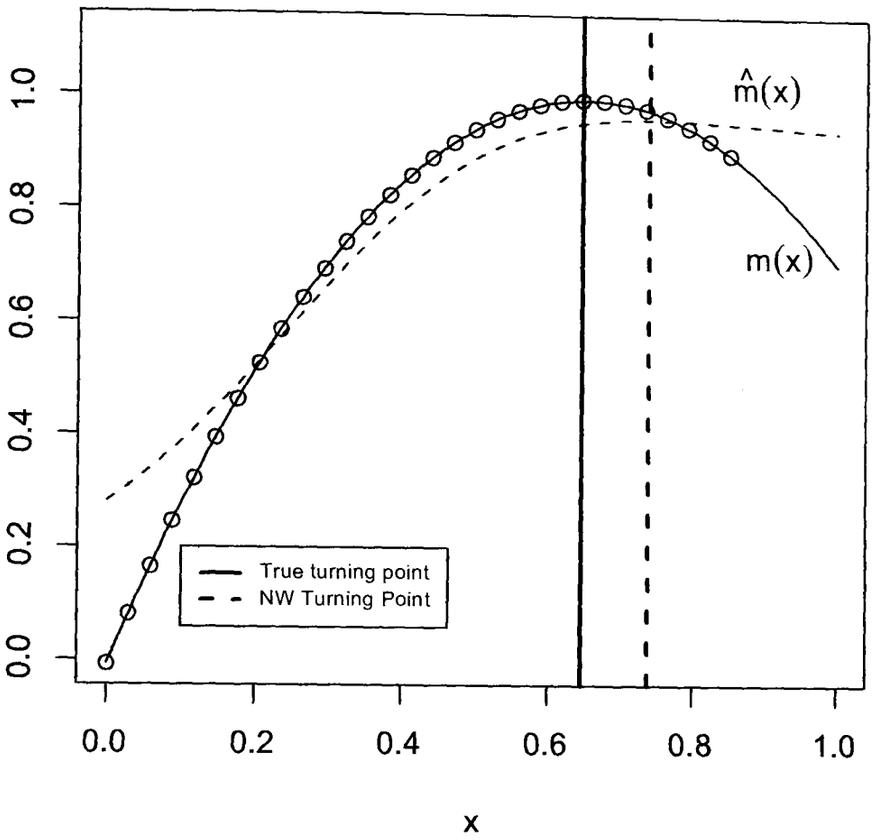



Figure 6.5: Combined effect of curvature of the mean function and boundary bias of the Nadaraya-Watson estimator on the estimated turning point.

have taken the emission per capita of three important air pollutant: sulphur dioxide ($SO_2$), carbon dioxide ($CO_2$) and nitrogen dioxide ($NO_x$). Our analysis is more focused on the first pollutant because it shows a clear bell shaped curve. $SO_2$ is a pollutant which action is mainly local (urban smog). $SO_2$ is emitted largely from burning coal (for heating purposes) and high-sulfur oil. Figures 6.8, 6.10 and 6.11 present the results of the nonparametric estimation[1]. Sulfur dioxide is the only pollutant among those considered here that displays a clear inverted-U relationship with per capita income. Figure 6.8 shows the Nadaraya-Watson and

[1]For the Local polynomial estimate we use direct plug-in methodology to select the bandwidth of a local linear Gaussian kernel regression estimate, as described by Ruppert, Sheather and Wand (1995) implemented in their own S library. For the NW estimate we use the technique of cross-validation to select a smoothing parameter as provide by the *sm* R library by Bowman, A.W. and Azzalini, A. (1997).

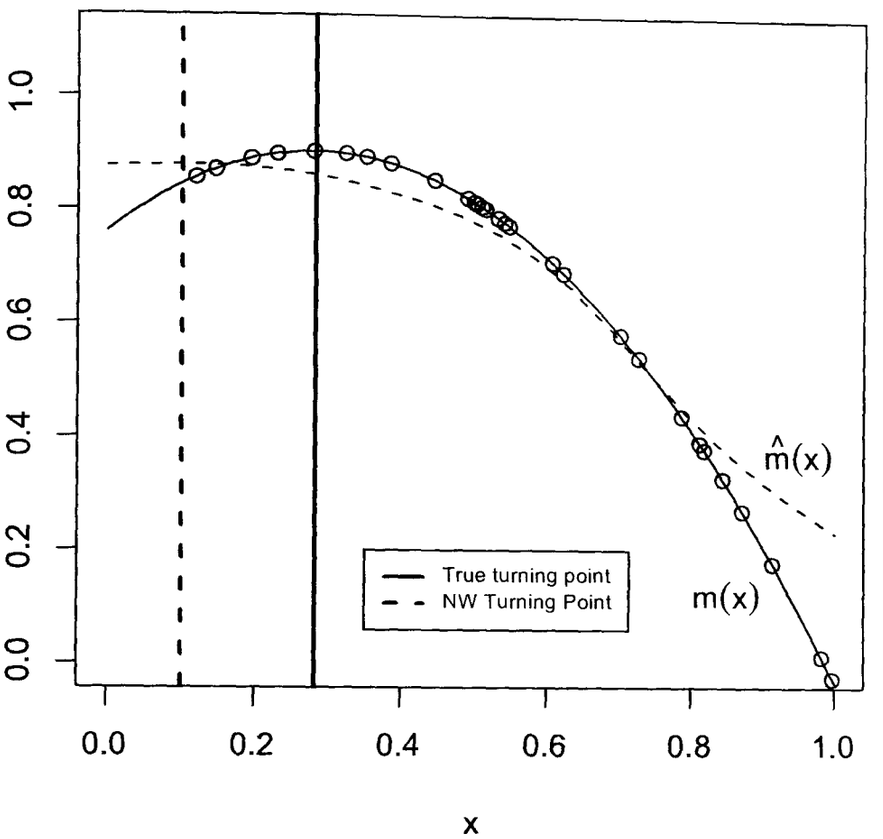



Figure 6.6: Effect of slope and boundary bias of the Nadaraya-Watson estimator on the estimated turning point. Points are a random sample from the uniform distribution. If the true turning point is located at low level of income the estimated turning point will be shifted to the left.

the Local polynomial estimate on the same graph. It is clear from the picture that the attenuating effect causes the NW estimate to be flatter than the locpoly estimate. A clearer illustration of this effect is provided by Figure 6.9 which shows the difference of the two curves. The difference is smoothed using a gaussian kernel with a bandwidth of 0.5 to enhance the interpretability. The shape is close to an inverted-U. This is consistent with the attenuation bias of the NW estimator hypothesis. One of the most important feature in Figure 6.9 is that the upward branch of the curve is considerably less prominent then the downward one. Based on the previous discussion this can be explained by the concentration of the rich industrialized countries in that branch. The negative slope of the curve and the concentration of countries determines a downward bias that partially compensate

the positive boundary bias. The picture also shows that the NW estimate between 1,851 and 23,522 dollars of per capita income differs by as much as 13 per cent.

From the picture it is clear that the asymmetric nature of the data in the sense that there are mostly high income countries and that they are mostly on the descending part of the curve. This concentration of high income countries on the descendent part of the curve seems to be responsible of the difference between the local polynomial and the NW estimate. The NW estimate of the turning point and the level of pollution associated with it is is lower then the locpoly one.

Table 6.2 reports the estimated turning points and the associated 'environmental price' for the two estimators considered. Since the variable are in logs the difference between two values given by different estimation methods gives an approximation to the percentage change of estimated concentration level that results from changing estimator correspondingly. The NW estimator gives a turning point that is more than 6 per cent smaller then the one obtained from locpoly estimator. Also, The associated environmental price of the NW estimator is more than 8 per cent smaller then the one computed from the local polynomial estimator. These observations provide evidence that in an actual example the bias is considerably affecting the estimates in agreement with the theoretical predictions.[2]

## 6.7 Nonparametric Testing the Inverted-U Vs. the N shaped EKC Hypothesis

We are interested in testing whether the Kuznets curve exists and what shape it takes, namely whether it is of an inverted-U shape or N shaped. Figure 6.8

[2]A rugplot is added to aid the interpretation. The data have been jittered to avoid mark's overlapping.

Figure 6.7: Local Polynomial and Nadaraya-Watson estimate for the $SO_2$

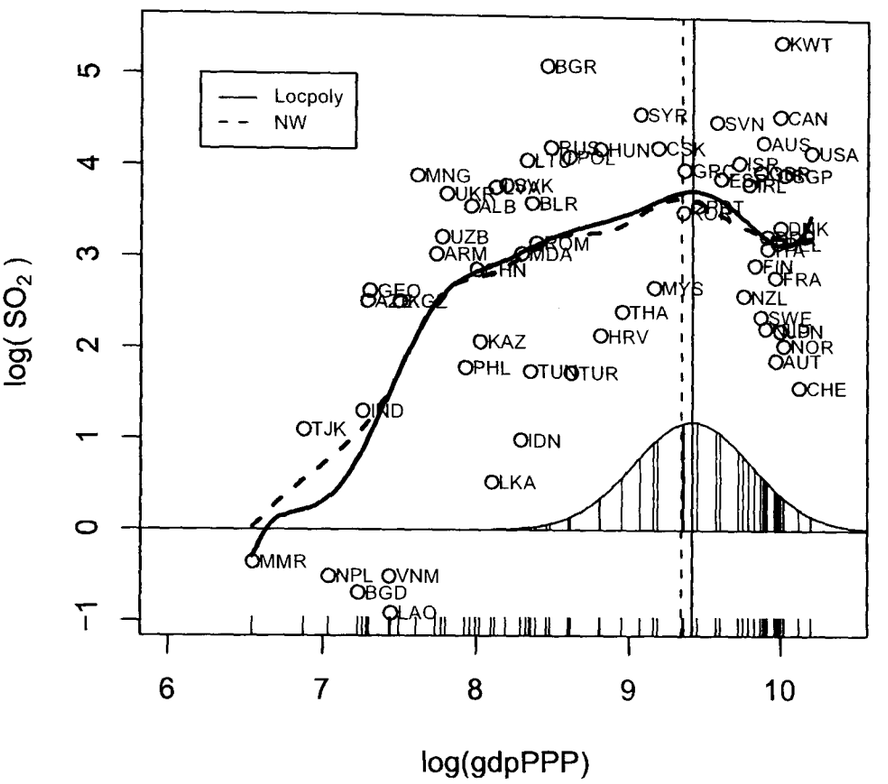



Figure 6.8: Local Polynomial and Nadaraya-Watson estimate for the $SO_2$. The two turning points data on the estimated turning point. The NW estimator assigns weights proportional to the heights of the rescaled kernel. A rugplot, which adds a mark for each observation on the x-axis, is added to aid the interpretation. The data have been jittered (a small amount of noise has been added to the data) to avoid mark's overlapping. The ISO-3166 3-letter identifications code has been used to label the countries. If the true turning point is located at high level of income the estimated turning point will be shifted to the left.

suggests the possibility that it might be N shaped. The question of interest is whether the N shape for the $SO_2$ reflects the shape of the underlying EKC or its caused by random fluctuations. Taskin and Zaim (2000) test the existence of the curve by testing whether the vector of partial derivatives of the conditional mean is equal to zero versus the alternative that it is not using Hotelling's $T^2$ test. This approach has at least two limitations. Firstly it is parametric and secondly does not allow to test for the shape of the EKC. We will approach the existence and shape problem in an unified way. The N shape is characterized by a second turning

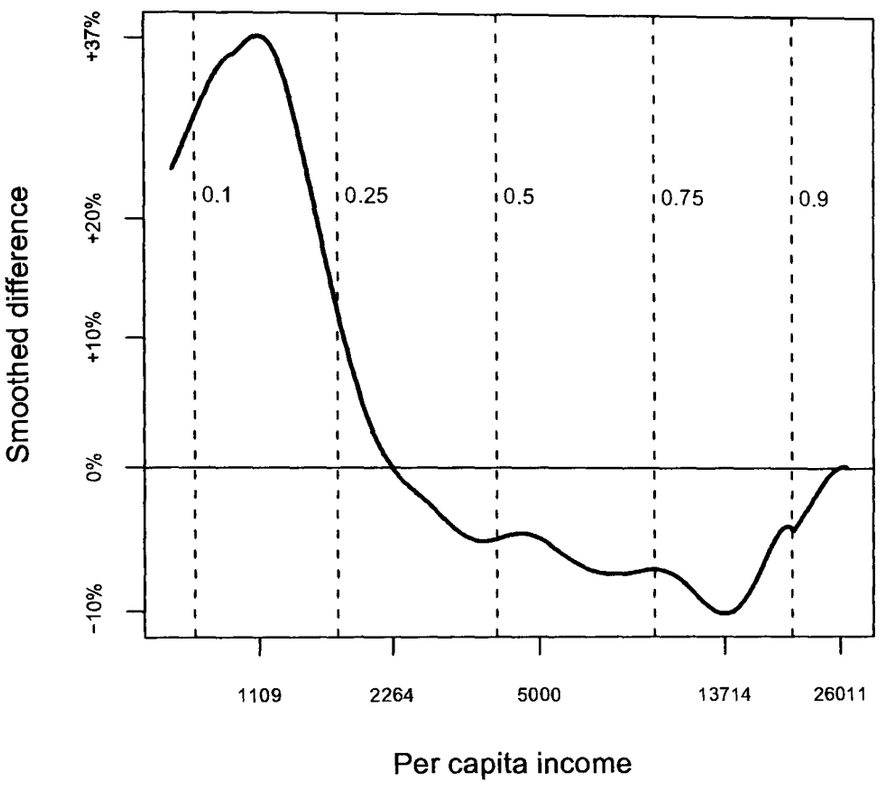



Figure 6.9: Smoothed difference between the Nadaraya-Watson and the Local polynomial estimates.

point. In order to proceed we need to define a nonparametric estimator for the second turning point. If we relabel the estimator for first turning point a $\widehat{TP}_1$, we can define the estimator for the second turning point as (the interior global minimum after the first turning point)

$$\widehat{TP}_2 = \arg \min_{x \in (\widehat{TP}_1, x_n)} \widehat{m}(x),$$

We device a procedure to test nonparametrically the inverted-U versus the N shaped EKC hypothesis. The test is in the spirit of the bootstrap based Silverman's (1981) test of multimodality of a probability density function, and of Bowman's et. al. (1998) adaptation of this to testing monotonicity in a nonparametric regression. To test for the inverted-U shape EKC hypothesis the idea is to see whether a relatively large $h$ is required to force an N shaped $\widehat{m}$ to an inverted-U

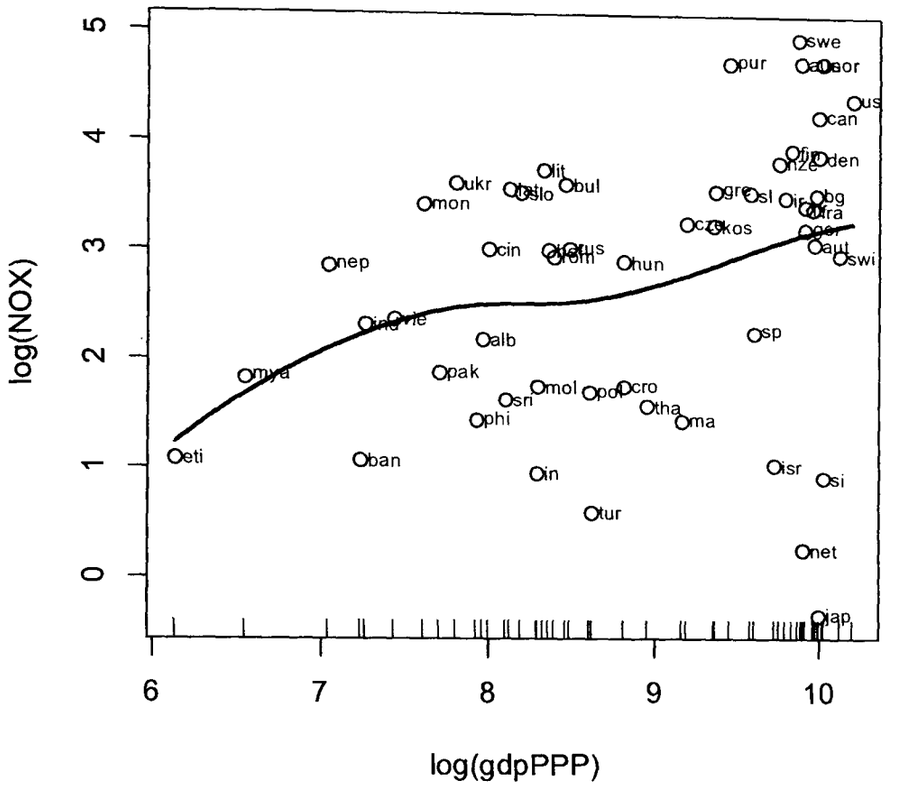



Figure 6.10: Local Polynomial estimate for the $NO_2$.

shape.

