THOMAS M. STOKER

■

Lectures on Semiparametric Econometrics

CORE
LECTURE
SERIES

CORE FOUNDATION
LOUVAIN-LA-NEUVE
UNIVERSITE CATHOLIQUE DE LOUVAIN

Lectures on Semiparametric Econometrics

The "CORE Foundation" was set up a few years ago with the goal of stimulating new initiatives and research activities at CORE.
One of these initiatives is the creation of a
CORE LECTURE SERIES
Every year, a young internationally renowned scientist is invited to give a set of lectures in one of the research areas of CORE.
The texts of these lectures are distributed in this new series.

CORE LECTURE SERIES

Martin Grötschel
Postmen, Ground States of Spin Glasses, Via Optimization and Cycles in Binary Matroids
(1989)

Drew Fudenberg and David M. Kreps
Learning and Equilibrium in Games
(1990)

Thomas M. Stoker
Lectures on Semiparametric Econometrics
(1991)

CORE, 34, Voie du Roman Pays,
B-1348 Louvain-la-Neuve, Belgium

1992

ISSN-0771 3894

LECTURES ON SEMIPARAMETRIC ECONOMETRICS

THOMAS M. STOKER
Sloan School of Management
Massachusetts Institute of Technology, Cambridge, MA

I wish to begin by thanking the faculty and staff at CORE for inviting me to present the 1991 CORE Lecture Series, and for their gracious hospitality during my stay in Belgium. This lecture series represents a unique opportunity to present research work in depth, which in my case represents work on semiparametric econometrics that I have been involved in since 1985. The attentive CORE audience made the lectures progress quite smoothly, and I thank them for the numerous comments and suggestions that I received. I would like to specifically thank Angus Deaton, Wolfgang Härdle, Steve Marron, Whitney Newey, James Poterba and James Powell for incisive comments that improved the lectures. Finally, I owe special thanks to Dale Jorgenson, for extensive comments on the lectures as well as encouragement and support throughout my career.

CONTENTS

INTRODUCTION

This series of lectures will discuss the semiparametric approach to econometric modeling and estimation. The material will focus on three main themes: i) relative merits of semiparametric approaches, ii) precision measurement and other statistical features, and iii) the use of semiparametric methods for model assessment. Standard models of limited dependent variables (discrete choice, censored or truncated samples, and duration processes) will be discussed, in part to motivate the use of index model restrictions. Average derivative estimators of index model coefficients provide the main example for general statistical results and practical issues. Certain specific applications will be discussed, emphasizing interpretation of estimates as well as graphical analysis.

LECTURE 1
BASIC ISSUES

This lecture gives a broad motivation for semiparametric methods in econometrics. After definitions, we give some basic illustration of how semiparametric methods involve estimation of parameters and functions in an empirical model. We then point out how semiparametric methods display the attractive statistical features of parametric and nonparametric modeling approaches. A simple model of labor supply illustrates in more detail how semiparametric methods arise in econometric models of limited dependent variables. Some statistical background is covered, including rates of convergence and precision properties of nonparametric estimators. We close with an overview of the remaining lectures.

I. Parameters Versus Functions in Empirical Analysis

In broadest terms, semiparametric econometric methods are measurement approaches that maintain structure in an empirical model that is most useful for interpreting the results, but do not rely on specific assumptions about features that are of secondary interest or importance. In particular, semiparametric methods are designed to permit estimation of parameters and auxiliary functions simultaneously, without specific assumptions on the forms of the unknown functions. As such, semiparametric estimators are by design less sensitive to auxiliary assumptions than parametric estimators, and are capable of depicting the data structure in clearer and richer ways than restricted parametric models.

A. Some Definitions and Basic Ideas

We begin at once with some definitions to focus our discussion.

A *parametric model* is a fully specified statistical model, known up to a finite vector of values θ, or parameters. As above, this requires a full specification of the distribution of the observations as a function of θ (conditional on ancillary information). Also, in typical practical situations, we are primarily interested in the values of only a subset of the parameters; for instance the sub vector β of $\theta = (\beta, \sigma)$. The remaining sub vector σ represents auxiliary or nuisance parameters, which are of less practical interest, but must be accounted for in the estimation of β.

A *semiparametric model* is a statistical model characterized up to (β, s), where again β is a finite vector of parameters of interest, but the auxiliary parameter s can take any value in an infinite dimensional set $\mathcal{S}$. For instance, s may represent the density function of the disturbances of a model, and $\mathcal{S}$ may consist of all density functions obeying certain properties, like continuity, differentiability or moment restrictions. A consistent semiparametric estimator of β is an estimator that is consistent regardless of the value of $s \in \mathcal{S}$. To characterize the properties of a semiparametric estimator of β, one's derivations must account for the fact that the nuisance parameter s may take on any value in $\mathcal{S}$.

"Nonparametric" estimation has been used to refer to several different concepts in statistics. For our purposes, a *nonparametric model* refers to a statistical model that is known only to obey general regularity conditions. Nonparametric estimation refers to estimation of such a model, or a function s that is only known to lie in an infinite dimensional set $\mathcal{S}$. In particular, nonparametric estimation methods can be applied to measure density functions and regression functions, and methods that are consistent are those with global approximating properties in suitably large data sets. Examples of nonparametric estimators include local average methods, such as kernel estimators, and truncated series expansion, such as polynomial or Fourier expansions.

The parametric-nonparametric distinction loosely characterizes the different styles of empirical analysis by econometricians and statisticians over the last few decades. Statistical analysis of data is, in crystalline form, a purely descriptive activity. While parameters may be measured as part of a method of parsimonious description, no systematic behavioral regularities are intentionally ascribed to the process generating the data. The aim of such an investigation is a cogent description of data patterns, and the value of such results rests on the notion that new data sets will exhibit the patterns previously observed and described. The nonparametric approach epitomizes the view "Why assume anything if you don't have to?" of Lehman (1975), or of methods of exploratory data analysis, such as in Tukey (1977) and Mosteller and Tukey (1977).

Econometric modeling makes use of economic theory and reasoning in two ways. First, for studying a particular response, economic reasoning suggests certain kinds of variables to be accounted for in the empirical analysis: e.g. prices and incomes are relevant for a study of expenditures on different commodity categories. Second, and more fundamental, is that econometric modeling regards the process generating the data as reflective of predictable behavioral regularities, namely those consistent with systematic economic responses of individuals, firms or other decision makers. As such, a complete econometric model specifies the economic behavioral process, as well as the bridge between that process and the observed data. The main goal is the interpretation of the economic process (or the basic structure), and the value of the results rests on whether the behavioral structure that has been uncovered will adequately represent behavioral responses in new data sets or new economic environments.

This distinction is well known from many standard problems that have guided the evolution of econometric methods. For instance, the original econometric problem of measuring the parameters of a simultaneous equation system highlights the use of modeling structure combined with statistical description. A complete ac-

counting of the joint configuration of quantity and price data from a competitive market may be interesting, but is unlikely to capture the actual market process without sufficient structure to identify separate demand and supply influences. In this spirit, the principal aim of many econometric techniques is to focus measurement on specific, well-defined influences.

The econometric tradition has evolved to the point where basic decision processes are the major focus of a model, with the connection of these processes to the observed data (say through distributions of unobserved heterogeneity or of available information) of secondary interest. Adherence to these distinctions can be seen from behavioral models of macroeconomic variables (albeit based on the ridiculous fiction of a single "representative agent" who makes decisions) to the extensive microeconomic models of consumption, labor supply and retirement decisions. As such, the models are frequently quite complicated,[1] and until recently, have been implemented as fully parametric statistical models.

Semiparametric approaches bridge these two extremes of parametric and nonparametric methods. We now develop the ideas of semiparametric estimation, and indicate why they are valuable.

B. Flexible Treatment of Functional Aspects: Distributions and Relationships Among Variables

A simple example will help to add concreteness to these ideas. Suppose that our job is to analyze observations on wage income (earnings) y_i of various individuals (say $i = 1, \ldots, N$), and we have observations on nonwage income I_i and various individual and occupational characteristics A_i , such as training, job experience, age, etc., that we expect may be associated with differences in earned wage income. Suppose that there are k predictor variables, summarized as $x_i = (I_i, A_i)$.

Consider first a model of this data where the semiparametric approach can be easily stated, but is of little consequence. In particular, suppose that we decided to summarize the differences in wage incomes y_i associated with differences in x_i by fitting a simple linear model, namely

$$y_i = \alpha + x_i^T \beta + \varepsilon_i, \tag{1.1}$$

[1]See Cowing and McFadden (1984) for a clear depiction of the practical issues that arise with extensive econometric models.

where the components of β gauge the "effects" of changing the components of x_i; for instance β_1 , the coefficient of nonwage income I_i, is interpreted as how the presence of nonwage income affects earned wage income. The term ε_i represents wage income differences that are not associated with x_i.

A fully parametric version of this model is obtained by specifying the density $f(\varepsilon)$ of ε_i, say as a normal density with mean 0 and variance σ^2. Then (1.1) gives a normal linear regression model, and we could estimate β using the least squares coefficients $\widehat{\beta}$ of y on x. Moreover, this structure of the model ascribes various optimality properties to $\widehat{\beta}$ as a measure of β (unbiased with minimum variance, normal, consistent, efficient, etc.).

