1

Introduction

1.1 Overview

The goal of this book is to help bridge the gap between applied economists and theoretical nonparametric econometricians/statisticians. The majority of empirical research in economics ignores the potential benefits of nonparametric methods and many theoretical nonparametric advances ignore the problems faced by practitioners. We do not believe that applied economists dismiss these methods because they do not like them. We believe that they do not employ them because they do not understand how to use them or lack formal training on kernel smoothing. Many theoretical articles and texts that develop nonparametric methods claim that they are useful to empirical researchers, which they often are, but many times the level of mathematics is too high for a typical economist or the detail with which the practical considerations are explained is not adequate except for those well versed in econometrics. At the same time, many of these articles and textbooks skip (or do not have room to include) the nuances of the methods which are necessary for doing solid empirical research.

Although nonparametric kernel methods have been around for nearly six decades, their use in economics journals did not become popular until the twenty-first century (noting that there were influential papers prior to 2000). In our opinion, two major developments have drastically increased the use of nonparametric methods in economics. The first is obvious: computing power. Without computers that can quickly provide estimates (coupled with efficient code), these methods are largely impractical for applied work. Of course, we cannot discount the importance of complementary statistical packages for nonparametric methods, such as the popular `np` package (Hayfield and Racine, 2008) in the `R` language (2012). The combination of higher-powered computers and available software has done much to popularize the methods across academic fields.

For economics, the second reason we believe this prevalence has increased of late is the assortment of new estimators which allow researchers to handle discrete data. We know that economic data is generally a combination of continuous and discrete variables. In the past, authors who wanted to use discrete data had to resort to semiparametric methods with little reasoning other than they did not know how to handle discrete data nonparametrically. This required stringent and sometimes unjustified assumptions on the data. For instance, having a dummy variable enter the regression linearly assumes that it is separable from the variables in the nonparametric function and that the only difference between groups is an intercept shift. Neither of these

assumptions need hold true for any particular data set. It is not as if authors using these methods necessarily believed this to be true; they simply did not have many options for how to handle this type of data (see Li and Racine, 2007, for a great introduction to nonparametric estimation with discrete data).

Nonparametric methods have advanced to a point where they are of use to applied economists and computers have advanced to a point where using the methods are feasible. In this book we plan to discuss in depth, and in terms that someone with one year of graduate econometrics can understand (say at the level of Greene, 2011), basic to advanced nonparametric methods. Our analysis starts with density estimation in the crudest sense and motivates the procedures through methods that the reader should be familiar with. We then move onto kernel regression, estimation with discrete data, and advanced methods like estimation with panel data and instrumental variables. We spend a lot of time discussing kernel choice and bandwidth selection as well as why one method may be preferable in one setting or another. We also pay close attention to the issues that arise with programming, computing speed, and application. In each chapter we keep derivations to a minimum, but make available on the web the derivations (without skipping steps) of our results. We will give the intuition in the text without the full brunt of the math, but the step-by-step derivation in the online appendix should be a useful learning tool for those who wish to gain additional insight.

Given that we wish to teach nonparametric methods to applied economists, we must include applications. However, as opposed to giving a bunch of simple applications without much insight, we focus on one particular topic that we have researched extensively: economic growth. In each chapter, we apply the methods we discuss to actual data. Given that our focus throughout the book is with respect to economic growth, we take publicly available data and attempt to perform proper applications. We not only show how the methods work in practice, but we also uncover results that have not been studied or that contradict the findings of previous studies. In this respect, we believe that the application sections are of interest by themselves. Also, the data and `R` code, which can be used to replicate the empirical results in the application section of each chapter (we have done our best to ensure this – e.g., set seeds), can be found on the text's website (http://www.the-smooth-operators.com).

Our hope is that once the readers have finished the first few parts of the book that they will be able to apply these methods to their specific problems, either in the field of economic growth or other areas of economics. We believe that it will be relatively straightforward to apply these methods to most data sets by taking the code available online and making minor modifications as necessary. We hope that this text will help increase the number of applications of nonparametric methods in economics. These powerful tools are widely available in today's applied environment, but we envision that they will be understood by a larger audience. Although statistical packages are essential to the promotion of one's research, this will not result in better research unless users are well-informed about the strengths and limitations of the methods.