As $h$ tends to infinity, the estimated curve tends to the least square regression line. This fact alone does not guarantee that the number of turning points is a monotone decreasing function of $h$. In fact, for the local polynomial estimator, this is indeed not the case. We will take comfort from the fact that Bowman et. al. (1998) find that departures from monotonicity are extremely unusual.

The test statistic is the critical bandwidth defined as,

$$h_{crit} = \min_{h>0}\{h|\ \widehat{m}(x;h) \text{ is of an inverted-U shape}\}.$$

Once for given observations we have computed $h_0$ we need to decide whether it is a "surprisingly" large value for the statistic $h_{crit}$. In order to do this $h_0$ has to be assessed against a suitable null sampling distribution.

A suitable choice of null sampling distribution should posses the following

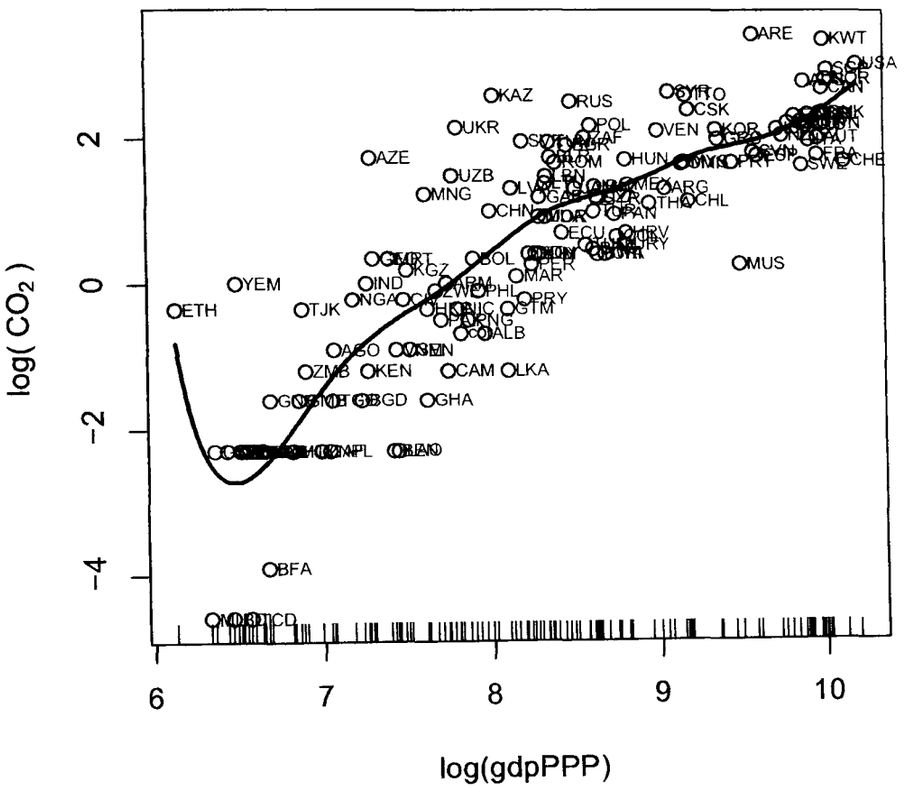



Figure 6.11: Local Polynomial estimate for $CO_2$.

desirable properties:

(i) The density of H should be such that $\hat{m}$ is of an inverted-U shape.

(ii) Subject to (1) the density of H should produce a plausible shape of $\hat{m}$ given the data, since, for example, large values of h would be from very flat inverted-U shaped curves.

(iii) among the densities satisfying (1) and (2) we should consider the "worst" of the infinite possibilities under the null, i.e., that alternative that would make the decision between an inverted-U shape and an N shape a most difficult one. Clearly, the the decision would be more difficult if $\hat{m}$ was the most nearly N shaped, amongst the infinite inverted-U shaped curves.

In order to determine the sampling distribution of H under the null of inverted-U shape we should consider the "worst" of the infinite possibilities under the null,

Table 6.2: Turning Points and Environmental Prices by Estimator

| | $\widehat{TP}$ (log) | $\widehat{EP}$ (log) | $\widehat{TP}$ (\$) | $\widehat{EP}$ (tons) |
|---|---|---|---|---|
| Locpoly | 9.416488 | 3.754450 | 12289.4 | 42.7 |
| Nadaraya-Watson | 9.353129 | 3.670684 | 11534.9 | 39.3 |
| Difference | 6.335975 % | 8.376582 % | 754.5 \$ | 3.4 tons |

i.e., that alternative that would make the decision between an inverted-U shape and an N shape a most difficult one. Clearly, the the decision would be more difficult if m was the most nearly N shaped, amongst the infinite inverted-U shaped EKC's.

Bootstrapping is used to provide a null distribution for the test statistic. Table 6.3 gives the critical bandwidths and P-values for the bootstrap test of the null hypothesis that EKC is of an inverted-U shape against the alternative that it is of an N shape. Using 10000 replication we find that the inverted-U shape cannot be rejected against the N shaped alternative hypothesis with a p-value of 0.326. This finding agrees with Shafik's (1994) parametric findings. Others (Grossman and Krueger, 1993, for ex.) have found weak evidence of an an N shaped EKC for $SO_2$. Though with the available data we cannot statistically detect the renewed positive relationship between per capita income and $SO_2$, it is remains a substantively important feature in our estimate that, because of its policy implications, cannot be easily dismissed. The high variability in the data and the conservative nature of these kind of tests Silverman (1983) might considerably bear upon the results. The table also reports a test of monotonic EKC versus a inverted-U shaped one. The monotone null is rejected at the 10 per cent significance level.

Table 6.3: Critical Bandwidths and their Estimated P-values

| EKC Hypothesis | $h_{crit}$ | P-value[a] |
|---|---|---|
| Monotone Vs. Inverted-U | 0.93 | 0.088 |
| Inverted-U Vs. N shaped | 0.24 | 0.326 |

[a]Ten Thousand replications were used to obtain the approximate null distribution.

## 6.8 Nonparametric Elasticity and Asymmetric Behaviour Around the Turning Point

Since the variables are in logarithms the derivative of the EKC has an important economic interpretation, the elasticity with respect of per capita income of the environmental quality indicator. Extending the idea of local polynomial fitting, one can estimate $\varepsilon(x)$ as the slope of the local polynomial fit.

$$\varepsilon(x) = b(x)'^{T} \left(B^{T} W(x) B\right)^{-1} B^{T} W(x) Y_{i}.$$

Figures 6.12 and 6.13 present the results of the nonparametric estimation of the elasticities. Sulfur Dioxide's nonparametric elasticity is relatively elastic for levels of income below the median and relatively inelastic for levels above the median. This finding is consistent with with the parametric elasticity found for Sulfur Dioxide by Shafik (1994b, p. 766). The nonparametric elasticity shows another interesting feature not identifiable in the parametric estimates. There is a "kink" at the turning point of the curve. Before the turning point the the elasticity changes very slowly whereas after reaching the turning point elasticity starts to decrease at a higher (more than double) rate. This is consistent with Panayotou's conclusions (2000). The curve appears to be flatter before the turning point then after it.This could be evidence of regime differences between countries on the increasing part of the curve and countries on the decreasing part. We attempt to verify theses preliminary findings in the following sections using a threshold approach, a more parsimonious nonparametric function estimation strategy. This approach, will allow to formally test parametrically some of the hypotheses concerning the relationship between economic growth and the environment.

## 6.9 Threshold Model Estimation and Testing Methodology

In this section we review the threshold estimation and testing procedure developed in Hansen (2000) that will be applied in this chapter to the EKC debate. Com-

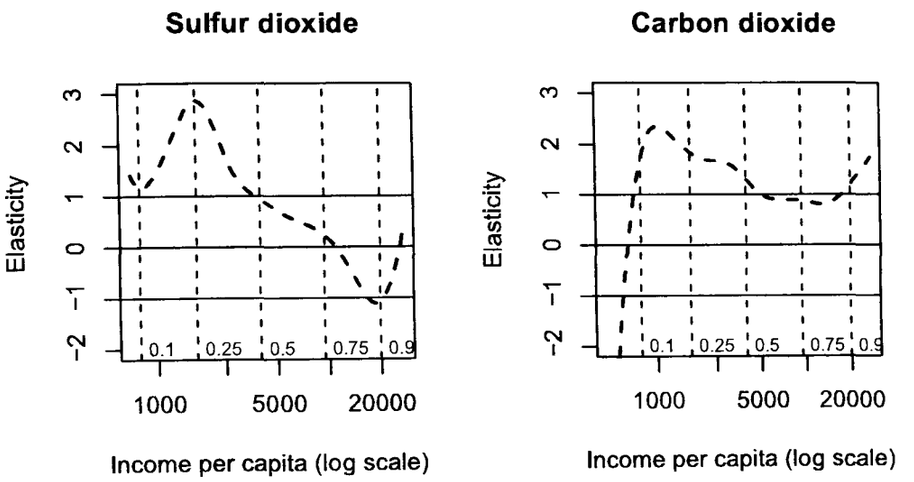



Figure 6.12: Changes in environmental elasticities with income.

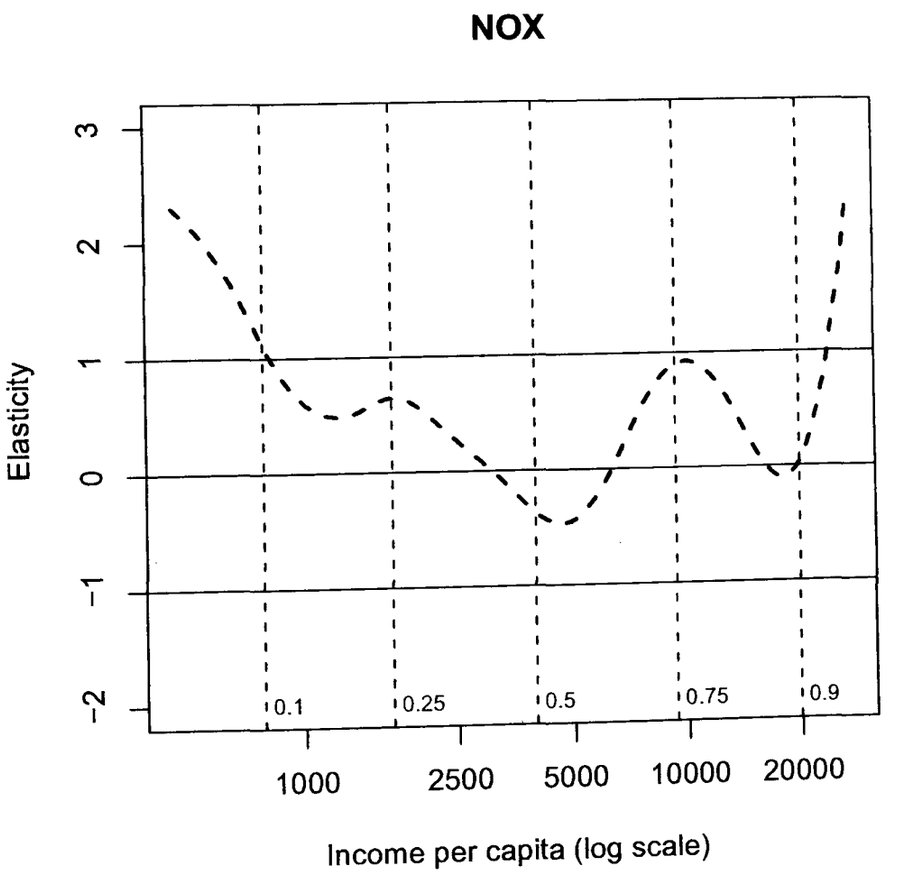



Figure 6.13: Elasticity of $NO_X$ with respect to income.

petitive methods to estimate thresholds, such as Breimans' (Breiman et al., 1984) tree regression, have no known distribution theory.

Let $\{y_i, \boldsymbol{x}_i, q_i\}$ be an observed sample, where $y_i$, $q_i \in \mathbb{R}$ and $\boldsymbol{x}_i = (1, x_{i2}, \cdots, x_{ik})^T$. The threshold variable $q_i$, which can be an element of $\boldsymbol{x}_i$, is assumed to have a continuous distribution. The threshold regression model

$$y_i = \boldsymbol{\vartheta}' \boldsymbol{x}_i + e_i, \quad q_i \leq \tau \tag{6.2}$$

$$y_i = \boldsymbol{\theta}' \boldsymbol{x}_i + e_i, \quad q_i > \tau \tag{6.3}$$

where $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2, \ldots, \vartheta_n)^T$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^T$. After defining the dummy variable,

$$d_i(\tau) = 1_{\{q_i \leq \tau\}},$$

the model (6.2)-(6.3), can be written as one equation

$$y_i = \boldsymbol{\theta}^T \boldsymbol{x}_i + \boldsymbol{\delta}^T \boldsymbol{x}_i d_i(\tau) + e_i, \tag{6.4}$$

where $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_n)^T$. Equation (6.4) allows all parameters to differ across regimes. The model can be expressed in matrix notation by defining the $n \times 1$ vectors $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ and $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)^T$, and matrices

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix}_{n \times k}, \quad \boldsymbol{X}_\tau = \begin{pmatrix} \boldsymbol{x}_1^T d_1(\tau) \\ \boldsymbol{x}_2^T d_2(\tau) \\ \vdots \\ \boldsymbol{x}_n^T d_n(\tau) \end{pmatrix}_{n \times k}.$$

Then (6.4) can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{X}_\tau \boldsymbol{\delta} + \boldsymbol{e}. \tag{6.5}$$

Let

$$S_n(\boldsymbol{\theta}, \boldsymbol{\delta}, \tau) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}_\tau \boldsymbol{\delta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}_\tau \boldsymbol{\delta}) \tag{6.6}$$

be the sum of squared errors. Keeping $\tau$ fixed, (6.5) is linear in $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$, yielding the conditional OLS estimators

$$\left(\widehat{\boldsymbol{\theta}}(\tau) \quad \widehat{\boldsymbol{\delta}}(\tau)\right)^T = (\boldsymbol{X}^{*T} \boldsymbol{X}^*)^{-1} \boldsymbol{X}^{*T} \boldsymbol{y} \tag{6.7}$$

where $\boldsymbol{X}_\tau^{*T} = \begin{pmatrix} \boldsymbol{X} & \boldsymbol{X}_\tau \end{pmatrix}$. The concentrated sum of squared error is

$$S_n(\tau) = S_n(\widehat{\boldsymbol{\theta}}(\tau), \widehat{\boldsymbol{\delta}}(\tau), \tau) = \boldsymbol{y}\boldsymbol{y}^T - \boldsymbol{y}^T \boldsymbol{X}^* (\boldsymbol{X}_\tau^{*T} \boldsymbol{X}_\tau^*)^{-1} \boldsymbol{X}_\tau^{*T} \boldsymbol{y}, \tag{6.8}$$

and $\widehat{\tau}$ can be defined as

$$\widehat{\tau} = \arg\min_{\tau \in \mathrm{T}_n} S_n(\tau).$$

where $\mathrm{T}_n$ is a suitably bounded set. Hansen (2000) showed that, under some regularity conditions, the distribution of $\widehat{\tau}$ is nonstandard but free of nuisance parameters.