With β the parameters of main interest, we could also note that $\widehat{\beta}$ has many valuable properties that are not connected to the normality of ε. In particular, under the conditional mean independence assumption

$$E(\varepsilon \mid x) = E(y - \alpha - x^T \beta \mid x) = 0, \tag{1.2}$$

the estimator $\widehat{\beta}$ is unbiased, consistent and, asymptotically normal. Using $\widehat{\beta}$ to measure β based on (1.2) can be regarded as a semiparametric method, which implements the structural equations (1.1), (1.2) without specifying the density $f(\varepsilon)$ of the disturbance.

In this familiar linear model context, provided $\widehat{\beta}$ is an adequate estimator, there seems little reason for concerning ourselves with the parametric-semiparametric distinction, for practical purposes. However, other partial distributional restrictions can motivate different methods of measuring β. For instance, under the conditional median independence assumption

$$\text{Median}(\varepsilon \mid x) = \text{Median}(y - \alpha - x^T \beta \mid x) = 0, \tag{1.3}$$

the natural estimator of β is the least absolute deviations (LAD) estimator $\tilde{\beta}$, found via

$$(\tilde{\alpha}, \tilde{\beta}) = \arg\min \sum |y_i - \alpha - x_i^T \beta|. \tag{1.4}$$

This posture can be generalized to apply to any quantile other than the median. Corresponding to this generalization is the quantile regression (weighted LAD) method of measuring β.[2]

[2]This method is studied by Bassett and Koenker (1978), and has been advocated by Chamberlain (1991) as a parsimonious method of summarizing data interrelationsships.

Instead of focusing on the distribution of the disturbance ε, we can consider more general relationships among the variables. For instance, we could begin with the model

$$y_i = G(x_i^T\beta) + \varepsilon_i \tag{1.5}$$

including the centering restriction

$$E(\varepsilon \mid x) = E[y - G(x^T\beta) \mid x] = 0. \tag{1.6}$$

These restrictions comprise an index model, summarized as $E(y \mid x) = G(x^T\beta)$. This form retains the coefficients β as relative measures of the contributions of predictor variables, but permits a nonlinear relationship between y and $x^T\beta$. Semiparametric methods, such as the average derivative method, estimate the coefficients β and the univariate function G nonparametrically. An estimate of G gives a compact depiction of the nonlinearity evidenced in the observed data.

Finally, we could effectively abandon all structure on relationships among the variables, by using the model

$$y_i = m(x_i) + \varepsilon_i, \tag{1.7}$$

including the centering restriction

$$E(\varepsilon \mid x) = E[y - m(x) \mid x] = 0, \tag{1.8}$$

where the k-dimensional function $m(\cdot)$ only obeys smoothness and regularity conditions. This constitutes a fully nonparametric model, with an appropriate estimator $\widehat{m}(x)$ exhibiting global approximation properties in large data sets. Such estimators could be computed via kernel regression, nearest neighbor methods, truncated polynomial or Fourier series expansions, to name a few. This approach could be taken further, by fitting the joint density of (y, x) using a nonparametric estimator.

C. Nonparametric Estimation in Econometrics

While nonparametric methods have only recently been applied in econometric analysis, the need for flexible treatment of functions has been recognized for many

years. The early calls for flexible econometric methods came in demand and production analysis, when it was noted that various empirical models would dictate effects that ought to be measured. For instance, the popular linear expenditure system of Stone and Geary implies extremely restrictive cross-price and income effects, as does the Cobb–Douglas or CES production models. This state of affairs led to the use of "flexible functional forms," which were parametric models that allowed for arbitrary values of first and second order derivatives to be measured at a point.[3] While this definition of pointwise accuracy was not upheld in the estimation methods used to fit these models, they represented a first step toward eliminating subtle or not-so-subtle impositions of parametric assumptions on the results of empirical analysis.

Recent applications of nonparametric methods in econometrics have been focused on low dimensional problems, which are most amenable to modern graphical methods for analysis. These applications demonstrate how nonparametric methods can give a richer and/or clearer depiction of the data than traditional parametric methods. As such, these methods are a welcome addition to any econometrician's tool box.

One example of this type of work is given in Schmitz(1989), Härdle(1991) and Hildenbrand and Hildenbrand (1986), who document the changing bimodal character of the UK income distribution over time. This structure would have been masked by typical Pareto or lognormal income distribution models, and further analysis has led Schmitz(1989) to uncover the modal structure as a mixture of the income distribution of pensioners and of all other individuals. Other kinds of examples include the demand studies of Bierens and Pott-Butler (1991), Härdle and Jerison (1989) and Hausman and Newey (1990), the latter using a nonparametric estimator of demand to measure welfare costs of gasoline taxation; and the study by Gallant, Hsieh and Tauchen (1991) of financial series.

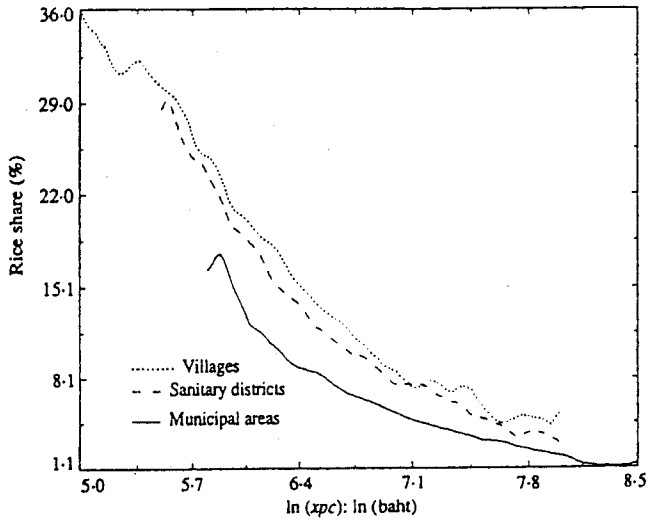
For illustration, we consider some of the analysis of Deaton (1989) on the impact of changing rice prices in Thailand. This study concerned a very simple question; whether raising rice prices would on average be beneficial or costly to families, and whether this distinction varied with whether the family was poor or rich, or whether the family lived in an urban or rural location. As such, a simple analysis is based on studying how many people were net buyers of rice and how

[3] Lau (1986) surveys this work.

many were net sellers, at different levels of income. With this in mind, Deaton constructs a simple net benefit measure as follows. If q denotes the amount of rice purchased, and y the amount of rice sold, then the amount of money required to compensate a family for a small price change dp is $dB = (q - y)dp$. This amount, as a proportion of income X, is

$$\begin{aligned} dB/X &= (pq/X - py/X)d\ln p \\ &= (e - s)d\ln p \end{aligned}$$

where $e = pq/X$ is the expenditure share of rice purchases and $s = py/X$ is the fraction (or multiple) of rice sales relative to total expenditure. The nonparametric analysis we discuss is an analysis of expenditure share e, and relative net purchases $e-s$, as related to $x \equiv \ln(xpc)$, or log total expenditure per family member. Deaton proxies welfare "benefits" by $w = -dB/x/d\ln p = -(e - s)$. We reproduce several of Deaton's diagrams, reporting kernel estimators computed with these data.[4]



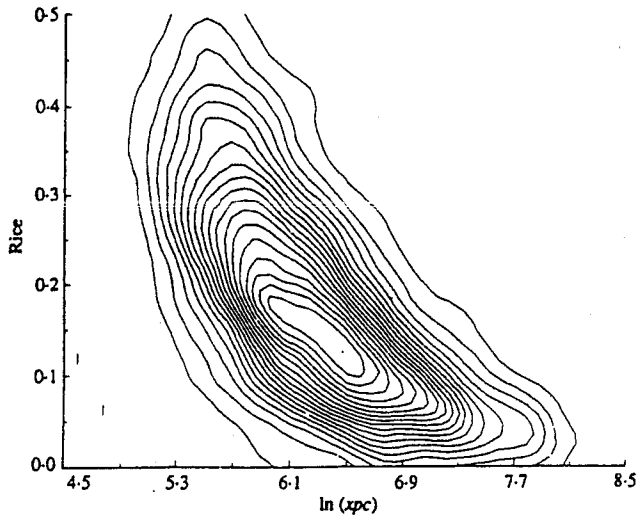
Rice share regressions. Bandwidths are 0·10, 0·15 and 0·10.

Figure 1.1a (From Deaton (1989))

In Figures 1.1 a,b we display pictures for a typical "Engel curve" or demand analysis, on e and $x \equiv \ln(xpc)$. The first diagram gives nonparametric regression

[4] I am grateful to A. Deaton and the publisher for permission to reprint these diagrams.

estimators of $E(e \mid x)$, with their downward slope indicative of Engel's law that rice shares decline with total expenditure. These curves are nonparametric versions of what one would fit in a regression analysis of the rice Engel curve. The second diagram gives a representative estimator of the joint density of e with x, and reveals a typical pattern of individual expenditure shares, namely that they are highly spread around the regression line. Figure 1.2 contains a simple nonparametric regression depicting some characteristics of rice farming, namely that poor families are more likely to be farmers, and given that a family is a farmer, richer families are more likely to sell their crop.



Villages: Bivariate density contours; rice share and ln (xpc). Bandwidth = 0·50.

Figure 1.1b (From Deaton (1989))

Figures 1.3 a,b,c contains diagrams of the joint distribution of relative net purchases $e - s$ with log expenditure x, as well as the regression of benefits $-(e - s)$ on x. Taking the regression first, we see that the mean benefits are positive for all income levels, but are relatively largest for families in the middle of the income distribution. This depiction could lead one to think that higher prices are beneficial to families at all income levels. However, the distributions of net purchases (the regression is on the negative side of these pictures) show that the limited regression

analysis misses the extreme amount of variation in the net purchases, and hence benefits, for families of all incomes. Aside from the fact that higher prices do not suggest any poor-to-rich income redistribution (or vice versa), at all income levels there are substantial numbers of net gainers and net losers. In particular, to make sense out of a "higher prices are good" policy prescription, one would require evidence of some kind of redistribution process (say through increased income of farm workers), that would spread the mean benefits across the wide range of families seen here. Again, this distributional richness would be missed by any standard, sensible regression analysis of the benefit measures.