1.2 Birth of the text

There was no single defining moment which prompted us to write this text. Most of the reasons came about as we conducted our own research. There were countless times

when we were presented with situations that we did not know how to resolve. This was particularly the case in our applied work. Earlier in our careers, a cheap way to get a publication was to read *Econometrica*, the *Journal of the American Statistical Association*, or other high-level statistics/econometrics journals and to code a newly published estimator and apply it to a well-known data set. We would first replicate the results of a Monte Carlo simulation and then we would take that same estimator and run it through a proper data set. We were often confronted with two situations: (1) many times the estimators worked well in simulations (often with a single covariate), but performed poorly with real data (this was often true with less "well-behaved" data); and (2) to analyze the economic results, we were often left without ways to appropriately dissect estimated gradients or it was unclear how to present the results we had. We therefore needed to determine ways to "empiricize" theoretical works. Now, while some of these empirical advances are better than others, we typically noted in our papers that the code was available upon request. We are happy to note that many authors have made use of these offers. The benefit is that this increased citations; the downfall is more referee reports (although these are also beneficial). The combination of these events led us to think that there would be a demand for this type of text. In fact, while writing the text, we also had to figure out ways to "empiricize" estimators that we had not used in the past or required us to think differently about estimators we had used before.

1.3 Who will benefit

In addition to economists, the book may also be useful to researchers in other fields who use typical econometric tools (e.g., regression), such as political science, history, and applied statistics. We feel it will be useful to faculty and graduate students alike. Specifically, the reader should have at least one course in mathematical statistics and one course in linear regression. It would also be helpful, but not necessary, to have had a course in nonlinear regression.

We expect that this book could be part of either a third or fourth semester econometrics course. This text could be used to teach an applied nonparametric econometrics course or as part of a course on applied nonlinear methods. It could be used by itself or paired with a complementary nonparametric book like Li and Racine (2007) or a book covering nonlinear regression methods like Cameron and Trivedi (2005). It is unlikely that this text will be used to teach a more theory-driven econometrics course. There are books in the literature that are more tailored to that approach and we are not pretending to be theoretical econometricians. We are applied econometricians and our expertise is in applying nonparametric methods to data. It is in this realm where our comparative advantage lies. Nevertheless, we make an attempt to explain theoretical concepts in an intuitive way.

1.4 Why this book is relevant

We have noted the problems that most applied economists have with applying nonparametric methods. We have run into many of these problems ourselves in our own work. Here we plan to lay everything out so that you will know how to apply them. We

feel that without this text, the cost of learning how to use nonparametric methods with actual data will be too high for many economists. We hope that this text will decrease this cost.

In addition to presenting the material at a different level, we also introduce or further discuss methods that are not in current nonparametric textbooks. This field changes rapidly and hence there appears to be reason to update or write new texts relatively often (note that the book is not meant to be comprehensive, so we will leave out many great papers). For example, our text spends a large amount of time on panel data methods. In particular, we present our panel data estimators for the unbalanced case (noting that the number of journal articles with unbalanced panel in the nonparametric setting is small). Another area which separates us from past textbooks is a chapter on constrained nonparametric estimation. Nonparametric methods relax functional form assumptions, but it is often the case that this leads to violations of economic theory that we assume to hold true. We discuss methods to impose constraints in a nonparametric framework in Chapter 12. In addition, in many of the chapters we provide useful, straightforward tables or derivations that do not exist in the literature and which we believe will be helpful to the practitioner.

Finally, we focus much more on the application of nonparametric methods and what the theory means in terms of the application of such methods. We not only talk about estimation and testing, but we also spend a lot of time on how to present the results. This is often overlooked or results are given for the univariate case (either by considering a single variable or by using counterfactual analysis). While there is nothing inherently wrong with this approach, most economic data sets contain many variables and this often makes presentation difficult. That being said, we do not want our text to be seen as a "cook book." Yes, we discuss how to estimate, test, and present results, but we also try to incorporate intuition from the theoretical underpinnings of these same estimators and tests so that authors can be well informed when they employ these methods.

1.5 Examples

To give a sense of what is to come, and to show how the methods can be applied in many fields, in this section we consider three simple examples. The first is with respect to density estimation and the latter two are via nonparametric regression. These are each univariate examples and so they allow us to show the results in two-dimensional figures. We clearly are ignoring many other factors in each case, but hope that they give you an idea of how powerful the estimators are and what they may be able to accomplish.