To test the hypothesis

$$\mathrm{H}_0 : \tau = \tau_0,$$

a likelihood ratio approach can be employed under the maintained hypothesis that $e_i$ is independently and identically distributed $N(0, \sigma^2)$. Let the test statistic

$$LR_n = n \frac{S_n(\tau) - S_n(\widehat{\tau})}{S_n(\widehat{\tau})} \tag{6.9}$$

For large values of the statistic (6.9) the null $\mathrm{H}_0$ is rejected. Hansen (2000) shows that under certain regularity conditions

$$LR_n(\tau_0) \xrightarrow{d} \eta^2 \xi,$$

where $\xi$ is a random variable with distribution $\Pr(\xi \leq x) = (1 - e^{-x/2})^2$ and $\eta^2$ is a nuisance parameter equal to 1 in the case of homoskedasticity

$$E(e_i^2 | q_i) = \sigma^2.$$

Confidence regions based on the likelihood ratio statistic can be obtained by inverting the likelihood ratio test of $\mathrm{H}_0 : \tau = \tau_0$. Denoting with $C$ the desired asymptotic confidence level, and with $c_\xi$ the $C$-critical value for $\xi$, the confidence set is defined as

$$\widehat{\mathrm{T}} = \{\tau | LR_n \leq c\}. \tag{6.10}$$

In case of heteroskedasticity, approximate confidence regions can be constructed

based on the scaled likelihood ratio statistic,

$$LR_n^* = \frac{LR_n(\tau)}{\widehat{\eta}^2} = \frac{S_n(\tau) - S_n(\widehat{\tau})}{\widehat{\sigma}^2\widehat{\eta}^2} \tag{6.11}$$

where $\widehat{\eta}$ is a consistent estimate of the nuisance parameter (see Hansen, 2000). In case of heteroskedasticity, the modified confidence set becomes

$$\widehat{\Gamma}^* = \{\tau | LR_n^* \leq c\}. \tag{6.12}$$

Since the estimator $\widehat{\eta}^2$ is consistent for the threshold parameter $\eta^2$, $P(\tau_0 \in \widehat{\Gamma^*}) \to C$ as $n \to \infty$, so that $\widehat{\Gamma}^*$ is a heteroskedasticity-robust asymptotic C-level confidence set for $\tau$.

## 6.10 Data and Estimation Results

We illustrate the usefulness of Hansen's threshold model to the EKC debate by fitting a standard model that seeks to explain pollution emissions as a function of GDP and trade related variables. The specification used to illustrate the procedure is the log-log functional form of

$$SO2_{i,1990} = \alpha + \beta_1\, INC_{i,1990} + \beta_2\, INC^2_{i,1990} + \beta_3\, OPEN_{i,1990} + \beta_4\, FDI_{i,1990} + u_i \tag{6.13}$$

where for each country $i$:

$SO2_{i,t}$ = Sulfur Emissions measured in tons of sulfur per capita, in year $t$.

$INC_{i,t}$ = Real GDP per capita (1985 intl. prices), in year $t$.

$OPEN_{i,t} = \frac{\text{Exports+Imports}}{\text{Nominal GDP}}$, in year $t$.

$FDI_{i,t}$ = Gross Foreign Direct Investment, in % of GDP, in year $t$.

Sulfur emissions were taken from the data from the *Historical Global Sulfur Emissions* data set of A.S.L and Associates, which includes the sulfur dioxide emissions from burning hard coal, brown coal, and petroleum, and sulfur emissions from mining and related activities for most of the countries of the world during the period 1850-1990 (Lefohn et al., 1999). The real GDP per capita came from the *Penn World Tables* (PWT) Mark 5.6. The values are all measured in 1985 international

US dollars. [3] Foreign Direct Investment data were obtained from the UN *World Trade Data Base* discussed in Feenstra et al. (1997). We use the 1990 data, the last year available for $SO_2$ emissions. over a sample of 45 OECD and non-OECD countries.

We use two possible threshold variables: INC and OPEN. To select among the two variables we use heteroskedasticity-consistent Lagrange multiplier test for a threshold developed by Hansen (1996). As the threshold is not identified under the null hypothesis of no threshold. the $p$-values are obtained by means of bootstrapping. Using 1000 bootstrap replications. the $p$-value for the threshold model using INC was significant at 0.0365 for the log-linear model and marginally significant at 0.0980 for the log-quadratic model. This results suggest that there might be a sample split based on per capita income.[4] No evidence of a split based on OPEN was found. Figure 6.14 shows the graph of the heteroskedasticity-robust likelihood ratio sequence $LR_n^*(\tau)$ against the threshold in natural log of INC. Income is graphed on a natural log scale. The least square estimate of $\tau$ is the value that minimizes the curve. which occurs at $\widehat{\tau} = \$15,329$. The 95 % critical value of 7.35 is also plotted (dashed line). The asymptotic 95 % confidence set is $\widehat{\mathrm{T}}^* = [\$15,326, \$15,503]$. which in the graph is given by the levels income where the $LR_n^*(\tau)$ sequence crosses the dashed line.

Table 6.4 presents the linear and quadratic OLS for the global sample and the two samples based on the spilt on INC. The OLS for the global sample shows an ITP of about \$5.700. in agreement with previous studies for $SO_2$ emissions. The income turning point of the global sample is much lower than the threshold income that divides the two regimes. Changes that might benefit the environment occur at much higher levels of income than those implied by standard EKC models. The ITP of the global sample is much lower than the threshold income that

[3]The PWT are described in Alan Heston and Robert Summers The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950-1988, Quarterly Journal of Economics, May 1991, pp. 327–368.

[4]To compute the estimates and confidence intervals for threshold model, we used Hansen's program written in GAUSS available from `http://www.ssc.wisc.edu/~bhansen/progs/progs_threshold.html`. The version of GAUSS used to run the program was the GAUSS for Windows version 6.0. The hardware used was a Dual Intel Pentium IV (Prestonia) Xeon Processors 3.06 GHz with HT Technology with 4 GB of RAM running on Microsoft Windows XP/2002 Professional (Win32 x86) 5.01.2600 (Service Pack 2).

| Variable | **Global OLS** | | **Regime 1** | | **Regime 2** | |
|---|---|---|---|---|---|---|
| Constant | −16.991 | −86.387 | −28.137 | −104.337 | −19.001 | −6774.782 |
| | (3.0469) | (16.3956) | (2.9606) | (14.8965) | (5.5303) | (2407.6255) |
| LGDP | −0.1765 | 15.8955 | 1.4806 | 20.3905 | −0.2504 | 1373.0482 |
| | (0.3906) | (3.8462) | (0.3276) | (3.5594) | (0.5413) | (491.1188) |
| LGDP2 | | −0.9192 | | −1.1702 | | −69.7848 |
| | | (0.2216) | | (0.2117) | | (25.0332) |
| LTRADE | 1.3455 | 1.2542 | 0.5694 | 0.6290 | 1.9999 | 2.1913 |
| | (0.3331) | (0.2628) | (0.3696) | (0.3013) | (0.2709) | (0.2830) |
| LFDI | −0.4550 | −0.3779 | −0.2503 | −0.1905 | −0.6350 | −1.1058 |
| | (0.1339) | (0.1413) | (0.1403) | (0.1209) | (0.2069) | (0.1742) |
| TP | | 8.647 | | 8.712 | | 9.838 |
| | | (5692) | | (6077) | | (1.873e+004) |
| $\widehat{\sigma}^2$ | | 0.891 | | 0.762 | | 0.366 |

Table 6.4: Regression coefficients

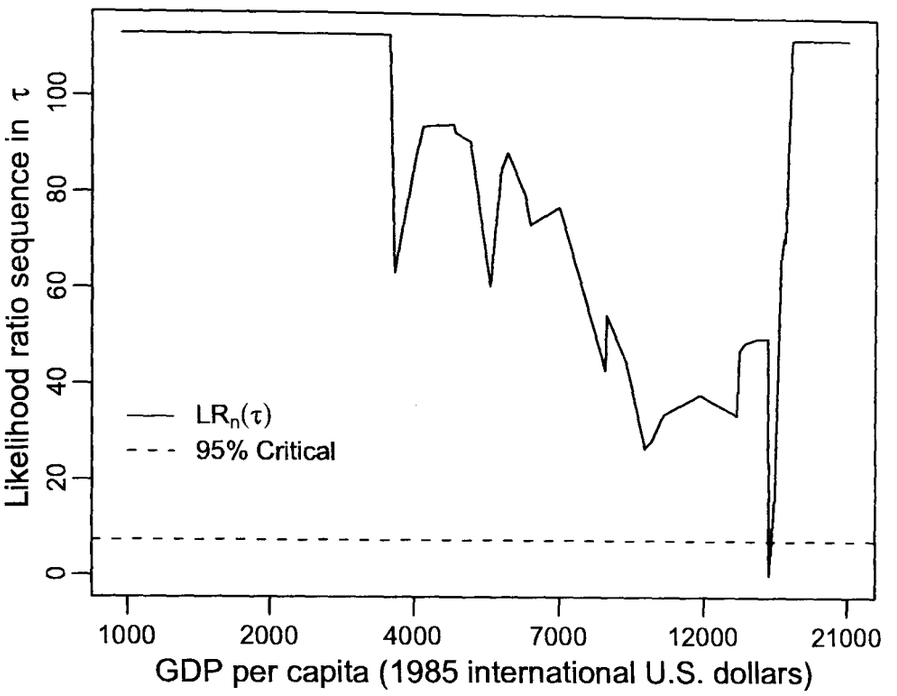



Figure 6.14: Confidence interval construction for threshold

divides the two regimes. On the other hand the impact of income on pollution is greater in regime 2 than in regime 1. This is illustrated in Figure 6.15 where the estimated quadratic relationships between income and emission for the two regimes are illustrated. Regime differences are also apparent from the estimated error variance. The estimated error variance of regime 1, the poorer countries, is more than twice that of regime 2, the richer countries. Panayotou (2000) after examining the evidence form Vincent (1997) and Carson et al. (1997a) concerning the existence of a Kuznets curve within individual countries concludes that: "whereby rising incomes result in a more effective regulatory structure by changing public preferences and making resources available to regulatory agencies. States with low-income levels have a far greater variability in emissions per capita than high-income states suggesting more divergent development paths. This has the implication that it may be more difficult to predict emission levels for low-income countries approaching the turning point." A formal test supports this hypothesis. The statistics for the Goldfeld-Quandt test for the null that the two variances are equal versus the alternative that the variance for regime 1 is higher than that of

regime 2 (Goldfeld & Quandt. 1997) is 2.325 with a $p$-value of 0.059.

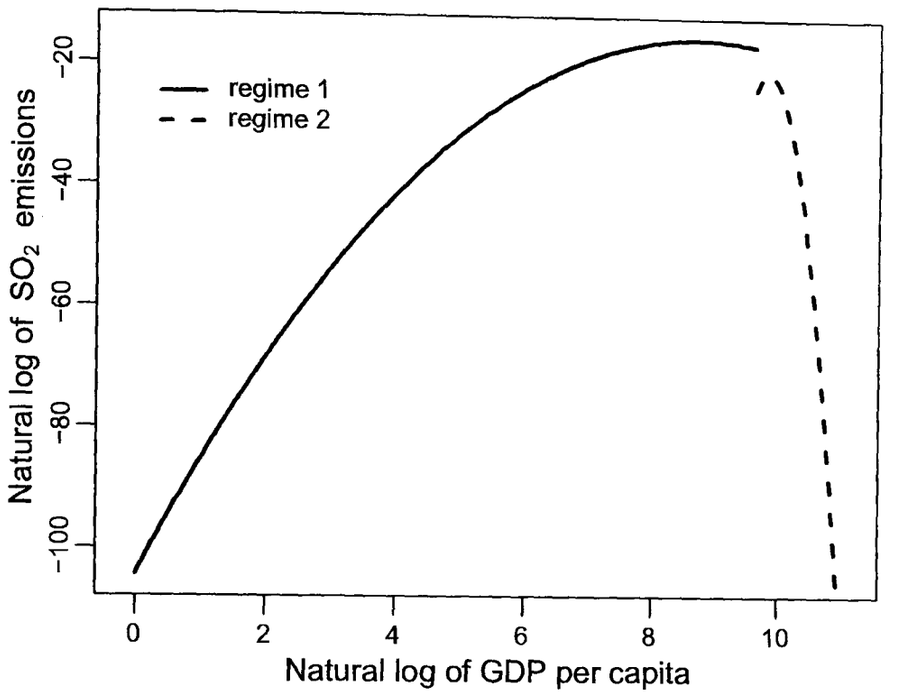



Figure 6.15: EKC estimated curves for different regimes

## 6.11 Conclusion and Further Studies

The environmental Kuznets curve, the inverted U-shaped relationship between economic development and environmental quality, is an empirical "law" of environmental economics that has been documented in many cross-country studies.

In this chapter we found that threshold estimation is a promising technique that can be used to test a different class of models of the environment-economic system and support a conscious policy intervention. Applying this methodology to the environmental Kuznets curve debate, we find support for threshold models that lead to different reduced-form relationships between environmental quality and economic activity when early stages of economic growth are contrasted with later stages. In agreement with the findings from Chapter 5, we find no evidence of

a common inverted U-shaped environment/economy relationship that all country follow as they grow economically. We also find that changes that might benefit the environment occur at much higher levels of income than those implied by standard models. These findings suggest that there is nothing automatic about these changes, improvements are a consequence of the deliberate introduction of policies addressing environmental problems.

Moreover, we find evidence that countries with low-income levels have a far greater variability in emissions per capita than high-income countries. This implies that it may be more difficult to predict emission levels for low-income countries that may be approaching a turning point.