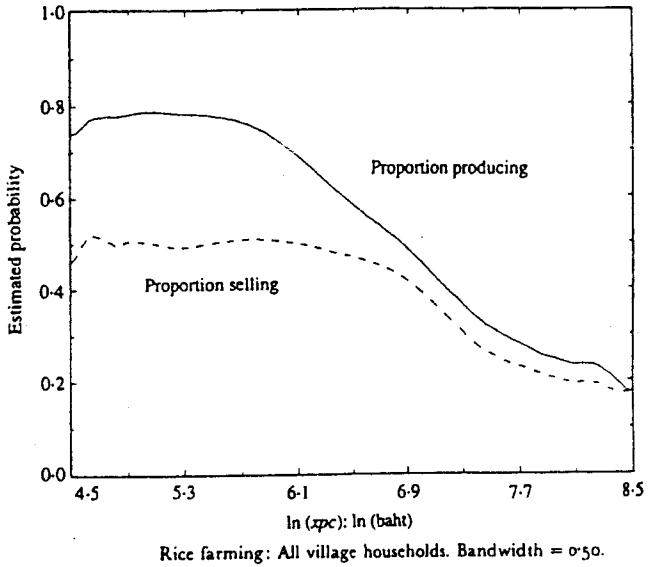


Figure 1.2 (From Deaton (1989))

This brief discussion certainly does not substitute for a detailed reading of Deaton's interesting study. However, it does serve our purposes of illustrating the potential role for enhancing econometric analysis by using nonparametric statistical methods.

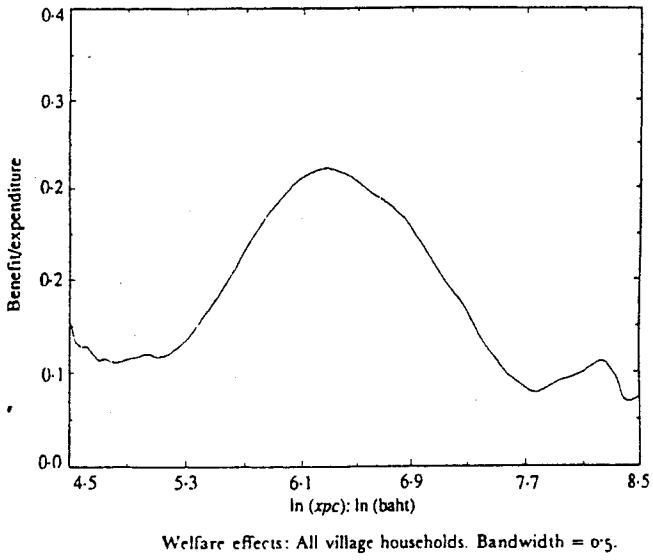


Figure 1.3a (From Deaton (1989))

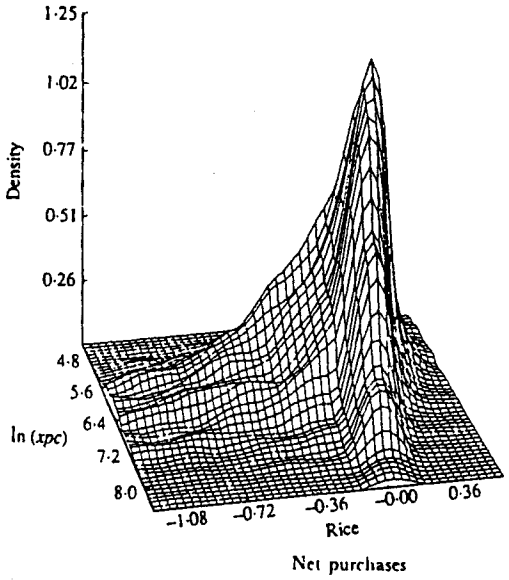
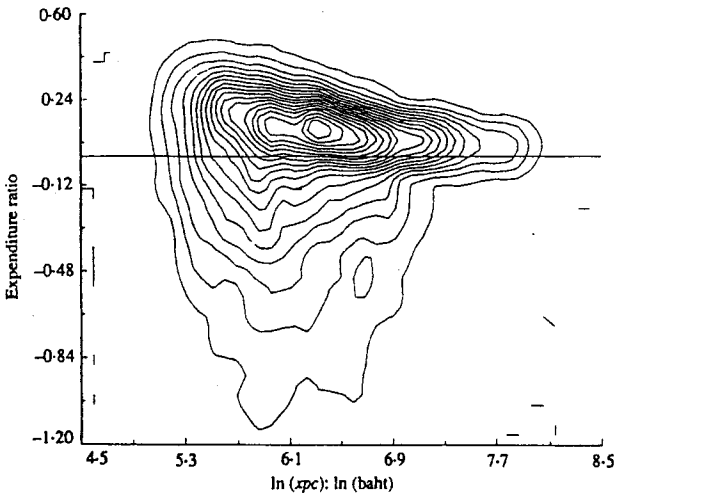


Figure 1.3b (From Deaton (1989))



Net purchases: 5,836 village households; 3,001 make net purchases; 2,677 make net sales.

Figure 1.3c (From Deaton (1989))

II. Why Use Semiparametric Methods?

A. Flexibility, Parsimony, and Precision

Nonparametric approaches provide the greatest range of flexibility in empirical analysis, representing "structure free" methods. The cost of this flexibility is twofold. First, in a problem with many predictors, a nonparametric estimator may be very difficult to interpret, or be effectively useless for parsimonious data summary. For instance, with regard to our discussion of analyzing wage income, if there are $k = 10$ predictors, a nonparametric regression $\widehat{m}(x)$ is a 10-dimensional function, with potentially many bumps and other nonmonotonic features. It is notoriously difficult to display such a function, even with modern graphical methods. The second cost is in statistical precision. Even with large data sets, the pointwise values of a nonparametric regression $\widehat{m}(x)$ are likely to be based on only a few data points, and therefore are not very precise measurements. This issue is formalized in statistical theory by slow rates of convergence, a problem which is exacerbated by many predictors; the so-called "curse of dimensionality" discussed later.

On the other hand, a parametric approach is potentially quite restrictive, but can display great advantages in terms of interpretability and precision. The design of a parametric model focuses attention on a few values, which summarize the data interrelationships. For instance, the great practical value of a linear model is its summary of the impact of each predictor by one number, its coefficient. More specifically, econometric analysis generally seeks to uncover basic economic structure, such as a decision mechanism, and a parametric model provides the easiest setting for separating the basic structure from incidental structure, such as the distribution of unobserved heterogeneous features. In addition, parametric estimators inherit the data usage properties of sample averages, with precision increasing with the square root of sample size ($\sqrt{N}$convergence). This feature of the precision of estimation cannot be matched by a nonparametric approach with many predictor variables.

Why use semiparametric methods? Because such methods exhibit the best features of the other approaches. Parameters are used for interpretation and simple functions are used for further data summary. With an index model $E(y \mid x) = G(x^T\beta)$, the relative impacts of predictor variables are summarized through β, and the connection of y to the index $x^T\beta$ is given by the one-dimensional function G. In a more explicit econometric model, such as the labor supply model given below, the decision structure can be modeled for interpretation, with the incidental features treated flexibly.

The main statistical question faced by semiparametric methods is whether the unrestricted functions in the basic model can be treated nonparametrically, in a way that avoids the nonparametric curse of dimensionality. The primary theoretical advance of the literature to date is that this question can be typically be answered positively. In particular, the estimators of parameters in a semiparametric model can and do converge at $\sqrt{N}$ rates of convergence, or utilize data with comparable efficiency to estimators based on fully parametric models.[5] For index models, we study average derivative estimators of the coefficients, that display these characteristics.

The functions estimated as part of a semiparametric approach must obey the precision features of nonparametric estimators, but they are naturally of reduced

[5]Such results have been shown for many specific semiparametric methods and models, such as those discussed in the next lecture. Various kinds of unifying theory are given in Andrews (1989), Newey (1991) and Goldstein and Messer (1990), among others.

dimension. Again, in an index model $E(y \mid x) = G(x^T\beta)$, the function $G(\cdot)$ is one-dimensional, and an estimate can be fully depicted on a graph.[6] In a low (one- or two-) dimensional problem, the data requirements for characterizing unknown functions are quite moderate. The Deaton pictures are effective because you can *see* the answers.

Enhanced methods of data analysis are only part of the advantages that semiparametric methods bring to empirical modeling. Since any parametric model can be generalized to a semiparametric model, the latter can be used to assess the adequacy of the parametric model. For instance, suppose that a key parameter was estimated using a parametric model, and a more general semiparametric model. If the estimated values were the same, then one could conclude that the parameter estimate was not heavily influenced by the restrictions of the parametric model. If the estimated values differed substantially, doubts would be cast on the adequacy of the restrictions of the parametric model. In either case, one can learn from graphing the estimates of unknown functions from the semiparametric model. For instance, these graphs could be compared to the specified functions of the parametric model.