1.5.1 CO_2 emissions

Our first example is in the area of environmental economics. Specifically, we have data (Boden, Marland, and Andres, 2011) on per-capita CO_2 emissions from 152 countries in 1960 and 2005 (balanced sample). The solid line in Figure 1.1 is the kernel density estimate (smoothed histogram) of per-capita CO_2 emissions across 152 countries in

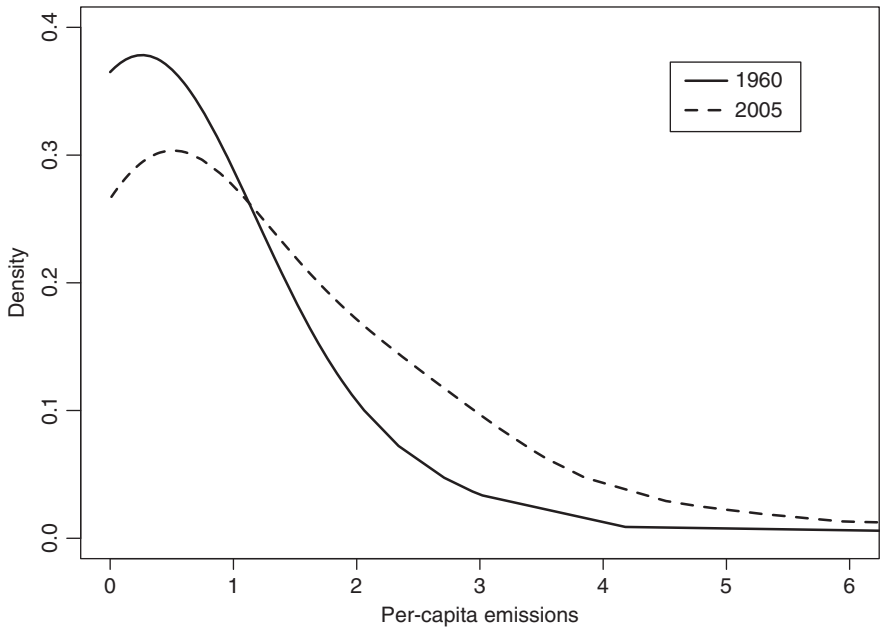

Figure 1.1. Density plots of per-capita CO_2 emissions for 152 countries in the years 1960 and 2005

1960. The dashed line is the kernel density estimate for the same countries in 2005. We see that the 2005 density of per-capita emissions has a shorter mode and is shifted to the right of the 1960 density. Note that each of the densities has a long right-hand-side tail that we cut off to make the figures easier to read. The increase in overall emissions is expected on a national level, but what this shows is that even though population growth has increased dramatically over the 45-year period (from around three billion inhabitants worldwide in 1960 to roughly seven billion in 2005), per capita emissions have still grown. This perhaps shows why there is so much interest in decreasing emissions across the globe.

1.5.2 Age earnings

The first regression example is an application known to many nonparametric econometricians. This example was first presented in Ullah (1985) and it appeared in other papers he has written as well as in his textbook (Pagan and Ullah, 1999, 155). It has also been used as an example in other textbooks (e.g., Ruppert, Wand, and Carroll, 2003, 117), research articles (e.g., Henderson and Parmeter, 2009, 463), and software packages (e.g., the `np` package in `R` – Hayfield and Racine, 2008, 11).

In a long line of literature, both econometricians and labor economists have argued over the relationship between earnings and experience. Heckman and Polachek (1974), among others, have argued for a polynomial function. Ullah (1985) examined this assumption by taking the 1971 Canadian Census Public Use Tapes to relate earnings versus experience (age in his application). To be able to illustrate the example in two dimensions, he selected the 205 males in the data set with thirteen years of education.