These findings suggest that policy maker should exercise extreme caution, particularly in developing countries as, as Arrow et al. (1995) pointed out, "policies that promote gross national product are not substitutes for environmental policy." Moreover, as there is evidence of more uncertainty about possible future development paths and the location of possible turning points, there are reasonable grounds for concern that potentially dangerous and irreversible effects on the environment may occur if appropriate precautionary action is not taken.

There are some important caveats to bear in mind. Estimated error variance might reflect the impact poorer quality data, omitted pollution determinants, and so on.

Possible directions for further research include the following aspects.

- Our threshold estimation has focussed only on one pollutant, sulfur dioxide emissions. As mention in previous sections, since it is one of the main pollutants, these result should be of interest to policy makers. Also, previous studies clearly show that sulfur dioxide emissions behave similarly to other local impact pollutant with serious health consequences such as nitrogen oxides, and particulates. It is likely that some of this results may be applicable in other cases, but it would be of interest, in a further study, to apply this methodology to a global pollutant affecting climate change such as carbon dioxide.

- Another area requiring further investigation is the choice of control variants. Based on the results in the previous chapter, it would be beneficial to re-estimate the threshold models using capital abundance, as defined in Antweiler et al. (2001). Other variables that have been used in the lit-

erature on the EKC that could be further investigated include, industrial composition of output (see, e.g., Grossman & Krueger, 1995), population density (see,e.g., Cropper & Griffiths, 1994; Selden & Song, 1994), openness to trade (see, e.g., Antweiler et al., 2001; Hettige et al., 1992; Grossman & Krueger, 1993b; Suri & Chapman, 1998), environmental regulation and control (see, e.g., Shafik, 1994a; Baldwin, 1995), democracy (see, e.g., Torras & J.K., 1998; Harbaugh et al., 2002), corruption (see, e.g., Lopez & Mitra, 2000), civil and political liberties (see, e.g., Barrett & Graddy, 2000; Torras & J.K., 1998), power inequality (see, e.g., Boyce, 1994), literacy (see, e.g., Torras & J.K., 1998), geographical factors (see, e.g., Neumayer, 2002), income inequality (see, e.g., Torras & J.K., 1998; Magnani, 2000; Ravallion et al., 2000), and so on.

- Extending this methodology to panel data observations would improve the reliability of the results by allowing to control for various unobserved effects so it would be interesting and important subject of future research.

# Part III

# Conclusion

Chapter 7

# Summary

The role of nonparametric methods in econometrics has increased in importance during the past several yeas. A quick search through through economic journal databases reveals that, though most economic application of nonparametric methods are recent, they are steadily growing in number. The choice between between traditional parametric and semiparametric/nonparametric method is rapidly tilting towards the latter, as computations become ever cheaper.

Their increase in popularity can be attributed in part to their flexible nature but also to the ever growing computational power, the availability of more powerful graphic devices, and their implementation many in off-the-shelf software. Many statistical and econometrics software application offer nonparametric density and regression estimators that can be accessed with few clicks of a mouse or with a simple function call at a prompt.

In this thesis emphasis was given to methods that enable the inclusion of multiple explanatory variables without suffering of the so called "curse of dimensionality" that severely limits the applicability of standard neonparametric methods. In this thesis we have seen, through relevant economic applications, how these methods can be used in conjunction with parametric methods to mutually support each others findings. Once a probabilistic structure has been identified by nonparametric means, we can adopt, whenever appropriate and on an independent sample, a fully parametric approach, to reinforce the nonparametric results and to test relevant economic hypothesis.

In Chapter 3 we have proposed some basic standards to improve the use and reporting of nonparametric methods in the statistics and economics literature for the purpose of accuracy and reproducibility. In particular, we made recommendations in five aspects of the process: computational practice, published reporting, numerical accuracy, reproducibility, and visualization. We have highlighted the fact that nonparametric methods are inherently computationally intensive and rely on a plethora of implementation details that can be built-in the software application, fixed as default settings, or determined by the researcher. The control available over these implementation details is a function of both the sophistication of the software and the user. More knowledgeable users and better designed software can give greater control over the nonparametric estimation procedure. Detailed control over the estimation procedure is often required to achieve more accurate results, for correct model selection strategy, for efficiency in computation, and to facilitate reproducibility and further research. We have also reflected on current

developments in the practice of computing, visualization, and open source software, and their potential usefulness in making empirical research in economics using nonparametric methods more easily reproducible.

In Chapter 4 we investigated the effect of demographic and socio-economic characteristics of households on income inequality in the UK. We started by estimating the conditional distribution of income over a broad set of determinants. We then devised a method for obtaining conditional inequality measures by inverting the estimated conditional distribution. Our results provided a visually clear representation of both the substantive and statistical impact of each factor on income inequality, keeping all others constant.

Our approach is novel in at least four respects. First, by estimating the entire conditional distribution of income over a broad set of determinants, our estimation procedure uncovers higher-order properties of the income distribution and non-linearities of its moments that cannot be captured by means of a "standard" parametric approach. For example, similar to the results obtained in the previous literature, we found that the shape of the age-income profiles agrees with the observable prediction of the life-cycle model, which assumes that resources are accumulated at a faster rate at a young age. Also, we found that income of families during the period of child rearing was higher than income in the retirement stage of the life-cycle, when economic responsibility is greatly reduced. In addition, we found that the age-income profiles peaked later for the wealthier households and appeared considerably non-linear, declining rapidly after the age of 60. Besides having important consequences for the policy maker as such, the asymmetry might also indicate the presence of different factors affecting the upward and downward branches of the age-income profile that have not been included in our and previous analysis. For instance, factors that determine a loss in earning capacity at retirement age of individuals, like deterioration of health and increasing aversion towards risk, could help in explaining the observed asymmetry.

Second, by estimating the whole distribution we were able to identify where in the distribution of income the various determinants exerted their greatest impact. This detailed analysis provided further insight into the determinants of inequality, of great importance to researchers as well as policy makers. For example, we found that, in agreement with previous published results, the impact of employment status was spread over the entire income distribution. However, in addition, we found that the impact on income was substantially greater for lower income families.

Third, we devised a method for obtaining nonparametric conditional inequality measures by inverting the estimated conditional distribution. Our estimates indicated, for example, that if average household size increased from 2 to 4, households in the top 90th percentile of the income distribution would move from earning 3.2 times more then households in the 10th percentile to earning about 2.5 times more. This amounted to a 20 per cent fall in inequality. This increase in inequality was obtained after controlling for other important factors, such as the age structure, the presence of a retired head and young children. Previous approaches, based on the "standardization" of inequality series, inequality decomposition by population sub-groups, or nonparametric methods, have not been to identify the contribution of individual factors on inequality, except for very simple cases.

Finally, our approach allowed us to establish consistency and to estimate asymptotic variances of the proposed inequality estimators, which was useful for inference purposes. It provided a visually clear representation of both the substantive and statistical impact of each individual factor on income inequality, keeping all others constant. For instance, we found that for the UK sample, household size, number of young children, age of head, and employment status, have a large substantive and statistical impact on inequality. Factors such as years of education, marital status, and urban versus rural households, on the other hand, did not significantly impact inequality. Combined with the recent trend of declining household size in the UK, this results could help explain the trend of increasing income inequality observed in the past decades in the UK.

Chapter 5 we re-examined the relationship between openness to trade and the environment, controlling for economic development, in order to identify the presence of multiple regimes in the cross-country pollution-economic relationship. We first identified the presence of multiple regimes, then we developed an easily interpretable measure, based on an original application of the Blinder-Oaxaca decomposition, of the impact on the environment due to differences in regimes, and finally we applied a nonparametric recursive partitioning algorithm to endogenously identify various regimes. Our conclusions were threefold. First, we rejected the null hypothesis that all countries obey a common linear model. Second, we found that quantitatively regime differences can have a significant impact. Thirdly, by using regression tree analysis we found subsets of countries which appear to possess very different environmental/economic relationships. In particular, we found that the impact of openness to foreign markets on sulfur and carbon dioxide emissions varies

according to the level of development, trade policies, and the productive structure of the economy. Our result also showed that there is substantial geographic homogeneity within each regime, giving some support to findings by geographical factors. Our finding also highlighted the importance of democracy, corruption, and civil and political liberties. We found support for studies that based on the poor environmental performance of Soviet economies and dictatorships established in Latin America, Asia and Africa, have been advocating democratic reforms as a way to promote both economic and environmental welfare.

In Chapter 6 investigate the existence of the so called *environmental kuznets curve* (EKC), the inverted-U shaped relationship between income and pollution, using nonparametric regression methods.

The flexible nature of nonparametric estimation allowed as to find evidence of an asymmetric behaviour of the curve before and after the turning point, consistently with threshold-effect models. This finding are also consistent with a strand of previous empirical evidence concerning the existence of a Kuznets curve within individual countries. We investigated these nonparametric findings further using a threshold estimation method. Our findings have considerable implications for the policy maker. Applying this methodology to the environmental Kuznets curve debate, we found support for threshold models that lead to different reduced-form relationships between environmental quality and economic activity when early stages of economic growth are contrasted with later stages. We found little evidence of a common inverted U-shaped environment/economy relationship that all country follow as they grow economically. We also found evidence that changes that might benefit the environment occur at much higher levels of income than those implied by standard models. These findings suggest that there is nothing automatic about these changes, improvements are a consequence of the deliberate introduction of policies addressing environmental problems.

We also found that regime differences are apparent from differences in the estimated error variance. The estimated error variance of the poorer countries regime was more than twice that of the richer countries regime. This implies that it may be more difficult to predict emission levels for low-income countries that may be approaching a turning point. This result is consistent with recent models of the EKC that assume that before crossing the turning point pollution in poorer countries may be completely unregulated.

We found that threshold estimation is a promising technique that can be used

to test a different class of models of the environment-economic system and support a conscious policy intervention. These findings suggest that policy maker should exercise extreme caution, particularly in developing countries, when promoting growth as a solution to environmental problems. As Arrow et al. (1995) pointed out, "policies that promote gross national product are not substitutes for environmental policy." Moreover, as there is evidence of more uncertainty about possible future development paths and the location of possible turning points, there are reasonable grounds for concern that potentially dangerous and irreversible effects on the environment may occur if appropriate precautionary action is not taken. With fast-growing developing countries experiencing increasing environmental problems like China and India this uncertainty makes inaction a very risky strategy for the future of our planet.

Based on this, necessarily short, application of nonparametric methods in economics, we conclude that these methods are having, and will continue to have a considerable impact on the discipline. In particular, we have seen that the development of semiparametric methods that overcome the "curse of dimensionality" problem that afflicted earlier nonparametric approaches, has ensured that richer and more interesting economic problems can be usefully investigated through their application.

Chapter 8

# Future Research Directions

There are several areas in the application of nonparametric methods to economics where additional research would be valuable. For results to become more detailed and useful to the policy makers the choice of models and variables can be extended. For policies to be reliable, improvement in the methodology, and further testing of the models and hypothesis using independent data is needed.

Income inequality is an important field where nonparametric methods have emerged and established themselves as a tool to advance the discipline. We have seen how applying nonparametric techniques can provide further insight into the determinants of inequality, that of great importance to researchers and policy makers alike.

To strengthen our results, an important aspect to address are potentially endogenous regressors. The estimation method assumes that regressors are exogenous. This can be certainly argued for age and possibly education. However, household income is an important determinant of the decision to have children, household formation, marriage, household dissolution, retirement to some extent, and so on. Some other econometric approach, such as instrumental variables, could be explored to obtain improved estimates.

There are several interesting hypothesis that have emerged from this study such as the possible effect of liquidity constraints on education and the possibility that the impact of worsening health condition or and changing attitudes toward risk. It would be interesting to formally test these hypotheses on an independent sample.

Another aspect for further research concerns the methodology. It would be useful to compare our method with other alternative approaches, such as the nonlinear quantile regression. Also, it would be useful to extend the approach we used to explore inequality to make use of the panel nature of the data. Though preliminary analysis did not show any significant change in the results, a panel approach would allow to track households over time and to model age and cohort effects. Ignoring cohort effects produces age-income profiles that could be biased. age-income profiles can vary across cohorts, particularly for cohorts that are distant in time.

More research directions on the relationship between the environment and economic growth based on our study have also emerged. We investigates the existence of the so called environmental kuznets (EKC) curve using nonparametric and semiparametric regression methods. The EKC features two variables of considerable

interests to economists and policy makers, namely an indicator of environmental quality and the level of per capita income.

One area of improvement concerns the dependent variables used. The Blinder-Oaxaca decomposition and the threshold estimation method both focused only on one pollutant, namely sulfur dioxide emissions. As it is one of the main pollutants, these result should be of interest to policy makers. Also, previous studies clearly show that sulfur dioxide emissions behave similarly to other local impact pollutant with serious health consequences such as nitrogen oxides, and particulates. It is likely that some of this results may be applicable in other cases. It would be interesting, in a further study, to apply this methodology to a global pollutant affecting climate change such as carbon dioxide and to other local pollutants to assess the robustness of our results.

The choice of regressors is also an area warranting further investigation. It became apparent during our analysis that the choice of variables can seriously affect results. This has also been established in recent published work. For instance, the variable used in this study for capital intensity behaves anomalously, as for richer countries we found that more capital intensity significantly reduces emissions, which contrasts with previously published findings. In fact, it is often assumed that capital intensity directly translates into pollution-intensity. This seems to be too simplistic. There appears to be the need to control for the level of "dirtiness" of an industry to improve our analysis. This variable we used was found to be highly correlated with income. Also, in order to improve the comparability of this study with others, it would be beneficial to re-estimate the models using capital abundance adjusted for differences in worker's productivity. Other variables that have been used in the literature on the EKC that could be further investigated include, industrial composition of output, population density, environmental regulation and control, democracy, corruption, civil and political liberties, power inequality, literacy, geographical factors, and income inequality.

For robust policy recommendations, the methodology used can also be refined further. A more thorough investigation on the sensitivity to our results to implementation details would greatly increase the value of our findings. Alternative threshold methods that have recently appeared in the literature can also be considered.

In the reporting of nonparametric results, possible directions for further research include extending the benchmark from the univariate density estimator

to other multivariate and regression settings. To benefit other researchers, the best way to report the benchmarks is to have them available via the web. An obvious choice seem to make them available through the Stanford site, "Econometric Benchmarks,"[1] maintained by Clint Cummins. "Econometric Benchmarks" makes some standard benchmark datasets and models for testing the accuracy of econometrics application software available for download. So far benchmarks are available for basic statistics, linear and nonlinear regression, simultaneous equations, time series, qualitative dependent variables, panel data models, and random number generation. After having constructed the benchmarks, the next step is to test popular statistics and econometric packages that support some of these methods and to disseminate reports on how close they come to the benchmarks.

[1] Accessible at `http://www.stanford.edu/~clint/bench/`.

# Bibliography


Amman, H., Kendrick, D., & Rust, J. (Eds.). (1996). *Handbook of Computational Economics*, volume 1. Amsterdam, The Netherlands: Elsevier North-Holland, Inc.

Andreoni, J. & Levinson, A. (2001). The Simple Analytics of the Environmental Kuznets Curve. *Journal of Public Economics*, *80*(2), 269–286.

Antweiler, W., Copeland, B., & Taylor, S. M. (2001). Is Free Trade Good for the Environment? *American Economic Review*, *91*(4), 877–908.

Arrow et al. (1995). Economic Growth, Carrying Capacity, and the Environment. *Science*, *268*, 520–521.