For this purpose, it is important to note that semiparametric models can be designed to focus on specific restrictions of a parametric model. With reference to the next lecture (and the example below), suppose that a parametric model was built around a latent index variable and unobserved factors that were normally distributed. A semiparametric model could retain the latent index variable, but allow an unrestricted distribution of the unobserved factors. Differences in estimates from these two models could only arise from undue influence of the restriction that the factors were normally distributed. A graph of the density of the unobserved factors (estimated nonparametrically) would show how the normality restriction fails. This outlines but one illustration of how semiparametric methods can be used to judge empirical results obtained from parametric models, as well as diagnose the reasons for any differences in results.[7]

[6] While we focus on index models, other methods have been developed that permit reduced dimension depictions of data (see Stone (1986) for a general discussion). Most notable is the use of additive models, such as $E(y|x) = \sum \zeta_j(x_j)$, where the impact of component x_j of x summarized as a nonlinear one-dimensional function ζ_j. See Hastie and Tibshirani (1990) for a discussion of such approaches.

[7] Semiparametric methods have begun to be used as diagnostic tools in certain economic applications. For instance, the parameter estimates for female labor supply obtained by Mroz (1987) were checked by Newey, Powell and Walker (1990) using various semiparametric estimators. Because checking the values of parameters represents only a very minimal use of semiparametric methods (as discussed in Lecture 5), we discuss more comprehensive methods of specification testing in Lecture 4.

To focus our discussion, we now turn to the role of econometric modeling in the design of semiparametric models. Return to our scenario of studying wage income y as a function of nonwage income and other attributes $x = (I, A)$. Suppose further that a linear model is clearly inadequate, because our data contains observations on many people who don't work, with $y_i = 0$ for such people. For this, we draw on a modeling framework that captures the decision to work, as well as how much to work. In so doing we illustrate several semiparametric methods, as well as motivate the estimation of an index model $E(y \mid x) = G(x^T\beta)$.

B. A Simple Model of Labor Supply

We construct a simple static model of labor supply as follows. Suppose that a person receives nonwage income I, and decides how much time to allocate to labor L and leisure $1 - L$. The decision is based on maximizing preferences between expenditures on goods G and leisure $1 - L$, subject to the constraint that income covers expenditures $G \leq WL + I$, where W is the wage rate available to the individual. Suppose further that preferences are represented by a utility function $U(G, 1 - L)$ of the Cobb–Douglas form, namely $U(G, 1 - L) = \phi \ln G + (1 - \phi)\ln(1 - L)$, where ϕ is a parameter gauging tastes between goods and leisure, and $0 < \phi < 1$.

Since leisure is valued, all income will be spent, so that $G = WL + I$. Substituting $G = WL + I$ into U, this simple setup predicts that labor hours are determined as the result of the solving the problem

$$\begin{aligned} \text{Maximize} \quad & U(L) = \phi \ln(WL + I) + (1 - \phi)\ln(1 - L) \\ \text{subject to} \quad & L \geq 0. \end{aligned}$$

Examining U, we have that

$$\begin{aligned} \partial U/\partial L &= \phi W/(WL + I) - (1 - \phi)/(1 - L) \\ \partial^2 U/\partial L^2 &= -\phi W^2/(WL + I)^2 - (1 - \phi)/(1 - L)^2 < 0. \end{aligned}$$

Consequently, $\partial U/\partial L$ declines as L increases. Therefore, the individual will not work:

$$L = 0$$

if $\partial U/\partial L < 0$ when $L = 0$. Otherwise, $\partial U/\partial L = 0$ determines L as in

$$WL = \phi W - (1-\phi)I.$$

These cases are combined as the basic model for wage income WL;

$$\begin{aligned} WL &= \max\{0, \phi W - (1-\phi)I\} \\ &= [\phi W - (1-\phi)I]\ 1[\phi W - (1-\phi)I \geq 0] \end{aligned}$$

where $1[\ ]$ is the indicator function, which is 1 or 0 depending on whether the condition in the brackets is true or false.

Wage rates are not observed, but rather just some indicators A of wage differences. We complete the model by assuming that wages are randomly distributed, with a common distribution around the mean $\mu + A^T\gamma$. In other words, W is summarized by a linear equation

$$W = \mu + A^T\gamma + u$$

where u is distributed independently of A and I. The model for observed wage income (earnings) is then

$$\begin{aligned} y \equiv WL &= [\phi W - (1-\phi)I]1[\phi W - (1-\phi)I \geq 0] \\ &= [\phi\mu - (1-\phi)I + A^T\gamma\phi + \phi u] \\ &\qquad 1[\phi\mu - (1-\phi)I + A^T\gamma\phi + \phi u \geq 0] \\ &= [a + x^T\beta + \varepsilon]1[\alpha + x^T\beta + \varepsilon], \end{aligned} \tag{1.9}$$

where we have made the assignments

$$\alpha = \phi\mu; \quad x = (I, A)^T; \quad \beta = [(\phi - 1), \gamma^T\phi]^T; \quad \varepsilon = \phi u.$$

This model is a censored regression model, frequently referred to as a censored Tobit model. It is clear that with estimates of α and β, one could solve for estimates of ϕ, μ, γ. In particular, if β_1 is the first component of β and β_2 the vector of remaining components, then $\beta_1 = \phi - 1$ and $\beta_2 = \phi\gamma$, so that β_2 is proportional to the coefficients in the wage equation. These equations also facilitate inference on the values of the basic parameters; for instance $\beta_2 = 0$ coincides with $\gamma = 0$.

The standard parametric approach to this problem requires a specification of the distribution of the unobserved wage distribution, or of the heterogeneity

term ε. If the variance of ε is σ^2 , denote the cumulative distribution function (c.d.f.) of ε/σ as $F(\cdot)$, and the density as $f(\cdot)$. For a parametric application, this would be taken as a normal distribution, or some other specified formulation. The probability $G(y \mid x)$ that $y \leq y_0$, or the c.d.f. of y, is then the probability that $\varepsilon \leq y_0 - \alpha - x^T\beta$, or

$$\begin{aligned} G(y \mid x) &= F[-(\alpha + x^T\beta)/\sigma] \quad \text{for} \quad y_0 = 0 \\ &= F[(y_0 - \alpha - x^T\beta)/\sigma] \quad \text{for} \quad y_0 > 0 \end{aligned}$$

with the density of y for $y > 0$ given as $(1/\sigma)f[(y - \alpha - x^T\beta)/\sigma]$. The conditional density of y given $y > 0$ is

$$f_c(y, x, \sigma, \alpha, \beta) = (1/\sigma)f[(y - \alpha - x^T\beta)/\sigma]/\{1 - F[-(\alpha + x^T\beta)/\sigma]\}.$$

If $d = 1[y > 0]$ indicates positive wages, then the log-likelihood of an observation (y, x) is

$$\begin{aligned} \ln \mathcal{L}(y, x, \alpha, \beta, \sigma) &= (1-d)\ln F[-(\alpha + x^T\beta)/\sigma] \\ &\qquad + d\ln\{(1/\sigma)f[(y - \alpha - x^T\beta)/\sigma]\} \\ &= (1-d)\ln F[-(\alpha + x^T\beta)/\sigma] \\ &\qquad + d\ln\{1 - F[-(\alpha + x^T\beta)/\sigma]\} \\ &\qquad + d\ln\{f_c[y, x, \sigma, \alpha, \beta]\} \\ &= \ln \mathcal{L}_d(d, x, \alpha/\sigma, \beta/\sigma) + d\ln \mathcal{L}(y, x, \alpha, \beta, \sigma) \end{aligned} \tag{1.10}$$

where $\mathcal{L}_d$ is the likelihood based on d, x (whether one works) and $\mathcal{L}_c$ is the conditional likelihood given $y > 0$ (that one does work). Standard maximum likelihood estimation measures α, β and σ as those values that maximize $\sum \ln \mathcal{L}(y_i, x_i, \alpha, \beta, \sigma)$.

A nonlinear regression analysis would be based on the regression function

$$\begin{aligned} E(y \mid x) &= \text{Prob}\ (y = 0 \mid x)E(y \mid y = 0, x) \\ &\qquad + \text{Prob}\ (y > 0 \mid x)E(y \mid y > 0, x) \\ &= \{1 - F[-(\alpha + x\beta)/\sigma]\} \\ &\qquad \{\alpha + x^T\beta + \sigma E[\varepsilon/\sigma \mid \varepsilon > -(\alpha + x^T\beta)]\} \end{aligned} \tag{1.11}$$

where the final expectation could be solved for from $F(\cdot)$.[8] The parameters α , β and σ could be estimated by fitting this function to the observed y values by least squares.

[8] If F is the normal c.d.f., this function is the inverse "Mills ratio" f/F evaluated at $(\alpha + x^T\beta)/\sigma$.

It is evidently clear that the choice of whether to work or not induces nonlinearity in the model for wages (0 is the value for several observations), and that the measurement of the parameters depends intrinsically on the assumed form of the heterogeneity distribution $F(\cdot)$. It also depends on the form of the wage equation, as well as on the form of preferences. As before, we could react in the extreme by nonparametrically estimating the regression; but again this may result in an uninterpretable smear of the various effects.

Semiparametric approaches provide a useful compromise to a full empirical specification here, because the heterogeneity distribution $F(\cdot)$ plays such a critical role in the model. "Model specific" semiparametric approaches are based on the entire model except for the specification of the distribution $F(\cdot)$ of the heterogeneity, or on (1.9). These include semiparametric maximum likelihood, where the likelihood function (1.10) is maximized with regard to the parameters α, β and σ, as well as the distribution F (which lies in some regular class).[9] Other methods that apply directly to (1.9), such as least absolute deviations, will be discussed in the next lecture. Suitable parameter estimates from any of these approaches retain the interpretability of the answers; namely α and β are interpretable as they would be if they were estimated using a correct parametric specification.