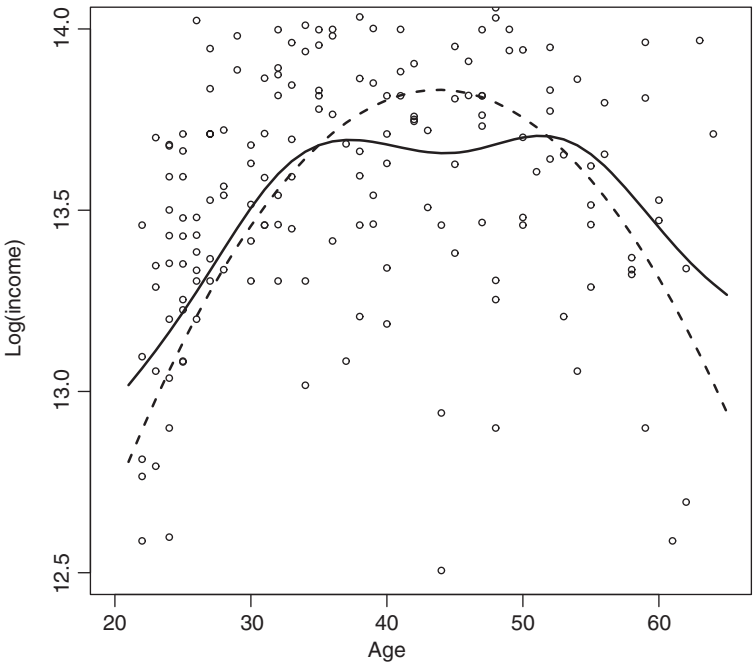

Figure 1.2. Scatterplot of log(income) and age with both local-constant (with rule-of-thumb bandwidth) and quadratic parametric regression fits (1971 Canadian Census Public Use Tapes)

Figure 1.2 is a replica (we additionally include a scatter plot) of that in Pagan and Ullah (1999). The dashed curve is the quadratic fit, whereas the solid line is a nonparametric fit (local-constant kernel fit with a rule-of-thumb bandwidth). The nonparametric fit has a less dramatic increase at lower ages and a less dramatic decrease at higher ages, but the main deviation is the flat portion of the curve from roughly age 35 to 55. If we ignore the slight dip, then this suggests that wages flatten out in mid-career as opposed to continuously changing as is assumed in the quadratic specification. Clearly, both models are simplistic given that there are many other factors which determine wages and hence a more comprehensive analysis is necessary in practice.

1.5.3 Hedonic price function

The final example we give comes from urban economics. Here we consider a hedonic price function for housing. In his seminal work on the theory of hedonic prices, Rosen (1974) suggested that "it is inappropriate to place too many restrictions on [the hedonic price] at the outset." Given that the form of the hedonic price function for housing is unknown (as well as the vast amounts of data available – preferred for nonparametric estimation), this is a great place for nonparametric methods to contribute.

Here we take a portion of the data in Anglin and Gençay (1996). In their paper they model the logarithm of housing price as a semiparametric function of standard

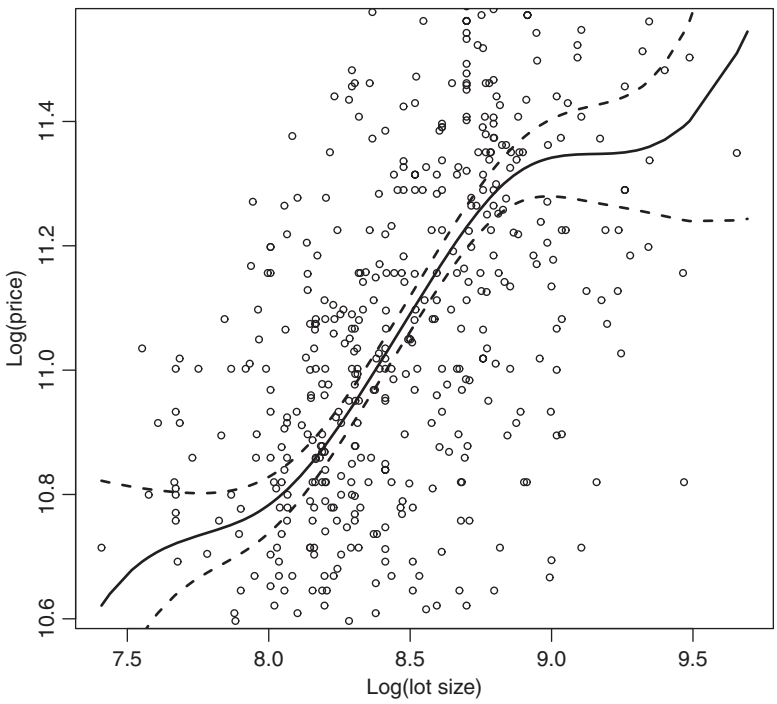

Figure 1.3. Scatterplot of log(price) and log(lot size) with a local-linear (with least-square cross-validation bandwidth) regression fit and confidence intervals (estimated via 399 bootstrap replications)