Arrow, K., Bolin, B., Costanza, R., Dasgupta, P., Folke, C., & Holling, C. S. a. (1995). Economic growth, carrying capacity, and the environment. *Ecological Economics*, *15*(2), 91–95. available at http://ideas.repec.org/a/eee/ecolec/v15y1995i2p91-95.html.

Atkinson, A. B. (1971). The Distribution of Wealth and the Individual Life-Cycle. *Oxford Economic Papers*, *23*(2), 1–37.

Atkinson, A. B. (1997). Bringing Income Distribution in from the Cold. *The Economic Journal*, *107*, 297–321.

Azomahou, T. & Phu, N. V. (2006). Heterogeneity in Environmental Quality: An Empirical Analysis of Deforestation. *Journal of Development Economics*. Forthcoming.

Azzalini, A. & Bowman, A. W. (1990). A look at some data on the Old Faithful Geyser. *Applied Statistics*, *39*(3), 357–365.

Baiocchi, G. (2005). Monte carlo methods in environmental economics. In R. Scarpa & A. Alberini (Eds.), *Applications of simulation methods in environmental and resource economics* chapter 16, (pp. 317–340). Dordrecht, The Netherlands: Springer Publisher.

Baiocchi, G. (2007). Reproducible Research in Computational Economics: Guidelines, Integrated Approaches, and Open Source Software. *Computational Economics*, *30*(1), 19–40.

Baiocchi, G. & Distaso, W. (2003). GRETL: Econometric Software for the GNU Generation. *Journal of Applied Econometrics*, *18*(1), 105–110.

Baldwin, R. (1995). Does sustainability require growth? In I. Goldin & L. Winters (Eds.), *The U.S.- Mexico Free Trade Agreement* (pp. 19–46). Cambridge: Cambridge University Press.

Banks, J. & Johnson, P. (1994). Equivalence Scale Relativities Revisited. *Economic Journal*, *104*(425), 883–90. available at http://ideas.repec.org/a/ecj/econjl/v104y1994i425p883-90.html.

Barbier, E. (1997). Introduction to the Environmental Kuznetz Curve. *Environment and Development Economics*, *2*(4).

Barrett, S. & Graddy, K. (2000). Freedom, Growth, and the Environment. *Environment and Development Economics*, *5*(4), 433–56.

Barrodale, I. & Roberts, F. D. K. (1973). An Improved Algorithm for Discrete l1 Linear Approximation. *SIAM Journal of Numerical Analysis*, *10*(5), 839–848.

Beckerman, W. (1992). Economic growth and the environment: Whose growth? whose environment? *World Development*, *20*.

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Blinder, A. S. (1973). Wage Discrimination: Reduced form and Structural Sstimates. *Journal of Human Resources*, *8*(4), 436–455.

Blinder, A. S. & Esaki, H. Y. (1978). Macroeconomic Activity and Income Distribution in the Postwar United States. *Review of Economics and Statistics, 60*, 604–609.

Bourguignon, F. (1979). Decomposable Income Inequality Measures. *Econometrica, 47*(4), 901–920.

Bowman, A. & Azzalini, A. (1997). *Density Estimation for Inference.* New York: Oxford University Press.

Bowman, A. W., Jones, M. C., & Gijbels, I. (1998). Testing monotonicity of regression. *Journal of Computational and Graphical Statistics, 7*(4), 489–500.

Boyce, J. K. (1994). Inequality as a Cause of Environmental Degradation. Technical report.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees.* Belmont: Wadsworth.

Brewer, M., Goodman, A., Myck, M., Shaw, J., & Shephard, A. (2005). Inequality in Britain: 2005. Commentary no. 99. London: Institute for Fiscal Studies.

Brock, W. A. & Taylor, S. M. (2004). The Green Solow Model. Technical report 988, National Bureau of Economic Research, Madison, WI.

Buckheit, J. B. & Donoho, D. L. (1995). *Wavelets and Statistics*, chapter Wavelab and Reproducible Research, (pp. 55–81). Berlin, New York: Springer-Verlag.

Burtless, G. (1995). International Trade and the Rise in Earnings Inequality. *Journal of Economic Literature, 33*, 800–816.

Buse, A. (1982). The Cyclical Behavior of the Size Distribution of Income in Canada: 1947–78. *Canadian Journal of Economics, 40*, 189–204.

Busovaca, S. (1985). *Handling degeneracy in a nonlinear l(,1) algorithm.* PhD thesis.

Carson, R., Jeon, Y., & McCubbin, D. (1997a). The Relationship Between Air Pollution Emissions and Income: US Data. *Environment and Development Economics, 2*(4), 433–450.

Carson, R. T., Jeon, Y., & McCubbin, D. (1997b). The Relationship Between Air Pollution and Income: US Data. *Environment and Development Economics, 2*, 433–450.

Cavlovic, T. A., Baker, K. H., Berrens, R. P., & Gawande, K. (2001). A Meta Analysis of Environmental Kuznets Curve Studies. *Agricultural and Resource Economic Review, 29*, 32–42.

Chambers, J. M. & Hastie, T. (Eds.). (1991). *Statistical models in S.* London: Chapman & Hall.

Checchi, D. (2001). Does Educational Achievement Help to explain Income Inequality? In G. A. Cornia (Ed.), *Inequality, Growth and Poverty in an Era of Liberalization and Globalization,* chapter 4. Oxford: Oxford University Press.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*, 829–836.

Cleveland, W. S. (1980). *The Elements of Graphing Data.* Wadsworth Advanced Books and Software.

Cleveland, W. S. (1985, 1994). *The Elements of Graphing Data.* New York: Chapman and Hall.

Cleveland, W. S. (1993). *Visualizing Data.* Wadsworth Advanced Books and Software.

Cleveland, W. S., Grosse, E., & Shyu, W. M. (1992). Local regression models. In J. M. Chambers & T. J. Hastie (Eds.). *Statistical Models in S* chapter 8. Pacific Grove, CA: Wadsworth and Brooks/Cole.

Cleveland, W. S. & McGill, R. (1987). Graphical Perception: The Visual Decoding of Qquantitative Information on Statistical Graphs (with Discussion). *Journal of the Royal Statistical Society Series A, 150*, 192–229.

Cole, M., Rayner, A., & Bates, J. (1997). The environmental kuznets crve: An empirical analysis. *Environment and Development Economics, 2*(4), 401–416.

Copeland, B. & Taylor, S. M. (2003). *Trade and the Environment: Theory and Evidence.* Princeton: Princeton University Press.

Cotton, J. (1988). On the Decomposition of Wage Differentials. *Review of Economics and Statistics*, *70*, 236–243.

Coulter, F. A. E., Cowell, F. A., & Jenkins, S. P. (1992). Equivalence Scale Relativities and the Extent of Inequality and Poverty. *Economic Journal*, *102*(414), 1067–82.

Cowell, F. A. (1980). On the Structure of Additive Inequality Measures. *The Review of Economic Studies*, *47*(3), 521–531.

Cropper, M. & Griffiths, C. (1994). The Interaction of Population Growth and Environmental Quality. *American Economic Review*, *84*(2), 250–54. available at http://ideas.repec.org/a/aea/aecrev/v84y1994i2p250-54.html.

David, H. (1995). First (?) Occurrence of Common Terms in Mathematical Statistics. *The American Statistician*, *49*(2), 121–133.

Davidson, R. & MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford: Oxford university Press.

de Bruyn, S. M., van den Bergh, J. C. J. M., & Opschoor, J. B. (1998). Economic growth and emissions: reconsidering the empirical basis of environmental kuznets curves. *Ecological Economics*, *25*(2), 161–175.

de Leeuw, J. (2005). On Abandoning XLISP-STAT. *Journal of Statistical Software*, *13*(7), 1–81.

Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. *American Economic Review*, *76*(4), 587–603.

Dijkgraaf, E. & Melenberg, B. (2005). Environmental Kuznets Curves for CO2: Homogeneity Versus Heterogeneity. Technical Report 2005.

Dijkgraaf, E. & Vollebergh, H. R. (2005). A Test for Parameter Heterogeneity in CO2 Panel EKC Estimations. *Environmental and Resource Economics*. Forthcoming.

DiNardo, J., Fortin, N., & Lemieux, T. (1989). Labour Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. *Econometrica*, *64*(5), 1001–1044.

DiNardo, J., Fortin, N., & Lemieux, T. (1996). Labour Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. *Econometrica*, *64*(5), 1001–1044.

Dinwiddy, R. & Reed, D. (1977). The Effects of Certain Social and Demographic Changes on Income Distribution. H.M. Stationery Off.: London.

Durlauf, S. N. & Johnson, P. A. (1995). Multiple Regimes and Cross-Country Growth Behavior. *Journal of Applied Econometrics*, *10*(4), 365–384.

Eddelbuettel, D. (2000). Econometrics with Octave. *Journal of Applied Econometrics*, *15*(5), 531–542.

Eddelbuettel, D. (2003). Quantian: A Scientific Computing Environment. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003) March 20–22, Vienna, Austria*, Vienna, Austria. Technische Universitt Wien.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, *87*, 998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, *21*, 196–216.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman and Hall.

Feenstra, R., Lipsey, R., & Bowen, H. (1997). World Trade Flows, 1970-1992, with Production and Tariff Data. Technical report, NBER Working Paper No. 5910.

Fisher, R. A. (1936). Has Mendel's work been rediscovered. *Annals of Science*, *1*, 115–137.

Fitzenberger, B., Koenker, R., & Machado, J. A. (2002). Economic Applications of Quantile Regression. *Series: Studies in Empirical Economics (guest editorial)*.

Ford, M. P. & Ford, D. J. (2001). Investigation of GAUSS' Random Number Generators. Technical report for aptech systems, inc., FORWARD Computing and Control Pty. Ltd., NSW Australia.

Foresi, S. & Peracchi, F. (1995). The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association*, *90*, 451–466.

Fox, J. (2002). *An R and S-Plus Companion to Applied Regression.* Thousand Oaks, CA, USA: Sage Publications. ISBN 0761922792.

Frankel, J. A. & Rose, A. K. (2005). Is Trade Good or Bad for the Environment? Sorting out the Causality. *The Review of Economics and Statistics*, *87*(1), 85–91.

Freedman, D. & Diaconis, P. (1981). On the Histogram as a Density Estimator: $L_2$ Theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *57*(4), 453–476.

Gallagher, K. S. (2003). Development of Cleaner Vehicle Technology? Foreign Direct Investment and Technology Transfer from the United States to China. Paper presented at United States Society for Ecological Economics 2nd Biennial Meeting 2.99, Saratoga Springs NY.

Gentle, J. E. (2003). *Random Number Generation and Monte Carlo Methods* (second ed.). New York: Springer-Verlag.

Gentleman, R. (2004). Some Perspectives on Statistical Computing. Technical report, Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA.

Geweke, J. (1996). Monte Carlo Simulation and Numerical Integration. In H. M. Amman, D. A. Kendrick, & J. Rust (Eds.), *Handbook of Computational Economics, Volume 1* (pp. 731–800). Amsterdam: North-Holland.

Glewwe, P. (1991). Household Equivalence Scales and the Measurement of Inequality : Transfers from the Poor to the Rich could Decrease Inequality. *Journal of Public Economics*, *8*(2), 211–216.

Goldfeld, S. M. & Quandt, R. E. (1997). Some Tests for Homoscedasticity. *Journal of the American Statistical Association*, *60*, 539–547.

Greene, W. (2000). *Econometric Analysis* (Fourth ed.). New York: Prentice Hall.

Grossman, G. & Krueger, A. (1993a). Environmental Impacts of a North American Free Trade Agreement. In P. Gaber (Ed.), *US-Mexico Free Trade Agreement.* Cambridge, MA: MIT Press.

Grossman, G. M. & Krueger, A. B. (1991). Environmental Impact of a North American Free Trade Agreement. Technical Report 3914, NBER Working Paper.

Grossman, G. M. & Krueger, A. B. (1993b). Environmental Impacts of a North American Free Trade Agreement. In P. Garber (Ed.), *The U.S.- Mexico Free Trade Agreement* (pp. 13–56). Cambridge, MA: MIT Press.

Grossman, G. M. & Krueger, A. B. (1995). Economic growth and the environment. *Quaterly Journal of Economics, 110*(2), 353–377.

Hall, R., Wolff, R. C. L., & Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association, 94*(445), 154–163.

Hansen, B. E. (1996). Inference When a Nuisance Parameter Is Not Identified under the Null Hypothesis. *Econometrica, 64*(2), 413–30.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica, 68*(3), 575–604.

Harbaugh, W. T., Levinson, A., & Wilson, D. M. (2002). Reexamining the Empirical Evidence for an Environmental Kuznets Curve. *Review of Economics and Statistics, 84*(3), 541–551.

Härdle, W. (1991). *Smoothing Techniques with Implementation in S.* New York: Springer-Verlag.

Hastie & Tibshirani (1990). *Generalized Additive Models.* London: Chapman and Hall.

Hastie, T. & Loader, C. (1993). Local regression: Automatic kernel carpentry. *Statistical Science, 8*, 120–129.

Hauck, W. W. & Anderson, S. (1984). A Survey Regarding the Reporting of Simulation Studies. *The American Statistician, 38*(3), 214–216.

Hettige, H., Lucas, R. E. B., & Wheeler, D. (1992). The Toxic Intensity of Industrial Production: Global Patterns, Trends, and Trade Policy. *American Economic Review*, *82*(2), 478–81.

Hoaglin, D. C. & Andrews, D. F. (1975). The Reporting of Computation-Based Results in Statistics. *The American Statistician*, *29*(3), 122–126.

Hollander, M. & Wolfe, D. (1973). *Nonparametric Statistical Methods*. New York, NY: John Wiley & Sons.

Hyndeman, R. J., Bashtannyk, D. M., & Grunwald, G. K. (1996). Estimating and Visualizing Conditional Densities. *Journal of Computational and Graphical Statistics*, *5*(4), 315–336.

Ihaka, R. & Gentleman, R. (1996). R: a Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, *5*, 299–314.

Jappelli, T. & Modigliani, F. (2005). The Age-Saving Profile and the Life-Cycle Hypothesis. In *The Collected Papers of Franco Modigliani - Vol. 6* (pp. 141–172). The MIT Press.

Jenkins, S. P. (1995a). Accounting for Inequality Trends: Decomposition Analyses for the UK, 1971–86. *Economica*, *62*, 29–63.

Jenkins, S. P. (1995b). Did the middle class shrink during the 1980s? UK evidence from kernel density estimetes. *Economic Letters*, *49*, 407–413.

Johnston, J. & DiNardo, J. (1997). *Econometric Methods* (fourth ed.). Singapore: McGraw-Hill.

Jones, L. & Manuelli, R. (1995). A positive model of growth and pollution controls. Technical Report 5205, Working Paper No.5205, National Bureau of Economic Research, Cambridge, MA.

Judd, K. & Tesfatsion, L. (Eds.). (2006). *Handbook of Computational Economics: Agent-Based Computational Economics*, volume 2. Amsterdam, The Netherlands: Elsevier North-Holland, Inc.

Knopper, K. (2003). Knoppix. Available at *http://www.knopper.net/knoppix/index-en.html,*.

Knüsel, L. (1989). Computergestützte Berechnung Statistischer Verteilungen. Technical report, Oldenburg, München-Wien (an English version of the program is available from *http://www.stat.uni-muenchen.de/~knuesel/elv*).