A weaker form of semiparametric approach arises from noting that the decision to work, and how much to work, is based on the latent variable $\alpha+x^T\beta+\varepsilon$, so that systematic differences in wage income y across individuals are based on the value of the index $x^T\beta$. In particular, The regression (1.11) gives primary motivation for estimation based on the index restriction

$$E(y \mid x) = G(x^T\beta) \tag{1.12}$$

discussed earlier. Semiparametric approaches based on estimation of the univariate function G and the coefficients β (up to scale) are discussed in the remaining lectures.

Consider the interpretability of estimates of G and β. First off, a strong interpretation for β arises from the full implications of the original model — if κ is a scale factor, then $\beta_1 = \kappa(1-\phi)$ and $\beta_2 = \kappa\phi\gamma$, so that the relative values of the components of β retain their connection to the parameters of the underlying

[9] Approaches based on semiparametric implementation of maximum likelihood include Buckley and James (1979), Heckman and Singer (1984), Gallant and Nychka (1987), Severini and Wong (1987), Ritov (1990), Klein and Spady (1990), and Ai (1991a).

model. A weaker interpretation of β is based on the restrictions of the index model (1.12). In particular, suppose that nonwage income I were altered slightly to $I + \Delta I$, then the mean wage income y would be altered by $\Delta E(y \mid x) = \partial E(y \mid x)/\partial I \Delta I = dG/d(x^T\beta)\beta_1 \Delta I = G'\beta_1 \Delta I$. If the components of A were likewise varied, then the change in conditional mean of y would be $G'\beta_2\Delta A$, where β_2 represents the remaining components of β. In other words, if the components of x are continuously varied, we see that β is proportional to the point wise derivatives of mean wage income $\partial E(y \mid x)/\partial x \equiv m'(x) = G'\beta$, or the (local) effects of x on y, although the proportionality constant G' varies with the value of x.

This is just a further verification of how the index model formulation retains the advantages of linear modeling in terms of parsimony, namely the contributions of variables are gauged by coefficient values. We note that what is definitive in the general form (1.11) are the relative values of the coefficients, not their levels, since any common scaling of all coefficients could be absorbed into the function G without changing any empirical implications of the equation (1.12).

Since the overall scaling of β is not important, we could normalize β to be representative of the basic effects. In particular, we could replace β by the mean effect; namely

$$\delta = E(m') \quad (= E(G')\beta) \tag{1.13}$$

and redefine G (absorbing the scale) as

$$E(y \mid x) = G(x^T\delta). \tag{1.14}$$

This redefinition imposes the normalization $E(G') = 1$, or that the average impact of the changing the index $x^T\delta$ on the mean of y is 1. As such, δ can be regarded as measured in units comparable to coefficients of a linear model. In this form, the mean effects δ are called the "average derivatives" of y on x. With this scaling, it is clear that estimation of δ and implementation of the index model (1.12) can be regarded in two ways; as a semiparametric generalization of linear models, with contributions of variables assessed by coefficient values; or as a flexible method of estimating parameters for a model that obeys an index restriction such as (1.11).

III. Statistical Antecedents: Precision and Rates of Convergence

With this introduction, we now discuss the modeling and statistical issues with a bit more formality, to clarify the range of issues faced in semiparametric estimation. Since our discussion will focus exclusively on nonlinear statistical models, our discussion of statistical properties will focus on approximate distributional theory obtained from asymptotic methods, which regard the sample size N as large. Throughout the lecture series we will discuss the use of such asymptotic theory, as well as some potential drawbacks such approximations may have.

The fundamental issues faced in asymptotic statistical theory can be loosely categorized into those of identification and consistent estimation, and those of characterizing the precision of estimators. Identification refers to a property of the statistical model, wherein a parameter is identified if altering it changes the empirical implications of the model. Consistent estimation refers to whether a statistic can be found that will measure the true parameter value exactly with an infinite amount of data (or has all mass in its distribution concentrated exactly on the true parameter value). These concepts are logically distinct and important for semiparametric estimation; however we will treat them somewhat casually here, only pointing out how identification and consistent estimation is achieved for index model restrictions, as well as certain specific models. See Manski (1988a), among many others, for a general treatment. Given a consistent estimator, the issues of precision are addressed via rates of convergence and limiting distribution theory, which we also treat casually, but in more detail below.

The cornerstone of asymptotic statistical theory are the Laws of Large Numbers and the Central Limit Theorems, which establish consistency and $\sqrt{N}$ asymptotic normality of sample averages. For instance, suppose that y_i is a k component random vector distributed with mean $\mu_y = E(y)$ and covariance matrix $\Sigma_y = E[(y-\mu_y)(y-\mu_y)^T]$, and that $\overline{y} = N^{-1}\sum y_i$ is the vector of sample averages. The weak law of large numbers asserts that plim $\overline{y} = \mu_y$, or

$$\overline{y} = \mu_y + o_p(1).$$

The central limit theorem asserts that departures of $\overline{y}$ from μ_y , scaled up by $\sqrt{N}$, approach a normal distribution:

$$\sqrt{N}(\overline{y} - \mu_y) \to \mathcal{Y} \sim \mathcal{N}(0, \Sigma_y).$$

Thus, $N(\overline{y} - \mu_y)\Sigma_y^{-1}(\overline{y} - \mu_y) \to \mathcal{Z} \sim \chi^2(k)$, which, given an estimate of Σ_y, provides one way of making approximate inferences on the value of μ_y .

This latter feature motivates the familiar asymptotic methods based on Taylor series expansions. In particular, if $B(\mu_y)$ is a continuous function of μ_y , then $B(\overline{y})$ consistently estimates $B(\mu_y)$ by Slutsky's Theorem, and if $B(\mu_y)$ is differentiable at μ_y with $b \equiv B'(\mu_y) \neq 0$, its limiting distribution is dictated by the first-order terms of the Taylor expansion. In particular, the "delta method" asserts that

$$\sqrt{N}[B(\overline{y}) - B(\mu_y)] = \sqrt{N}b^T(\overline{y} - \mu_y) + o_p(1)$$

which implies that

$$\sqrt{N}[B(\overline{y}) - B(\mu_y)] \to \mathcal{B} \sim \mathcal{N}(0, b^T\Sigma_y b)$$

from the properties of $\overline{y}$.[10] This equivalence to a sample average is often called the "influence representation;" in this case

$$B(\overline{y}) - B(\mu_y) = N^{-1}\sum \zeta(y_i, \mu_y) + o_p(1/\sqrt{N})$$

where $\zeta(y_i, \mu_y) = b^T(y_i - \mu_y)$ is the "influence" of the i-th observation.[11]

The approximate distributional theory available for estimators from suitably regular parametric models derives from influence representations of the above form. For instance, suppose that the data is a random sample from a distribution with density $p(y, x, \theta)$, and that θ is estimated by maximum likelihood; namely

$$\widehat{\theta} = \arg\max \mathcal{L}(\theta^*)$$

where $\mathcal{L}(\theta^*) = N^{-1}\sum \ln p(y, x, \theta^*)$ is the log likelihood function. Provided the true value θ is the unique maximizer of $E[\ln p(y, x, \theta^*)]$, and there is sufficient uniformity to connect $\mathcal{L}(\theta^*)$ to $E[\ln p(y, x, \theta^*)]$ as $N \to \infty$, then $\widehat{\theta}$ can be seen to be a consistent estimator of θ. Moreover, if $\mathbf{1}(y_i, x, \theta) = \partial \ln p(y_i, x_i, \theta)/\partial\theta^*$ denotes the score, and $\mathbf{i} = E(-\partial^2 \ln p/\partial\theta\partial\theta^T) = E(\mathbf{1}\mathbf{1}^T)$ denotes the Fisher information matrix, then

$$\widehat{\theta} - \theta = N^{-1}\sum \zeta_{m\ell}(y_i, x_i, \theta) + o_p(1/\sqrt{N})$$

[10] The second-order and higher terms vanish since $N(\overline{y}-\mu_y)(\overline{y}-\mu_y)^T$ approaches a stable limiting distribution, and each component of $\sqrt{N}(\overline{y}-\mu_y)(\overline{y}-\mu_y)^T = o_p(1)$. Rao (1973) is one standard reference for such properties.

[11] See Huber (1981) for this terminology, as well as the connection of the influence terms to Frechet and Gâteaux derivatives.

where $\zeta_{m\ell}(y_i, x_i, \theta) = \mathrm{i}^{-1}\mathbf{1}(y_i, x_i, \theta)$; as consistent with the first-order terms of the Taylor expansion of $\partial\mathcal{L}/\partial\theta$ around θ (using $\partial\mathcal{L}(\widehat{\theta})/\partial\theta \equiv 0$). Therefore, $\sqrt{N}(\widehat{\theta}-\theta)$ has a limiting normal distribution with mean 0 and variance $\mathrm{var}(\zeta_{m\ell}) = E(\mathbf{1}\mathbf{1}^T)^{-1} = \mathrm{i}^{-1}$.

Virtually all estimators used in parametric econometric modeling have influence representations and are asymptotically normal at rate $\sqrt{N}$, as with maximum likelihood. Any suitably regular estimator based on maximization, such as nonlinear least squares, generalized method of moments or the like, is analyzed in this fashion.[12] Later we will discuss the influence representation of average derivative estimators, in our demonstration of their approximate distributional properties.