attributes of a home (number of bedrooms, bathrooms, etc.). The data comprise 546 observations from 1987 in the Windsor (Canada) housing market. Here we take one of their primary variables of interest (logarithm of lot size) as our single explanatory variable.

Figure 1.3 shows the results of a (local-linear) nonparametric regression (with a data-driven bandwidth – the least-squares cross-validation bandwidth is 0.2408). If we were to believe this simple regression, it would suggest that there are plateaus in the price of the home with respect to lot size. In other words, at relatively small and relatively large lot sizes, we see very small changes in price, whereas for intermediate lot sizes we see larger increases in price (which also appears linear in this range). Although this example is simplistic, the function is monotonic (as theory would predict) and we see intuition consistent with hedonic theory. See Parmeter, Henderson, and Kumbhakar (2007) for a full nonparametric examination with the Anglin and Gençay (1996) data.

1.6 Examples in the text

Our approach to present applications illustrating each of the estimators/tests discussed in the text differs from what is typically done. Instead of taking many different applications, we decided to have a singular focus with our illustrations of the methods.

Specifically, we will look at economic growth/output; however, we will be even more specific than that. In the density chapters we will examine the distribution of output per worker and in the regression chapters we will (primarily) look at modeling the worldwide production function.

There are several benefits to this approach, which include the following:

1. The Penn World Table (PWT) is a well-known publicly available data set.
2. Our own research lies in this area and thus we should be able to explain the applications better than we could if we were to examine many different applications.
3. The sample sizes and number of covariates are relatively small and hence replication of our results will not be as time consuming.
4. We are able to uncover findings not previously seen or discussed in the literature.

We feel that this approach will show the methods in their true light (both positive and negative). It is the case that they are not always perfect (as no estimation methodology is). However, the estimators and tests that we discuss are ones that we feel (typically) work well in practice. It also gives authors help on what to do when they are stuck in such a situation. This actually may be more helpful than showing when they do work.

1.6.1 Density

In the beginning of the text, we will focus on the worldwide distribution of labor productivity (output per worker). The main motivation for such an analysis comes from the work of Danny Quah. In a series of papers (Quah, 1993a,b; 1996; 1997), he argues that standard regression methods cannot always address issues relating to aspects of the entire distribution of the world's economies. To assume that the evolution of the distribution of output per worker can be boiled down to a single parameter estimate is naïve (noting that a growth regression with high explanatory power would be very informative regarding the evolution over time). These distributions are often multimodal, and examining the first one or two moments of the distribution is only appropriate when the distribution is Gaussian. He instead advocates examining kernel density estimates where he visually finds two modes in the distribution. These are later confirmed statistically by Bianchi (1997); Henderson, Parmeter, and Russell (2008); and Parmeter (2008), among others.

Our examples in the chapters concerning density estimation examine what was shown in the aforementioned papers, albeit with more recent data, but we also extend the discussion in several dimensions. For example, we look at multidimensional and conditional densities. In other words, we realize that output per worker is related to several other variables and we include some of those in our analysis. We also consider tests for more than just multimodality, including testing for differences between different groups, such as OECD and non-OECD.

1.6.2 Regression

The primary objective in the regression chapters is to attempt to estimate the worldwide production function. Although most research in economic growth has been concerned with growth regressions (Mankiw, Romer, and Weil, 1992), more recent research is concerned with determining the form of the worldwide production function (Duffy and Papageorgiou, 2000). This is another area where nonparametric econometrics may prove useful, as many economists are not comfortable with the standard Cobb–Douglas constant-returns-to-scale production function, which is common in the theoretical literature. There are several researchers who have considered more flexible functions, such as the constant-elasticity-of-substitution production function, but it is quite possible that the worldwide production function is more complicated.