Knüsel, L. (1995). On the Accuracy of Statistical Distributions in GAUSS. *Computational Statistics and Data Analysis*, *20*, 699–702.

Koenker, R. (1994). Confidence Intervals for Regression Quantiles. In Mandl. P. & Huskova, M. (Eds.), *Proceedings of the 5th Prague Symposium on Asymptotic Statistics*, (pp. 349–359). Physika Verlag, Heidelberg, Germany.

Koenker, R. (1996). Reproducible Econometric Research. Technical report, Department of Econometrics, University of Illinois, Urbana-Champaign, IL.

Koenker, R. (2006). Reproducibility in Econometrics Research. Technical report, Department of Econometrics, University of Illinois, Urbana-Champaign, IL. `http://www.econ.uiuc.edu/~roger/repro.html`.

Koenker, R. & Bassett (1978). Regression quantiles. *Econometrica*, *46*, 33–49.

Koenker, R. & Hallock, K. (2001). Quantile Regression: An Introduction. *Journal of Economic Perspectives*, *15*(4), 43–56.

Koenker, R. W. & d'Orey, V. (1987). Computing Regression Quantiles. *Applied Statistics*, *36*(3), 383–393.

Koenker, R. W. & d'Orey, V. (1994). Remark on Alg. AS 229: Computing Dual Regression Quantiles and Regression Rank Scores. *Applied Statistics*, *43*(2), 410–414.

Kuznets, S. (1955). Economic Growth and Income Inequality. *American Economic Review*, *45*, 1–28.

L'Ecuyer, P. (1999). Good Parameters and Implementations for Combined Multiple Recursive Random Number Generators. *Operations Research*, *7*(1), 159–164.

Lefohn, A. S., Husar, J. D., & Husar, R. B. (1999). Estimating historical anthropogenic global sulfur emission patterns for the period 1850–1990. *Journal Atmospheric Environment*, *33*(21), 3435–3444.

Leontief, W. W. (1966). Input-Output Economics. In W. W. Leontief (Ed.), *Input-Output Economics* chapter 2, (pp. 13–29). New York: Oxford University Press.

Li, H., Grijalva, T., & Berrens, R. P. (2007). Economic Growth and Environmental Quality: a Meta-Analysis of environmental Kuznets Curve Studies. *Economics Bulletin*, *17*(5), 1–11.

Li, Q. & Racine, J. S. (2006). *Nonparametric Econometrics: Theory and Practice.* Princeton: Princeton University Press.

Loader, C. (1999). *Local Regression and Likelihood.* New York: Springer-Verlag.

Loh, W.-Y. (2005). GUIDE (version 3) User Manual. NBER Working Paper 10557, Department of Statistics, University of Wisconsin-Madison, Cambridge, MA.

Lopez, R. (1994). The Environment as a Factor of Production: The Effects of Economic Growth and Trade Liberalization. *Journal of Environmental Economics and Management*, *27*, 163–184.

Lopez, R. & Mitra, S. (2000). Corruption, Pollution, and the Kuznets Environment Curve. *Journal of Environmental Economics and Management*, *40*(2), 137–150. available at http://ideas.repec.org/a/eee/jeeman/v40y2000i2p137-150.html.

Love, R. & Wolfson, M. (1976). Income Inequality: Statistical Methodology and Canadian Illustrations. Ottawa: Statistics Canada.

Lucy, D., A. R. & Pollard, A. (2002). Nonparametric calibration for age estimation. *Applied Statistics*, *51*(2), 183–196.

MacKinnon, J. G. (1999). The Linux Operating System: Debian GNU/Linux. *Journal of Applied Econometrics*, *14*(4), 443–452.

Magnani, E. (2000). Environmental Kuznets Curve, Environmental Protection Policy and Income Distribution. *Ecological Economics*, *32*(3), 431–443.

Manski, C. (1988). *Analog Estimation Methods in Econometrics.* London: Chapman and Hall.

Marsaglia, G. (1996). DIEHARD: A Battery of Tests of Randomness. Technical report, The University of Texas, M.D. Anderson Cancer Center, Department of Biomathematics. Available at *http://stat.fsu.edu/pub/diehard/*.

McClements, L. (1977). Equivalence Scales for Children. *Journal of Public Economics*, *44*(2), 191–210.

McCloskey, H. J. (1983). *Ecological Ethics and Politics.* Totowa, N. J.: Rowman and Littlefield.

McCullough, B. (1998). Assessing the Reliability of Statistical Software. *The American Statistician*, *52*, 358–366.

McCullough, B. (1999). Assessing the Reliability of Statistical Software: Part II. *The American Statistician*, *53*(1), 149–159.

McCullough, B. (2000). Is It Safe to Assume That Software is Accurate? *International Journal of Forecasting*, *16*(3), 349–357 (doi:10.1016/S0169–2070(00)00032–7).

McCullough, B. & Renfro, C. (1999a). Benchmarks and Software Standards: A Case Study of GARCH Procedures. *Journal of Economic and Social Measurement*, *25*(2), 59–71.

McCullough, B. & Vinod, H. (1999a). The Numerical Reliability of Econometric Software. *Journal of Economic Literature*, *37*(2), 633–665.

McCullough, B. & Vinod, H. (1999b). The Numerical Reliability of Econometric Software. *Journal of Economic Literature*, *XXXVII*, 633–665.

McCullough, B. & Vinod, H. (2003). Verifying the Solution from a Nonlinear Solver: A Case Study. *The American Economic Review*, *93*(3), 873–892.

McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2006). Lessons from the JMCB Archive. *Journal of Money, Credit, and Banking*, *38*(4), 1093–1107.

McCullough, B. D. & Renfro, C. G. (1999b). Benchmarks and Software Standards: A Case Study of GARCH Procedures. *Journal of Economic and Social Measurement*, *25*(2), 59–71.

Min, I. & Kim, I. (2004). A Monte Carlo Comparison of Parametric and Nonparametric Quantile Regressions. *Applied Economics Letters*, *11*(2), 71–74.

Mookerjee, D. & Shorrocks, A. F. (1982). A Decomposition Analysis of the Trend in UK Income Inequality. *The Economic Journal*, *92*, 886–902.

Munasinghe, M. (1999). Is Environmental Degradation an Inevitable Consequence of Economic Growth: Tunneling Through the Environmental Kuznetz Curve. *Ecological Economics*, *29*(1), 89–109.

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability and its Applications*, *9*, 141–142.

Neumark, D. (1988). Employers Discriminatory Behavior and the Estimation of Wage Discrimination. *The Journal of Human Resources*, *23*(3), 279–295.

Neumayer, E. (2002). National Carbon Dioxide Emissions: Geography Matters. *Area*, *36*(1), 33–40.

Nolan, B. (1988–1989). Macroeconomic conditions and the size distribution of income: evidence from the United Kingdom. *Journal of Post-Keynesian Economics*, *11*, 196–221.

Oaxaca, R. (1973). Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, *14*(3), 693–709.

Pagan, A. & Ullah, A. (1999). *Nonparametric Econometrics*. United Kingdom: Cambridge University Press.

Paglin, M. (1975). The measurement and trend of inequality: a basic revision. *American Economic Review*, *65*, 598–609.

Panayotou, T. (1995). Environmental Degradation at Different Stages of Economics of Economic Development. In I. Ahmed & J. Doeleman (Eds.), *Beyond Rio: The Environmental Crisis and Livelihood in the Third World*. London: Mac Millan Press.

Panayotou, T. (1997). Demystifing the Environmental Kuznetz Curve: Turning a Black Box to a Policy Tool. *Environment and Development Economics*, *2*(40).

Panayotou, T. (2000). Economic Growth, Environment, Kuznets Curve. Technical report, CID Working Paper No. 56.

Payne, R. A. (1995). Freedom and the Environment. *Journal of Democracy*, *6*(3), 41–55.

Peracchi, F. (2001). On estimating conditional quantile and conditional distribution functions.

Pudney, S. (1993). Income and wealth inequality and the life cycle: A nonparametric analysis for china. *Journal of Applied Econometrics*, *8*, 249–276.

Racine, J. (2000). The Cygwin Tools: a GNU Toolkit for Windows. *Journal of Applied Econometrics*, *15*(3), 331–341.

Racine, J. & Hyndman, R. (2002). Using R to Teach Econometrics. *Journal of Applied Econometrics*, *17*(2), 175–189.

Ramanathan, R. (2002). *Introductory Econometrics with Applications* (Fifth ed.). Orlando, Florida: Harcourt College Publishers.

Ravallion, M., Heil, M., & Jalan, J. (2000). Carbon emissions and income inequality. *Oxford Economic Papers*, *52*(4), 651–69. available at http://ideas.repec.org/a/oup/oxecpp/v52y2000i4p651-69.html.

Rogers, J., Filliben, J., Gill, L., Guthrie, W., Lagergren, E., & Vangel, M. (1998). Statistical Reference Datasets for Assessing the Numerical Accuracy of Statistical Software. NIST 1396, National Institute of Standards and Technology, Bethesda.

Rossini, A. J., Heiberger, R. M., Sparapani, R., Mächler, M., & Hornik, K. (2004). Emacs Speaks Statistics: A Multiplatform, Multi-package Development Environment for Statistical Analysis. *Journal of Computational and Graphical Statistics*, *13*(1), 247–261.

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995a). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, *87*(420), 998–1004.

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995b). An effective bandwidth selector for local least squares regression (Corr: 96V91 p1380). *Journal of the American Statistical Association*, *90*, 1257–1270.

Samanta, M. (1989). Non-parametric Estimation of Conditional Quantiles. *Statistics and Probability Letters*, *7*, 407–412.

Sawitzki, G. (1994a). Report on the numerical reliability of data analysis systems. *Computational Statistics and Data Analysis*, *18*(2), 289–301 (doi:10.1016/0167-9473(94)90177-5).

Sawitzki, G. (1994b). Testing Numerical Reliability of Data Analysis Systems. *Computational Statistics and Data Analysis*, *18*(2), 269–286 (doi:10.1016/0167-9473(94)90176-7).

Schmalensee, R., Stoker, T. M., & Judson, R. A. (1998). World Carbon Dioxide Emissions: 1950–2050. *The Review of Economics and Statistics*, *80*(1), 15–27.

Schmitz, H. & Marron, J. (1992). Simultaneous Estimation of Several Size Distributions of Income. *Econometric Theory*, *8*, 476–88.

Scott, D. W. (1979). On Optimal and data-Based Histograms. *Biometrika*, *66*, 605–610.

Scott, D. W. (1992). *Multivariate Density Estimation Theory, Practice, and Visualization*. New York: Wiley.

Selden, T. M. & Song, D. (1994). Environmental Quality and Development: Is There a Kuznets Curve for Air Pollution Emissions? *Journal of Environmental Economics and Management*, *27*(2), 147–162.

Semple, M. (1975). The Effect of Changes in Household Composition on the Distribution of Income 1961–73. *Economic Trends*, *266*, 99–105.

Shafik, N. (1994a). Economic Development and Environmental Quality: An Econometric Analysis. *Oxford Economic Papers*, *46*(0), 757–73. available at http://ideas.repec.org/a/oup/oxecpp/v46y1994i0p757-73.html.

Shafik, N. (1994b). Economic Development and Environmental Quality: An Economic Analysis. *Oxford Economic Papers*, *46*, 757–773.

Sheather, S. J. (2004). Density Estimation. *Statistical Science*, *19*, 588–597.

Sheather, S. J. & Jones, M. C. (1991). A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society, Series B, Methodological*, *53*, 683–690.

Shorrocks, A. F. (1982). Inequality Decomposition by Factor Components. *Econometrica*, *50*(1), 193–212.

Shorrocks, A. F. (1984). Inequality Decomposition by Population Subgroups. *Econometrica*, *52*(6), 1369–1386.

Silverman, B. (1983). Some properties of a test for multimodality based on kernel density estimates. In J. Kingman & G. Reuter (Eds.), *Probability, Statistics and Analysis* (pp. 248–259). Cambridge University Press.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society*, *43*, 823–836.

Silverman, B. W. (1982). Algorithm as 176: Kernel density estimation using the fast fourier transform. *Applied Statistics*, *31*(1), 93–99.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Anaysis*. New York: Chapman and Hall.

Simonoff, J. S. (1999). *Smoothing Methods in Statistics*. New York: Springer-Verlag.

St. Laurent, A. M. (2004). *Understanding Open Source & Free Software Licensing: A Straightforward Guide to the Complex World of Licensing*. Sebastopol, CA, USA: O'Reilly & Associates, Inc.

Stallman, R. (1985). The GNU Manifesto. *10*(3), 30–35.

Stern, D., C. M. B. E. (1996). Economic Growth and Environmental Degradation: The Environmental Kuznets Curve and Sustainable Development. *World Development*, *24*(7).

Stern, D. I. (2004). The Rise and Fall of the Environmental Kuznets Curve. *World Development*, *32*(8), 1419–1439.

Stern, D. I. & Common, M. S. (2001). Is there an environmental kuznets curve for sulfur? *Journal of Environmental Economics and Management*, *41*(2). 162–178.

Stokes, H. (2004). On the Advantage of Using Two or More Econometric Software Systems to Solve the Same Problem. *Journal of Economic and Social Measurement*, *29*, 307–320.

Stokey, N. L. (2001). Are There Limits to Growth? *International Economic Review*, *39*(1), 1–31.

Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, *10*, 1040–1083.

Sturges, H. (1926). The Choice of Class Interval. *journal of the American Statistical Association*, *21*, 65–66. Reprinted in *Statistics in the 21st Century*, pp. 214-228, edited by A. E. Raftery, M. A. Tanner, and M. T. Wells, Chapman & Hall/CRC, New York, 2002.

Suri, V. & Chapman, D. (1998). Economic Growth, Trade and Energy: Implications for the Environmental Kuznets Curve. *Ecological Economics*, *25*(2), 195–208.

Taskin, F. & Zaim, O. (2000). Searching for a Kuznets Curve in Environmental Efficiency Using Kernel Estimation. *Economics Letters*, *68*(2).

Taskin, F. & Zaim, O. (2001). The Role of International Trade on Environmental Efficiency: A DEA Approach. *Economic Modelling*, *18*(1), 1–17.

Torras, M. & J.K., B. (1998). Income, Inequality, and Pollution: a Reassessment of the Environmental Kuznets Curve. *Ecological Economics*, *25*(2), 147–160.

Trede, M. (1998a). Making mobility visible: A graphical device. *Economics Letters*, *59*, 77–82.

Trede, M. M. (1998b). The Age Profile of Mobility Measures: an Application to Earnings in West Germany. *Journal of Applied Econometrics*, *13*(4), 397–409.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information.* Graphics Press.

Ueberhuber, C. W. (1997). *Numerical Computation: Methods, Software, and Analysis*, volume 1. Berlin Heidelberg, Germany: Springer-Verlag.

Varian, H. R. (Ed.). (1996). *Computational Economics: Economic and Financial Analysis with Mathematica.* TELOS/Springer-Verlag.

Venables, W. & Ripley, B. (1999). *Modern Applied Statistics with S-PLUS* (third ed.). New York: Springer-Verlag.

Vincent, J. R. (1997). Testing for Environmental Kuznets Curve within a Developing Country. *Environment and Development Economics*, *2*(4), 417–432.