Nonparametric estimators of density and regression functions are well studied, and we will make no attempt to cover their properties in any sort of detail (see Silverman(1986) for a good introduction to density estimation and Härdle(1991) for a discussion of regression estimators). However, because nonparametric estimators are used for estimating the unknown functions in semiparametric models, some understanding of their statistical features is warranted. First, however, we need some terminology used for gauging the precision of such estimators, namely via their "rate of convergence."

The $\sqrt{N}$ term appropriate for deviations of sample averages captures the notion of how quickly accuracy is enhanced by increases in sample size N. More precisely, an estimator $\widehat{\theta}$ is "N^{r} consistent" for θ if

$$N^{\mathrm{r}}(\widehat{\theta} - \theta) = O_p(1).$$

It is clear that if this relation holds for a rate r, then it likewise holds for $\mathrm{r}^* > \mathrm{r}$. Therefore, the smallest value r of such rate values, or more precisely, the greatest lower bound, gauges how quickly the difference $\widehat{\theta}-\theta$ vanishes with sample size, and correspondingly, N^{r} is the "rate of convergence" of $\widehat{\theta}$ to θ. All of our estimators based on sample averages above had $r = 1/2$; with $\sqrt{N}(\widehat{\theta} - \theta)$ stochastically bounded (since it has a limiting normal distribution).

To add flavor to how the convergence rate is connected to the notion of how well data is processed by the estimator, consider the following silly example. In particular, suppose that for estimating $E(y) = \mu_y$, we (randomly) chose $M = N^{2r}$ observations, $r < 1/2$, and computed the sample average $\overline{y}_r$ from those observations (i.e. we throw away a fraction of the data, and increase the fraction as N increases). Central limit theory states that $\sqrt{M}(\overline{y}_r - \mu_y) = N^r(\overline{y}_r - \mu_y)$ has a limiting normal distribution, so that N^r is the rate of convergence of $\overline{y}_r$ to μ_y.

[12] Standard theoretical treatments include Huber (1981), Bickel and Doksum (1977) and Amemiya (1985).

This example is silly because we throw away good observations, with each having mean μ_y, as did the included observations. However, suppose that the observations tossed out were not so good; namely they contained biases that might justify their exclusion. In this case, the lower rate of convergence may be the best available, because including all N observations could induce systematic mismeasurement.

In very crude terms, this is the state of affairs that exists for nonparametric estimation of a density function or regression curve. In particular, to estimate a regression value $E(y \mid x) = m(x)$, points close to x are valuable, but points distant from x are not (unless, of course, a parametric model assumption tells you how to combine them). To illustrate this, consider a loose argument based on a kernel density estimator.

Consider the estimation of a density function $f(x)$, as illustrated in Figure 1.4. Fix attention on the point x_0, and consider measuring $f(x_0)$ by a weighted average of counts of x values close to x_0. To define "close," consider applying nonzero weights in a "window" of only those x values whose components differ from x_0 by at most h; or $x_{0j} - x_{ij} \leq h$, where x_{0j}, x_{ij} are the j-th components of x_0, x_i. Second, consider weights that depend only on the proximity of x_i to x_0 relative to h, or depend on the value of $(x_0 - x_i)/h$. If $\mathcal{K}(u)$ denotes a function with support $[-1, 1]^k$ such that $\int \mathcal{K}(u)du = 1$, then the above features are satisfied by the (Rosenblatt–Parzen) kernel density estimator defined as

$$\widehat{f}(x_0) = \frac{1}{Nh^k}\sum_{i=1}^{N} \mathcal{K}\left(\frac{x_0 - x_i}{h}\right),$$

which applies a weight of $h^{-k}\mathcal{K}[(x_0 - x_i)/h]$ to the count of an observation with $x = x_i$ in close proximity to x_0, and a weight of 0 to counts of distant observations. We consider the mean-squared error of $\widehat{f}(x_0)$, divided in typical form into squared bias and variance

$$E(\widehat{f}(x_0) - f(x_0))^2 = (\mathrm{Bias})^2 + \mathrm{Variance},$$

where

$$\text{Bias} = E[\widehat{f}(x_0)) - f(x_0)]$$
$$\text{Variance} = E[\widehat{f}(x_0) - E(\widehat{f}(x_0))]^2$$

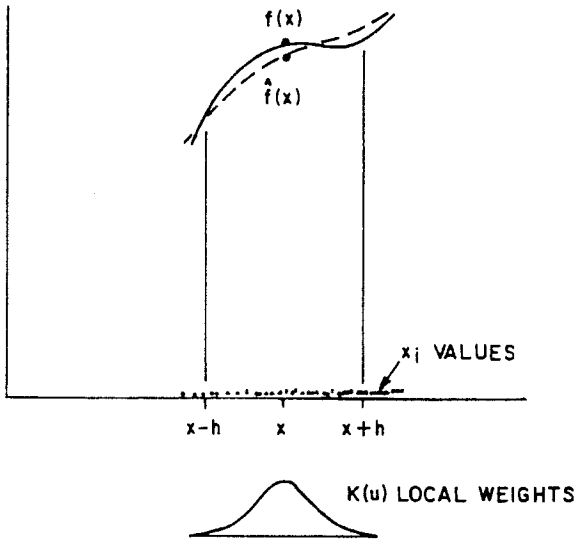


Figure 1.4
Local Averages and Kernel Density Estimation.

Taking the variance first, we may inquire as to how many nonzero terms are in the summation above, say M. The percentage of terms M/N varies proportionately with the level of density around x_0 , as well as the volume of the "window" $(2h)^k$. Therefore, the number of nonzero terms M is approximately $O(Nh^k)$, with $\widehat{f}(x_0)$ being an average based on this number of terms. As can be verified directly, the variance of $\widehat{f}(x_0)$ follows in line with this local average interpretation, or $\text{var}[\widehat{f}(x_0)] = O(M^{-1}) = O(Nh^k)^{-1}$. The variance is smaller with more data (larger N) or a larger window (larger h), and will vanish in the limit only if the number of nonzero terms grows, or $Nh^k \to \infty$.

The cost of a larger window is in mixing in distant counts, which adversely

creates bias in the estimated density. In particular, the bias is

$$\begin{aligned} E[\widehat{f}(x_0)] - f(x_0) &= h^{-k}E\{\mathcal{K}[(x_0 - x_i)/h]]\} - f(x_0) \\ &= \int h^{-k}\mathcal{K}[(x_0 - x_i)/h]f(x_i)dx - f(x_0) \\ &= \int \mathcal{K}(u)[f(x_0 - hu) - f(x_0)]du \end{aligned}$$

which will evidently be small when h is small, provided f is continuous at x_0 . The impact of further smoothness of f on bias is seen by a Taylor series expansion; namely if f is twice differentiable at x_0 with second derivatives uniformly Lipschitz, then

$$\begin{aligned} \int \mathcal{K}(u)[f(x_0 - hu) - f(x_0)]du = h\left(\frac{\partial f(x_0)}{\partial x}\right)\int u\mathcal{K}(u)du \\ + \frac{h^2}{2}\text{tr}\left[\left(\frac{\partial^2 f(x_0)}{\partial x \partial x^T}\right)\int uu^T\mathcal{K}(u)du\right] + O(h^3). \end{aligned}$$

Provided a condition is applied so that the remainder is negligible, bias is $O(h)$ in general, $O(h^2)$ if $\mathcal{K}$ is chosen such that $\int u\mathcal{K}(u)du = 0$ (say a density function with zero mean), or $O(h^4)$ if $\mathcal{K}$ symmetric and required to exhibit $\int uu\mathcal{K}(u)du = 0$ (a "higher order" kernel, with positive and negative local weighting).

In this way, the bias is determined by the smoothness of the function to be estimated, and the method of local averaging (order of the kernel). From a similar argument, if f is p-th order differentiable, the kernel can be chosen so that the bias is $O(h^p)$.[13]

As the sample size increases, the best pointwise mean square error is obtained by balancing bias squared $O(h^{2p})$ and variance $O(Nh^k)^{-1}$. This is achieved by choosing the window width h to decline as $O(N^{-1/2p+k})$. This results in a mean squared error of order $O(N^{-2r})$, where $r = p/(2p + k)$. Of course, this rate is less than $\sqrt{N}$ because we only want to use points close to x_0 , and to eliminate asymptotic bias, the criterion of "closeness" is sharpened as more data is used.

Identical issues arise if we were interested in estimating the derivatives of f at x_0, where we could form local differences of counts. If $\mathcal{K}$ is differentiable, and vanishes on the boundary of its support, this is accomplished with the kernel

[13]Using th exact analogy to the above formulæ, if the p-th order derivatives are also uniformly Lipschitz, then the bias can be shown to be $O(h^{p-1})$.

density derivative estimator

$$\hat{f}'(x_0) = \frac{1}{Nh^k} \sum_{i=1}^{N} h^{-1} K' \left(\frac{x_0 - x_i}{h} \right).$$

All the same arguments go through as above, and with the optimal bandwidth, the mean squared error is $O(N^{-2r})$, with $r = (p-1)/(2p+k)$.