When possible, our plan is to estimate production functions with limited assumptions. First, we generally estimate the models in levels and not in logs. Sun, Henderson, and Kumbhakar (2011) discuss the problems with estimating production functions in logs. Second, we generally will not estimate the models in per-worker terms. In other words, we will not divide output or capital by labor input. The reason for this is that this implies a constant returns production function and we are also interested in testing for this assumption. We will spend time comparing our results to several parametric specifications and between the nonparametric models of different chapters.

As this approach to estimating production functions is relatively new, this process (i.e., the end-of-chapter applications) can be thought of as research in action. When we started this book we were unsure where this approach was going to take us and hence there is a lot of trial and error. This "experiment" to applications may produce some unexpected results (in fact, one of the main results seems to be our Moby Dick), but our belief is that these applications will lead to a better understanding of the methods, which is our primary purpose.

1.7 Outline of the remainder of the book

The remainder of the book can be generally broken down into four sections: (1) density estimation and testing, (2) regression analysis, (3) handling discrete data, and (4) advanced nonparametric methods. The first two sections can be used to teach a semester course if the instructor also dives into the technical appendices. An applied course should probably include the first three sections with selected portions of Section 4 (as desired). It is possible to start a course focusing solely on regression with Section 2, but we recommend spending some time in the first section (say a week or two) to help with intuition.

Chapter 2 outlines univariate density estimation. It is relatively lengthy as it introduces terms, ideas, functions, estimators, and asymptotic properties necessary to develop the intuition of smoothing, which will appear throughout the remainder of the book. We also spend a lot of time describing the smoothing parameter (bandwidth), as it is crucial to nonparametric estimation in any form. We believe that this chapter will be very intuitive and will help with understanding the constructs of later chapters.

Univariate density estimation can be useful in applied settings, but most economic data require methods which can handle multivariate data. Chapter 3 outlines multivariate density estimation. Here you will learn how to estimate and plot multidimensional densities. A lengthy discussion will be devoted to the computational and theoretical concerns when switching from univariate to multivariate data in a nonparametric setting. We also suggest how these methods can be extended in order to perform nonparametric regression.

Chapter 4 introduces various forms of hypothesis tests within the density framework. The understanding of these tests on an intuitive level will lead naturally to testing in other nonparametric realms. For example, we first consider testing a nonparametric density versus a known parametric density. This approach will be useful when trying to compare parametric regression models versus their nonparametric counterparts. We also consider tests for the difference between unknown densities as well as tests for uncovering the number of modes in the underlying density.

Our first real taste of regression analysis is given in Chapter 5 (estimation). We focus on several different estimators of the conditional mean and discuss the role of kernel choice and methods for bandwidth selection. We further discuss how automated bandwidth selectors can tell us something about relevance and linearity in different estimation frameworks. Finally, we consider estimation of the derivative of the conditional mean, which can be used to assess partial effects as well as other interesting economic phenomena (e.g., returns-to-scale).

In Chapter 6, we outline methods for testing in nonparametric regression. We start with methods to test for correct parametric specification. These tests can be used to validate economic theory or discredit past research based on (perhaps) restrictive functional form assumptions. We then consider tests similar to those in the parametric literature, such as tests for omitted variables. Finally, we highlight some tests which are under-utilized, but may be useful.

We start our discussion of discrete data in Chapter 7 by analyzing their role in density estimation. We consider estimation of univariate densities (probability mass functions) with discrete data. We also explain their role in joint densities and conditional densities. We pay special attention to kernel choice here as we cannot use the same kernel functions as before. Bandwidth estimation is also discussed, and we show how automated methods can determine whether or not the discrete data are relevant for prediction. Finally, we detail the analogous testing procedures for univariate and multivariate density estimators with discrete data.

In Chapter 8 we turn to nonparametric regression in the presence of discrete covariates. Using the same estimators as in Chapter 5, we discuss how to incorporate discrete regressors. Implementation is relatively straightforward. We again discuss estimation, bandwidth selection, and testing in a manner similar to what was outlined in Chapters 5 and 6 so that the discussion goes smoothly.

In our advanced topics section (the name of which necessitates that they be taught at a slightly more rigorous level), we decided to discuss four different areas: semiparametric methods, instrumental variable estimation, panel data methods, and constrained regression. There is clearly one very important area of econometrics that we are