Vinod, H. D. (2000). Review of GAUSS for Windows, Including its Numerical Accuracy. *Journal of Applied Econometrics*, *14*(2), 211–220.

Vinod, H. D. (2001). Care and Feeding of Reproducible Econometrics. *Journal of Econometrics*, *100*(1), 87–88.

Wand, M. P. (1994). Fast computation of multivariate kernel estimators. *Journal of Computational and Graphical Statistics*, *3*(4), 433–445.

Wand, M. P. & Jones, M. C. (1995). *Kernel Smoothing.* London: Chapman and Hall.

Watson, G. S. (1964). Smooth Regression Analysis. *Sankhyā Ser.*, *26*, 359–372.

Wolfowitz, J. (1942). Additive Partition Functions and a Class of Statistical Hypotheses. *Annals of Mathematical Statistics*, *13*, 247–279.

Wooldridge, J. (2002). *Introductory Econometrics: A Modern Approach* (Second ed.). Mason, Ohio: Thomson, South-Western.

World Bank (1992). *The World Banks World Development Report 1992: Development and the Environment.* Washington, DC: Oxford University Press.

Yatchew, A. (1999). *Semiparametric Regression for the Applied Econometrician.* United Kingdom: Cambridge University Press.

Zhang, Z. (2000). Decoupling Chinas carbon emissions increase from economic growth: An economic analysis and policy implications. *World Development*, *28*(4), 739–752.

# Appendices

Appendix A

# Perl Code for LRE Routine

```
sub re {
    my ( $est, $cert ) = @_;
    return abs( $cert - $est ) / abs($cert);
}

sub log10 {
    my $n = shift;
    return log($n) / log(10);
}

sub lre {
    my ( $est, $cert, $nosd ) = @_;
    my $aest = abs(est);
    if ( $cert == 0 ) {
        if ( abs($est) > 1 ) {
            return 0;
        }
        else {
            ( -log10($aest) < $nosd )
              ? return -log10($aest)
              : return $nosd;
        }
    }
```

```
    elsif ( $cert == $est ) {
        return $nosd;
    }
    elsif ( abs( $est / $cert ) > 2 || abs( $est / $cert ) < 1 / 2 ) {
        return 0;
    }
    else {
        ( -log10( re( $est, $cert ) ) < $nosd )
          ? return -log10( re( $est, $cert ) )
          : return $nosd;
    }
}
```

Appendix B

# Old Faithful geyser data

The original dataset contains two variables: eruption duration and waiting times. We retain only one for our purposes. $X_i$ is the duration in minutes of an eruption of the Old Faithful geyser in the Yellowstone National Park.

Table B.1: Old Faithful Test Data

| $i$ | $X_i$ | $i$ | $X_i$ | $i$ | $X_i$ | $i$ | $X_i$ |
|---|---|---|---|---|---|---|---|
| 61 | 2.233 | 62 | 4.500 | 63 | 1.750 | 64 | 4.800 |
| 65 | 1.817 | 66 | 4.400 | 67 | 4.167 | 68 | 4.700 |
| 69 | 2.067 | 70 | 4.700 | 71 | 4.033 | 72 | 1.967 |
| 73 | 4.500 | 74 | 4.000 | 75 | 1.983 | 76 | 5.067 |
| 77 | 2.017 | 78 | 4.567 | 79 | 3.883 | 80 | 3.600 |
| 81 | 4.133 | 82 | 4.333 | 83 | 4.100 | 84 | 2.633 |
| 85 | 4.067 | 86 | 4.933 | 87 | 3.950 | 88 | 4.517 |
| 89 | 2.167 | 90 | 4.000 | 91 | 2.200 | 92 | 4.333 |
| 93 | 1.867 | 94 | 4.817 | 95 | 1.833 | 96 | 4.300 |
| 97 | 4.667 | 98 | 3.750 | 99 | 1.867 | 100 | 4.900 |
| 101 | 2.483 | 102 | 4.367 | 103 | 2.100 | 104 | 4.500 |
| 105 | 4.050 | 106 | 1.867 | 107 | 4.700 | 108 | 1.783 |
| 109 | 4.850 | 110 | 3.683 | 111 | 4.733 | 112 | 2.300 |
| 113 | 4.900 | 114 | 4.417 | 115 | 1.700 | 116 | 4.633 |
| 117 | 2.317 | 118 | 4.600 | 119 | 1.817 | 120 | 4.417 |
| 121 | 2.617 | 122 | 4.067 | 123 | 4.250 | 124 | 1.967 |
| 125 | 4.600 | 126 | 3.767 | 127 | 1.917 | 128 | 4.500 |
| 129 | 2.267 | 130 | 4.650 | 131 | 1.867 | 132 | 4.167 |
| 133 | 2.800 | 134 | 4.333 | 135 | 1.833 | 136 | 4.383 |
| 137 | 1.883 | 138 | 4.933 | 139 | 2.033 | 140 | 3.733 |
| 141 | 4.233 | 142 | 2.233 | 143 | 4.533 | 144 | 4.817 |
| 145 | 4.333 | 146 | 1.983 | 147 | 4.633 | 148 | 2.017 |
| 149 | 5.100 | 150 | 1.800 | 151 | 5.033 | 152 | 4.000 |
| 153 | 2.400 | 154 | 4.600 | 155 | 3.567 | 156 | 4.000 |
| 157 | 4.500 | 158 | 4.083 | 159 | 1.800 | 160 | 3.967 |
| 161 | 2.200 | 162 | 4.150 | 163 | 2.000 | 164 | 3.833 |
| 165 | 3.500 | 166 | 4.583 | 167 | 2.367 | 168 | 5.000 |
| 169 | 1.933 | 170 | 4.617 | 171 | 1.917 | 172 | 2.083 |
| 173 | 4.583 | 174 | 3.333 | 175 | 4.167 | 176 | 4.333 |
| 177 | 4.500 | 178 | 2.417 | 179 | 4.000 | 180 | 4.167 |
| 181 | 1.883 | 182 | 4.583 | 183 | 4.250 | 184 | 3.767 |

*Continued on next page*

| $i$ | $X_i$ | $i$ | $X_i$ | $i$ | $X_i$ | $i$ | $X_i$ |
|---|---|---|---|---|---|---|---|
| 185 | 2.033 | 186 | 4.433 | 187 | 4.083 | 188 | 1.833 |
| 189 | 4.417 | 190 | 2.183 | 191 | 4.800 | 192 | 1.833 |
| 193 | 4.800 | 194 | 4.100 | 195 | 3.966 | 196 | 4.233 |
| 197 | 3.500 | 198 | 4.366 | 199 | 2.250 | 200 | 4.667 |
| 201 | 2.100 | 202 | 4.350 | 203 | 4.133 | 204 | 1.867 |
| 205 | 4.600 | 206 | 1.783 | 207 | 4.367 | 208 | 3.850 |
| 209 | 1.933 | 210 | 4.500 | 211 | 2.383 | 212 | 4.700 |
| 213 | 1.867 | 214 | 3.833 | 215 | 3.417 | 216 | 4.233 |
| 217 | 2.400 | 218 | 4.800 | 219 | 2.000 | 220 | 4.150 |
| 221 | 1.867 | 222 | 4.267 | 223 | 1.750 | 224 | 4.483 |
| 225 | 4.000 | 226 | 4.117 | 227 | 4.083 | 228 | 4.267 |
| 229 | 3.917 | 230 | 4.550 | 231 | 4.083 | 232 | 2.417 |
| 233 | 4.183 | 234 | 2.217 | 235 | 4.450 | 236 | 1.883 |
| 237 | 1.850 | 238 | 4.283 | 239 | 3.950 | 240 | 2.333 |
| 241 | 4.150 | 242 | 2.350 | 243 | 4.933 | 244 | 2.900 |
| 245 | 4.583 | 246 | 3.833 | 247 | 2.083 | 248 | 4.367 |
| 249 | 2.133 | 250 | 4.350 | 251 | 2.200 | 252 | 4.450 |
| 253 | 3.567 | 254 | 4.500 | 255 | 4.150 | 256 | 3.817 |
| 257 | 3.917 | 258 | 4.450 | 259 | 2.000 | 260 | 4.283 |
| 261 | 4.767 | 262 | 4.533 | 263 | 1.850 | 264 | 4.250 |
| 265 | 1.983 | 266 | 2.250 | 267 | 4.750 | 268 | 4.117 |
| 269 | 2.150 | 270 | 4.417 | 271 | 1.817 | 272 | 4.467 |

Appendix C

# R Code for Banking to 45 degrees

```
### Banking to 45 degrees function
b45 <- function( a, x, y ) {
    vbi <- abs(diff(y)) / ( max(y) - min(y) )
    hbi <- abs(diff(x)) / ( max(x) - min(x) )
    sum( atan( a * vbi/hbi ) * sqrt( hbi^2 + a^2*vbi^2) ) /
    sum( sqrt( hbi^2 + a^2*vbi^2 ) )  - pi/4
}

ye <- c(2.75, 3.75, 3)
xe <- c(7.8,  9.4, 10.1)
ar <- uniroot( b45 , x=xe, y=ye, lower=-10, upper=10 )$root
```

Appendix D

# Parametric Quantile Regression Approach

In this section we use a parametric approach to estimate conditional measures of inequality following an analogous approach to the one developed in This chapter. We use the parametric quantile regression method of Koenker & Bassett (1978) to directly estimate the conditional quantiles. Parametric versions corresponding to our conditional measures of income inequality are then derived from the parametric conditional quantiles.

Quantile regression has emerged as an influential tool of empirical economics in recent years. For a recent series of applications of quantile regression in economics see, e.g., Koenker & Hallock (2001) and Fitzenberger et al. (2002).

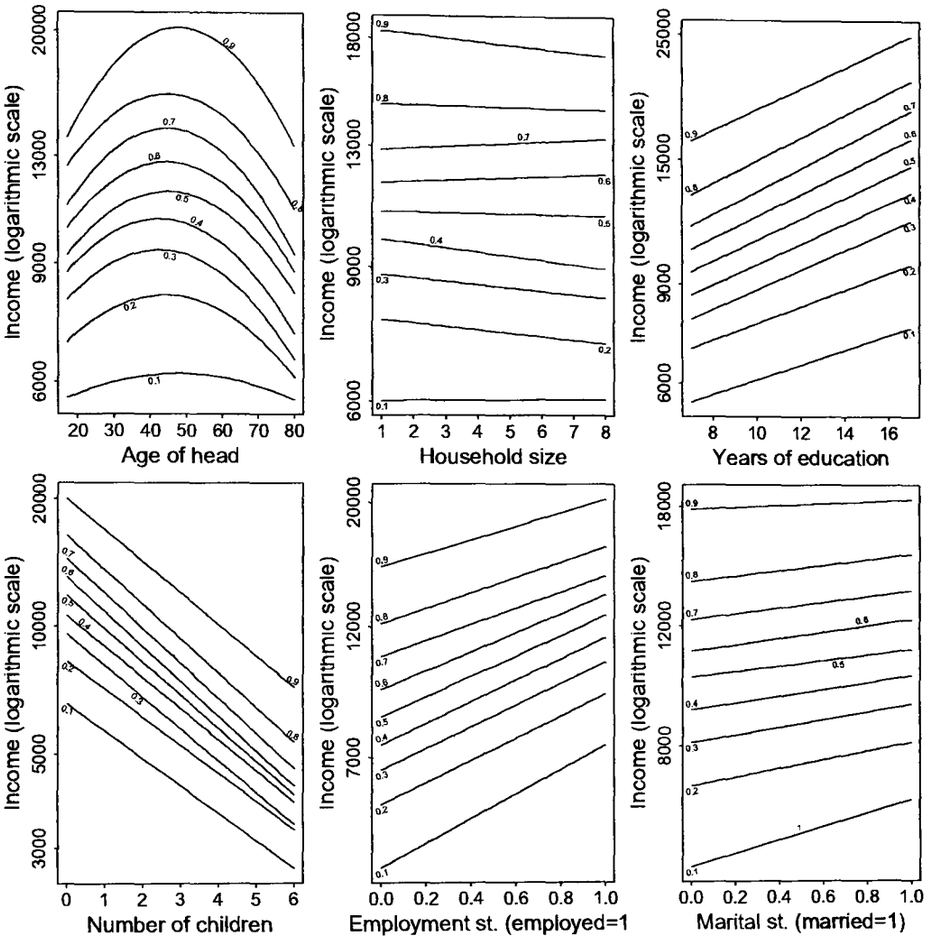
To estimate the conditional quantiles we used GNU R's implementation of the Barrodale & Roberts (1973) algorithm for least absolute deviation regression extended to linear quantile regressions as described in Koenker & d'Orey (1987, 1994). This particular algorithm can handle problems involving up to several thousand observations. It also implements a scheme for computing confidence intervals for the estimated parameters, based on inversion of a rank test described in Koenker (1994).[1]

Figure D.1, shows the estimated quantiles of log income conditional on age of household head, household size, years of education of the households head, number of young children, employment status of the head, and on the marital status of the households head, obtained with the parametric regression quantile approach. These estimates are used to derive, together with asymptotic confidence intervals, parametric equivalents to our conditional inequality measures shown in Figures D.2 and D.3 on page 240 and page 241, respectively.

We can see that the overall trend of the parametric conditional deciles is almost identical to the corresponding nonparametric quantile estimates presented in Figure 4.6 on page 119 of This thesis. In particular, age-income profiles have both an inverted-U shape, the number of young children-income profiles are negatively sloped, household size-income profile are relatively flat and both with increasing and decreasing deciles, and education-income, employment status-income, and marital status-income profiles are all increasing. This finding is reassuring and provided support reinforces our findings.

[1] We used R release 2.3.0, the standard Win32 release available at the time of writing the chapter, together with the routines to obtain quantile regression coefficients and standard errors provided by the `quantreg` R package, version 4.08 developed by Roger Koenker.

Figure D.1: Estimated parametric regression deciles of the conditional distribution function of log income




The results for the derived conditional inequality measures are less clear. Figure D.2 and D.3 display the estimated inequality profiles. There seems to be a similar trend for the Conditional Relative InterQuartile Range (CRIQR) measures of inequality, looking at the center of the distribution, of the nonparametric approach with the the Conditional Decile Dispersion Ratio (CDDR) measures of inequality, focusing on the tails, of the parametric approach. In particular, the parametric conditional measures show that older household heads, larger household sizes, more years of education, and being married, have all a negative impact on inequality for the center of the income distribution, having more young children has a slightly positive impact. This findings agree with the results from the nonparametric conditional deciles. Only the result for the employment status differ

markedly as it appears to be both economically and statistically non significant for the parametric case. Though a more careful analysis would be required to support this argument more convincingly, these preliminary findings seem to suggest that the parametric approach based on least absolute deviations might be too sensitive to influential observations, so that smaller local changes are swamped by "non-local" effects.

As results differ quite substantially, it is reasonable to conclude that the parametric specification would require difficult *ad-hoc* assumptions to match the nonparametric results, but further analysis is required to support this view. The parametric approach can be more efficient assuming that the underlying maintained model assumptions hold, but can be potentially misleading otherwise. A recent study comparing parametric and nonparametric quantile regression methods using a Monte Carlo approach by Min & Kim (2004), found some evidence of superiority of the nonparametric quantile regression approach particularly when the underlying model is nonlinear or the error terms are not normally distributed. Based on the above considerations, it would seem more appropriate to use a nonlinear quantile regression approach (see, e.g., Busovaca, 1985, and references therein) in this case for a more fruitful comparison. This is an interesting investigation worthy of further research.