These arguments are intended to motivate the idea that because of the need for increasingly fine local approximation, nonparametric estimators necessarily display a rate of convergence slower than averages based on the full sample, or parametric estimators. In fact, none of these issues are particular to kernel estimators, or estimators solely based on local averaging, as the best (optimal) rate of convergence for estimating the d-th order derivative of a density or regression function is N^r where $r = (p-d)/(2p+k)$. This is shown in the seminal work of Stone (1980,1982),[14] and applies equally well to any nonparametric method, including series expansions like polynomial or Fourier series. This rate depends on the order of the derivative to be estimated d, the smoothness of the function to be estimated p and the dimension k. The latter dependence, due to slow addition of data points in k-dimensional neighborhoods, is termed the "curse of dimensionality." As we have stressed, an important feature of understanding semiparametric estimators is how this "curse" is deflected.

IV. Overview of the Lecture Series

The purpose of the lecture series is to illustrate how the semiparametric approach can be used in practice, as well as cover many of the issues raised here. Lecture 2 will survey many standard modeling paradigms for limited dependent variables, to show the wide range of application areas for index models and to survey various "model specific" semiparametric methods. Lecture 3 covers average derivative estimation of index models, with an empirical example and discussion of the appropriate statistical theory. Lecture 4 shows how the specification of a model can be assessed through statistical tests in another empirical application, as well as covering some recent theory that clarifies how precision theory for semiparametric

[14] Härdle (1991) discusses this work and related work, including the modifications for Lipschitz conditions such as discussed earlier.

methods works in general. Lecture 5 concludes with broad suggestions for the future development of the field, as well as special attention to some small sample problems.

A short lecture series cannot hope to cover the lion's share of work in this rapidly evolving field.[15] As such, I will limit my remarks on statistical issues to those for independent random samples, where individual observations are not systematically connected or otherwise dependent. The methods discussed are applicable mainly to survey style data sets, and omit the issues of time series or other types of dependent data problems (although many of the ideas extend quite naturally to such data[16]). Second, I will be quite selective about topic coverage, with no intention of giving comprehensive treatment to all of the questions under current study. The most notable omission is a systematic treatment of semiparametric efficiency theory. This theory systematizes how semiparametric estimators can capture information in a way comparable to parametric estimators, although a clear coverage is beyond the scope of the lecture series.[17] Finally, we do not address issues of Bayesian or other non-classical approaches to statistical analysis.

[15] Fairly comprehensive surveys include Robinson (1988a) and Delgado and Robinson (1991).

[16] See Robinson (1991) for several such extensions.

[17] A good survey is provided by Newey (1989).

LECTURE 2

INDEX MODELS AND VARIOUS SEMIPARAMETRIC APPROACHES

In this lecture we will discuss various standard types of models that give rise to the index model structure

$$E(y \mid x) = G(x^T \beta). \tag{2.1}$$

Our focus will be on how latent index variables are used in modeling limited responses, such as those that are discrete, bounded or transformed. This gives one context for index models in econometric work, as well as helping readers unfamiliar with limited response models; in part, our coverage is a response to the fact that most published work on semiparametric methods just begins by spelling out one of these models formally. Also, in our discussion we will make reference to certain structural and distribution functions; say $\phi(x)$ and $F(\varepsilon)$, discussing specific properties of these functions where relevant. A parametric approach to any of the frameworks discussed below requires that all such unknown functions are stated explicitly, say up to a set of unknown parameters to be measured. Semiparametric methods will be applicable when some of these functions are left unspecified, as made clear in the specific cases discussed below.

While our emphasis is on the index restriction (2.1) and its variations, it is possible to catalog partial distributional restrictions and associated estimators for each of the specific models here. For instance, in the first lecture, we discussed how conditional mean independence was associated with OLS coefficient estimators with linear models, and how conditional median independence was associated with LAD estimators. While we review many of the specific procedures that could be systematized in this fashion, we do not catalog methods as such.[1] Moreover, as before, we will only make occasional reference to semiparametric identification issues; in part because of our focus on estimation of the index regression (2.1), where such issues are typically straightforward.

[1] For instance, see Manski (1988b) for this kind of comprehensive treatment of estimation methods for binary response models.

I. Various Limited Dependent Variable Models

The standard normal linear regression model

$$y = \alpha + x^T\beta + \varepsilon \tag{2.2}$$

takes $\alpha + x^T\beta$ as the typical value of y, with departures reflected by ε. When the distribution of ε is in a general sense "centered" or "symmetric", it is natural to measure α and β by minimizing the distance between y and the typical value $\alpha + x^T\beta$ over all data points, with the precise distribution of ε a secondary feature. Limited dependent variables arise when the response y is subject to limitations — it may be bounded, or take only values 0 and 1, etc. In this case, the stochastic features of the model (ε above) take on a central role for the "typical value" of y given x. Consequently, for such models, semiparametric methods have a natural appeal, as ways of measuring parameters that are not sensitive to the stochastic features. For this reason, we now familiarize ourselves with these kinds of models.

A. Binary Response Models

Here the object is to model a response y that takes on the value 1 or 0, and we assume that the actual value is the response to a predictable choice. If $\tilde{U}_1$ and $\tilde{U}_0$ denote the utility corresponding to each option, then 1 is chosen if the net benefits $B_{10} = \tilde{U}_1 - \tilde{U}_0 > 0$ and 0 is chosen otherwise.[2] If utility is modeled as $\tilde{U}_j = U_j(x) + u_j$, then net benefits are $B_{10} = \tilde{U}_1 - \tilde{U}_0 = \phi(x) - \varepsilon$, where $\phi(x) = U_1(x) - U_0(x)$ and $\varepsilon = u_0 - u_1$. Consequently, the empirical model is defined from

$$y = 1[\varepsilon < \phi(x)]$$

and a specification of the c.d.f. F of ε (conditional on x). The likelihood is based on the conditional mean

$$\begin{aligned} E(y \mid x) &= \text{Prob}\{1 \text{ chosen} \mid x\} = \text{Prob}\ \{\varepsilon < \phi(x)\} \\ &= F[\phi(x)]. \end{aligned}$$

If the distribution of ε varies with x, this dependence would be represented by an additional x argument in F.

[2]Typically, assumptions are made to guarantee that the probability of ties is zero, or where $\tilde{U}_1 = \tilde{U}_0$.

The linear index model (2.1) arises when the net benefits are modeled as dependent on a linear index $\phi(x) = \psi(x^T\beta)$ (and F depends on the value of x only through $x^T\beta$); whence we have

$$E(y \mid x) = F[\psi(x^T\beta)] = G(x^T\beta).$$

Probit models take F as the normal c.d.f.; logit models are likewise included. An example of a binary response model of this kind is available from the labor supply example of the first lecture, namely by studying the discrete response that indicates whether an individual works or not. A semiparametric estimator that measures β up to scale would be applicable to problems of this kind, with any transformation ψ or distribution of heterogeneity $F(\varepsilon)$.

An alternative to estimating β using the index model restriction (2.1) is to use least absolute deviations, which was proposed as one of the first substantive semiparametric methods to be studied. Assume that the index is linear, as in

$$y = 1(x^T\beta + \varepsilon > 0) \quad \varepsilon \text{ has c.d.f } F(\varepsilon).$$

Manski (1975)'s maximum score estimator is based on the following idea: assume that ε has conditional median 0; Prob $\{\varepsilon > 0 \mid x\} = 1/2$, then

$$\text{Prob}\ \{y = 1 \mid x\} \ \lesseqgtr\ 1/2 \quad \text{if} \quad x^T\beta \lesseqgtr 0,$$

with Prob $\{y = 1 \mid x^T\beta > 0\} > 1/2$. Thus $y = 1$ is "most likely" when $x^T\beta > 0$, and conversely. The maximum score estimator implements this logic, namely for the i-th observation, score 1 if $y_i = 1$ and $x_i^T\widehat{\beta} > 0$ or $y_i = 0$ and $x_i^T\widehat{\beta} < 0$, and score 0 otherwise, and define the estimator as any $\widehat{\beta}$ that maximizes the total score. A later interpretation of this estimator shows it to be a least absolute deviations estimator, since

$$\text{Score}(\beta) = N - \sum |y_i - 1(x_i^T\beta > 0)|$$

so that any maximum score estimator minimizes the absolute deviations of y from $1(x^T\beta > 0)$. Also, we have

$$\text{Score}(\beta) = 1/2(N + \sum \text{sgn}(y_i - 1/2)\ \text{sgn}(x_i^T\beta))$$

so that any maximum score estimator maximizes the sample correlation between $\text{sgn}(y - 1/2)$ and $\text{sgn}(x^T\beta)$.

Notice, however that maximum score estimators are not unique; even with a normalization $|\beta| = 1$, β can be varied slightly without changing $\text{sgn}(x_i^T\beta)$ for a finite number of x's; which means that maximum score estimators from a finite sample fall in defined ranges. While this entails severe restrictions on using the empirical estimates in practice, Manski(1985) has shown strong consistency for these estimators when the basic model is identified (or that the ranges shrink to the true value with increases in sample size). Related issues arise from considering the identification of β in these models, which can be quite delicate. For instance, β will not be identified when all predictors are themselves discrete: see Chamberlain (1986a) and Manski(1988b).[3] Aside from this, the maximum score estimator does not exhibit $\sqrt{N}$ consistency for β up to scale; the rate of convergence is slower, as shown in Kim and Pollard (1990). Horowitz (1990) has proposed a smoothed method of maximum score estimation, with better convergence properties. Matzkin (1990) has recently proposed a variation of this idea that relaxes the linear index, fitting a difference $U_0(x) - U_0(x)$ to $\phi(x)$, where $U(x)$ is a "least concave" function. Finally, Manski (1987) proposes applying maximum score estimators to differences observed in binary panel data; this idea has much in common with the "maximum rank correlation" estimator of Han (1987) for index models.