Figure D.2: Parametric based conditional measures of income inequality on age of head, household size, and years of education with one standard deviation confidence intervals

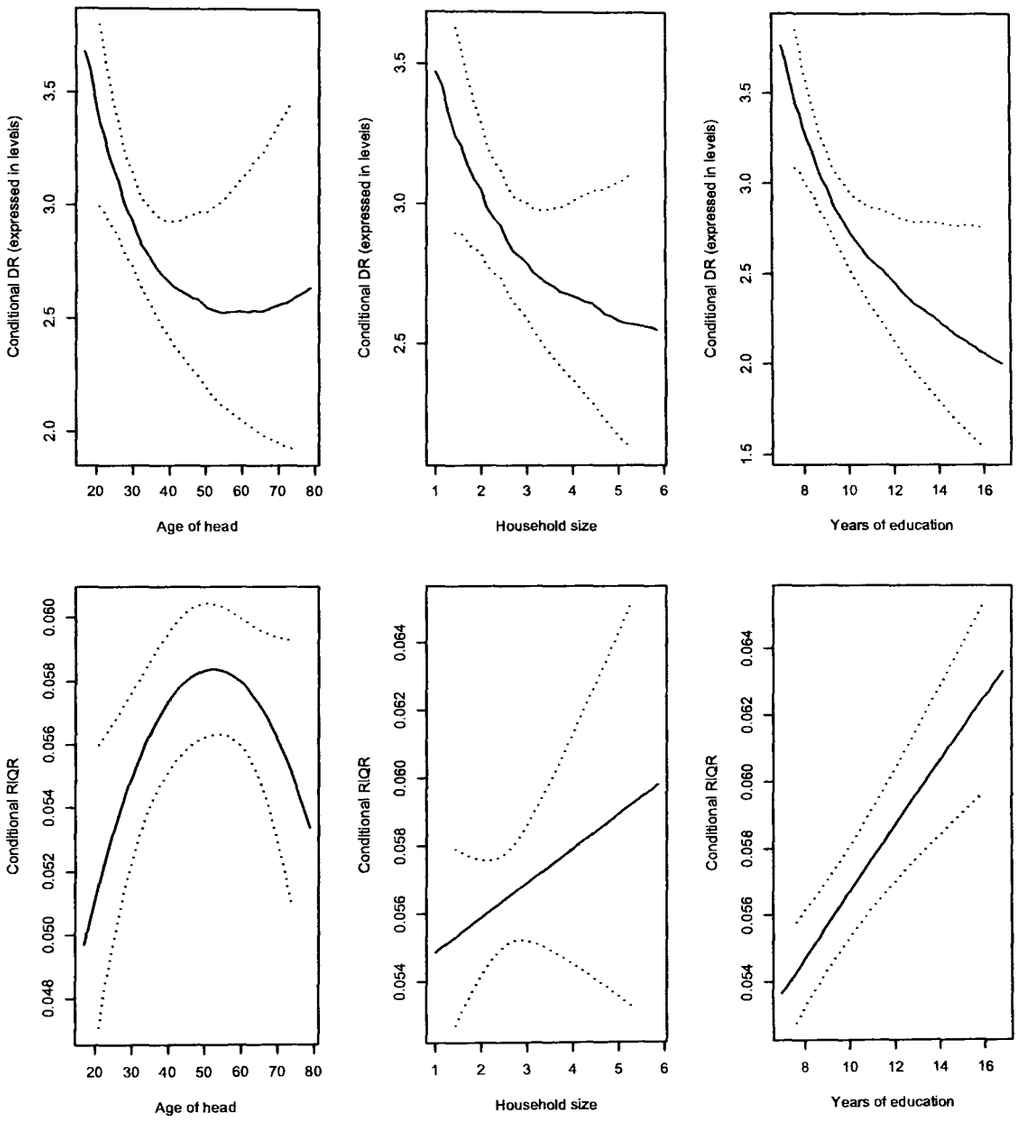
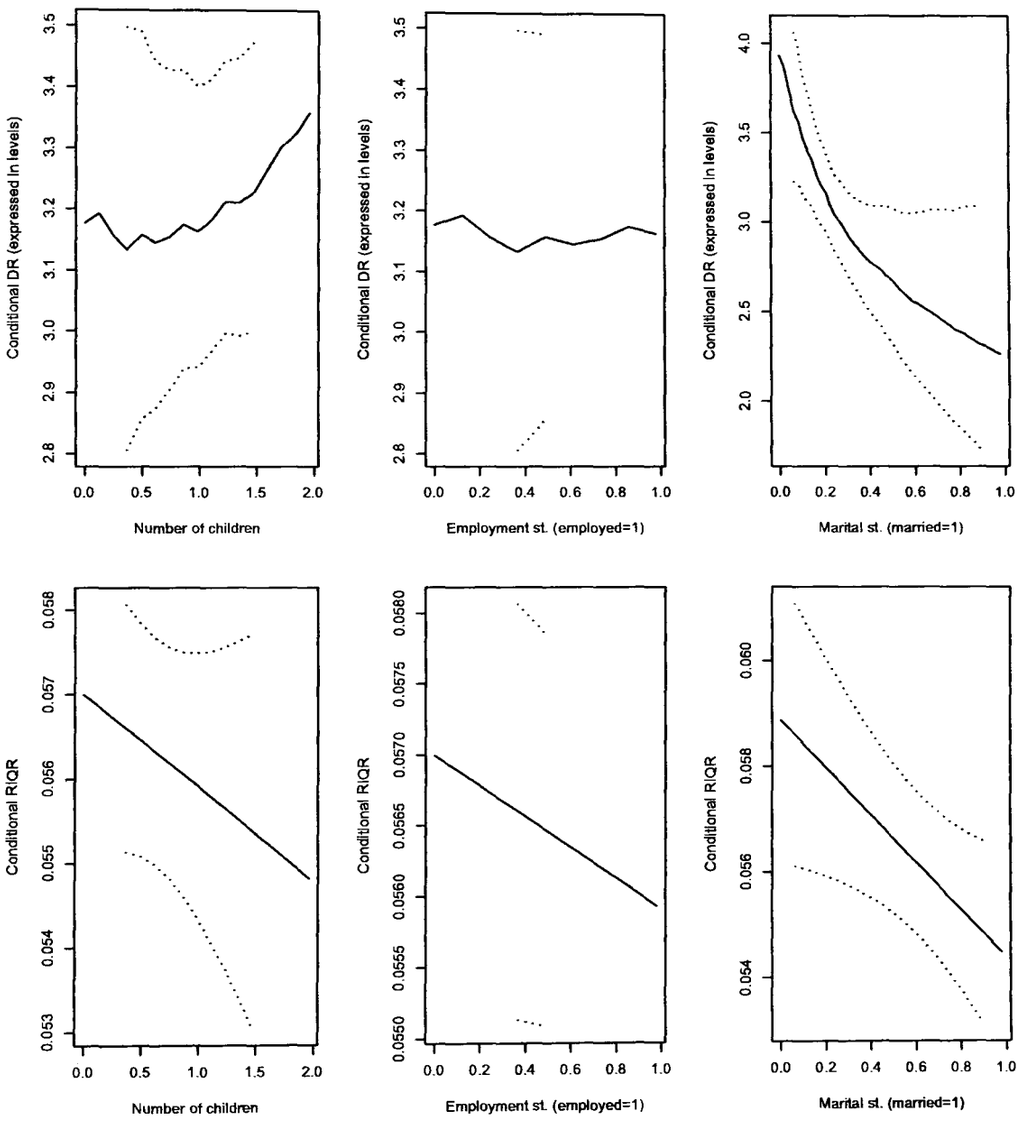


Figure D.3: Parametric based conditional measures of income inequality on number of children, employment and marital status, keeping all other determinants fixed at their respective mean values



Appendix E

# Logit Parameter Estimates

Figure E.1 plots the estimated parameters $\widehat{\beta}$s for the ($J = 25$) chosen evaluation points. The dotted lines represent the confidence bands ($\pm 1.96 \times$standard errors) calculated for each evaluation point.

In general, if the sign of the estimated coefficient is negative (positive) and its value is significantly different from zero, it means that an increase in the variable shifts the distribution to the right (left).

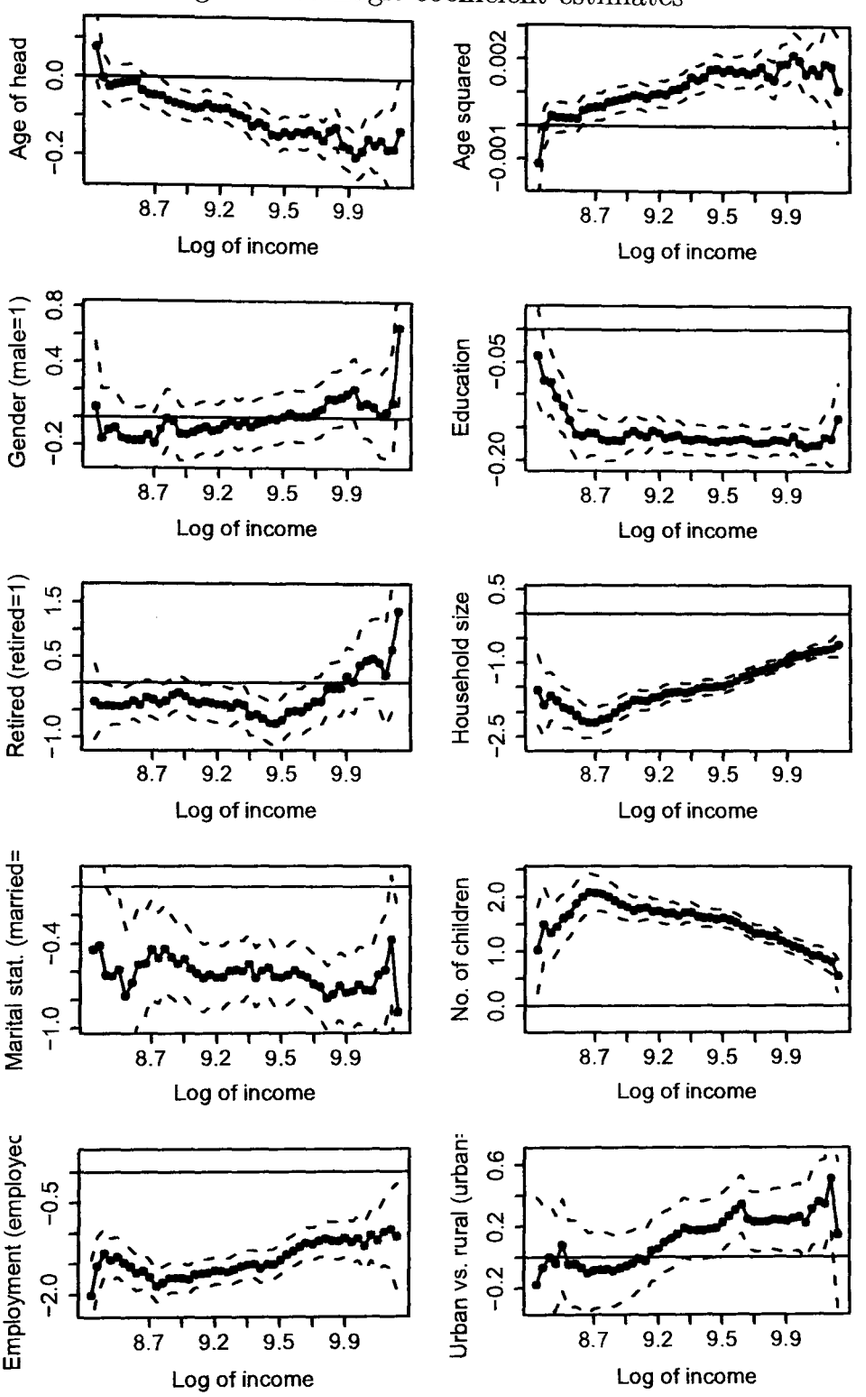
Thus, as expected, an increase in education and in the percentage of being employed determines a shift of the distribution to the right. Also, this means that it will be less probable for the individual to fall below the vertical line highlighted in the pictures. Other variables, whose increase determines a shift to the right in the distribution are the family size and the fact of being married.

Conversely, the only variable whose increase causes a shift to the left of the distribution of income is the number of people in the household with less than 16 years of age.

For gender, retired and urban/rural indicator a different pattern is observed. The sign of the estimated coefficients change from negative to positive and the values are frequently not statistically different from zero. This implies that the distribution of income shrinks as the value of those variables increases.

Finally, age starts from a negative value, then sharply turns positive and starts decreasing again, turning negative and also losing statistical significance. The result in the conditional distribution of an increase of age is an increase in the spread of the distribution (especially in the tails) and a shift to the left of the centered 60% of the probability mass (comprised between the 20% and 80% quantiles).

Figure E.1: Logit coefficient estimates



Appendix F

# Countries Included in the Dataset

Table F.1: Country Codes

| | | | |
|---|---|---|---|
| 1 | ALGERIA | 95 | JAPAN |
| 14 | EGYPT | 97 | KOREA, |
| 18 | GHANA | 98 | KUWAIT |
| 22 | KENYA | 100 | MALAYSIA |
| 25 | MADAGASCAR | 102 | MYANMAR |
| 30 | MOROCCO | 106 | PHILIPPINES |
| 31 | MOZAMBIQUE | 108 | SAUDI ARABIA |
| 32 | NAMIBIA | 109 | SINGAPORE |
| 34 | NIGERIA | 110 | SRI LANKA |
| 41 | SAFRICA | 111 | SYRIA |
| 44 | TANZANIA | 112 | TAIWAN |
| 46 | TUNISIA | 113 | THAILAND |
| 48 | ZAIRE | 116 | AUSTRIA |
| 49 | ZAMBIA | 117 | BELGIUM |
| 50 | ZIMBABWE | 119 | CYPRUS |
| 52 | BARBADOS | 120 | CZECHOSLOVAKIA |
| 54 | CANADA | 121 | DENMARK |
| 60 | GUATEMALA | 122 | FINLAND |
| 62 | HONDURAS | 123 | FRANCE |
| 64 | MEXICO | 125 | WGERMANY |
| 65 | NICARAGUA | 126 | GREECE |
| 71 | TRINIDAD&TOBAGO | 129 | IRELAND |
| 72 | U.S.A. | 130 | ITALY |
| 73 | ARGENTINA | 131 | LUXEMBOURG |
| 74 | BOLIVIA | 133 | NETHERLANDS |
| 75 | BRAZIL | 134 | NORWAY |
| 76 | CHILE | 136 | PORTUGAL |
| 77 | COLOMBIA | 137 | ROMANIA |
| 81 | PERU | 138 | SPAIN |
| 83 | URUGUAY | 139 | SWEDEN |
| 84 | VENEZUELA | 140 | SWITZERLAND |
| 88 | CHINA | 141 | TURKEY |
| 89 | HONG KONG | 142 | U.K. |
| 90 | INDIA | 143 | USSR |
| 91 | INDONESIA | 144 | YUGOSLAVIA |
| 92 | IRAN | 145 | AUSTRALIA |
| 94 | ISRAEL | 147 | NZ |

# Index