The basic motivation for estimators based on median restrictions, or more general quantile restrictions, is based on how the median operator "commutes" with continuous functions. If $\gamma(x)$ is the conditional median of z and $w(\cdot)$ is a

[3]This problem is formally treated by showing that semiparametric information is zero in these cases. The problem with discrete regressors can easily be seen from a simple example. Suppose that there are two predictors D_1 and D_2 that are both discrete, each taking on the values 0 and 1. Suppose further that the true model is an index model, namely $E(y|D_1,D_2)=G(D_1+\beta D_2)$, where we suppose that G is strictly increasing. The question is whether β and G can be uniquely solved for from the distribution of y, D_1 and D_2, which could be measured precisely in a large sample. In particular, we could estimate the four values $P_{mn}=E(y|D_1=m,\ D_2=n)=\text{Prob}\ (y=1|\ D_1=m,\ D_2=n)$ for $m, n = 0, 1$. The model gives the restrictions $P_{00}=G(0)$, $P_{01}=G(\beta)$, $P_{10}=G(1)$ and $P_{11}=G(1+\beta)$. Therefore, if β is known, then the values of G can be solved for. Inverting these restrictions gives $0=G^{-1}(P_{00})$, $\beta=G^{-1}(P_{01})$, $1=G^{-1}(P_{10})$ and $1+\beta=G^{-1}(P_{11})$. Therefore, if G is known, the value of β is determined.
However, if β and G are unknown, the restrictions of the model are $0=G^{-1}(P_{00})$, $1=G^{-1}(P_{10})$ and $1=G^{-1}(P_{11})-G^{-1}(P_{01})$. These equations are not sufficient to determine $\beta=G^{-1}(P_{01})$ (nor $G^{-1}(P_{11})$). Consequently β and G are not identified separately with discrete regressors.
The presence of continuous regressors can permit identification of discrete variable coefficients. Suppose that the model is $E(y|D,x)=G(\beta D+x^T\delta)$ where D is discrete, x is continuous, and δ is normalized (say with one component set to 1). Then G and δ are generally identified in the sample with $D=0$ (where $E(y|x)=G(x^T\delta)$ can be estimated using index model methods), and $\beta=G^{-1}[\text{Prob}\ (y=1|d=1,x)]-x^T\delta$ identified from G, δ and the $D=1$ sample.

nondecreasing function that is continuous at $\gamma(x)$, then $w(\gamma(x))$ is the conditional median of $w(z)$. This feature is coupled with the property that medians are the solutions to problems of minimizing absolute deviations, as in

$$g(x) = \arg\min_{\Gamma} E[|z - \Gamma| \mid x]$$

and that the sample analog of such an objective function coincides with least absolute deviations in the observed sample. For instance, when $y = w(x^T\beta + \varepsilon) = 1(x^T\beta + \varepsilon > 0)$ and the conditional median of ε given x is 0, then the conditional median of y is $1(x^T\beta > 0)$. As noted above, the maximum score estimator minimizes the absolute distance between y and its median $1(x^T\beta > 0)$, or $\sum |y_i - 1(x_i^T\beta > 0)|$. We will return to these ideas when discussing estimators of Tobit models below.

Other approaches to estimating binary response models are based on semiparametric implementations of maximum likelihood. In particular, the likelihood function[4] could be maximized for β and F. Coslett (1983) proposes choosing F as a monotonic step function, and Klein and Spady (1990) use a kernel estimator in this way. Klein and Spady demonstrate that their estimator is $\sqrt{N}$ consistent, asymptotically normal, and asymptotically efficient.

B. Multinomial and Ordered Discrete Response Models

Multinomial response models are developed on the same theme. Suppose that one models choice of one among $J + 1$ alternatives, numbered 0, 1, 2, ..., J. As above, the utility of option j could be modeled as $\tilde{U}_j = U_j(x) + u_j$, and the net benefits of choosing j over l are $B_{j1} = \phi_{j1}(x) - \varepsilon_{j1}$, with $\phi_{j1}(x) = U_j(x) - U_1(x, \theta)$ and $\varepsilon_{j1} = u_1 - u_j$. Option j is chosen if its net benefits are positive across all the alternatives; i.e., if $B_{j1} > 0$ for all 1. Therefore, if y indicates the choice of j, $y = 1$ [j chosen], then the probability that j is chosen is given as

$$\begin{aligned} E(y \mid x) &= \text{Prob}\ \{\varepsilon_{j0} < \phi_{j0}(x); \cdots \varepsilon_{jJ} < \phi_{jJ}(x)\} \\ &= F[\phi_{j0}(x); \cdots; \phi_{jJ}(x)], \end{aligned}$$

where $F(\cdots)$ is the joint c.d.f. of $\varepsilon_{j0}, \ldots, \varepsilon_{j,j-1}, \varepsilon_{j,j+1}, \ldots, \varepsilon_{jJ}$.

[4]With reference to (1.10) of the first lecture, the likelihood function is $\sum \ln \mathcal{L}_d(d_i,\ x_i, \alpha/\sigma, \beta/\sigma)$, for whether one works (d_i=1) or not (d_i=0). To identify β and F, the parameters σ and α must be normalized (for instance, σ=1 and α=0).

Suppose the choice of another option, say j', were studied. It is easy to see that the same index variables used for option j determine the probability of choosing $j' \neq j$, since the net benefits of option j' over any option k' are given as $B_{j'k'} = B_{jk'} - B_{jj'}$. At any rate, information on the choice of different options could be combined to characterize the basic J index variables above.

Parametric implementation of a multinomial response model requires specification of the index functions $\phi_{j0}(x), \ldots, \phi_{jJ}(x)$ as well as the joint distribution $F(\cdots)$, up to a set of parameters to be estimated. Semiparametric approaches can be based on partial specifications of either of these features. For instance, Abe (1991) uses additive approximations to the index functions within the context of a multinomial logit specification of $F(\cdot)$.

In the spirit of methods applicable to binary response models, the utility differences could be assumed to depend on linear forms: $\phi_{j1} = \psi_1(x^T\beta_{j1})$, giving rise to a multiple index regression model, as in

$$E(y \mid x) = F[\psi_0(x^T\beta_{j0}), \ldots, \psi_J(x^T\beta_{jJ})] = G(x^T\beta_{j0}, \ldots, x^T\beta_{jJ}).$$

These coefficients could be estimated by applying a method applicable to a multiple index model for $E(y \mid x)$, giving an approach that does not rely on assumptions on the distribution F. As of this time, no other semiparametric methods have been devised for multinomial response models that make use of specific features of the choice model, but do not require a specification of the distribution F.

When the discrete responses are ordered in intensity, then it is natural to model the responses as the result of a continuous index falling into various ranges. For instance, the response could take on the values $y = 0, 1, 2, \ldots$ with 0 "less" than 1, 1 "less" than 2, etc. For concreteness, suppose that in our labor supply example, we observed hours of work for individuals in coarse ranges; not working ($y = 0$), part-time work ($y = 1$, say 25 or less hours per week) and full time work ($y = 2$, more than 25 hours per week). In this setting, an ordered response model is defined by a continuous latent variable $y^* = \phi(x) + \varepsilon$, the distribution of the unobserved term ε, and limits $c = -\infty, c_1, \ldots, c_J, c_{J+1} = \infty$, such that

$$y = j \quad \text{when} \quad c_j < \phi(x) + \varepsilon \leq c_{j+1}.$$

The stochastic model (and likelihood function) is determined by the probabilities that $y = j$ for each j; if $d_j = 1[y = j]$, the probabilities

$$E(d \mid x) = \text{Prob}\ \{y = j \mid x\} = \text{Prob}\ \{c_j - \phi(x) < \varepsilon \leq c_{j+1} - \phi(x)\}$$

which are determined by c_1, c_{j+1}, $\phi(x)$ and the distribution of ε. The regression of y on x determined as

$$E(y \mid x) = \sum jE(d_j \mid x) \equiv G^*(\phi(x), c_1, \ldots, c_J).$$

When the limits $c_1, \ldots, c_J$ are of secondary interest, the regression is a nonlinear function of the systematic term $\phi(x)$; if the term is a linear index $\phi(x) \equiv \psi(x^T\beta)$, then we have

$$E(y \mid x) = G(x^T\beta)$$

or the standard index model form. It is clear that approaches based on conditional medians and least absolute deviations could be applied here as well.

C. Censored and Truncated Regression Models

Standard treatment of variables that are bounded involve explicit modeling of the bounding process, as in the work-don't work distinction of the labor supply example. In particular, a variable y that varies smoothly but has several observations at a bound (say a lower bound of $y = 0$) is modeled as a censored version of an underlying continuous latent variable. A variable that is observed only when it obeys a bounding condition is likewise modeled as a truncated version of an underlying continuous random variable.

Censored regression, or Tobit models refer to a model of the form

$$y = [\phi(x) + \varepsilon]1[\varepsilon > -\phi(x)]$$

where $y = \phi(x) + \varepsilon$ is observed when positive, or zero otherwise. In the linear case $\phi(x) = a + x^T\beta$, this model is

$$y = [a + x^T\beta + \varepsilon]1[\varepsilon > -\alpha - x^T\beta] \tag{2.3}$$

as with the example of the last lecture. With $F(\varepsilon)$ the c.d.f of ε, the regression function in this case is

$$\begin{aligned} E(y \mid x) &= \{1 - F[-\alpha - x^T\beta]\}\{\alpha + x^T\beta + E[\varepsilon \mid \varepsilon > -\alpha - x^T\beta]\} \\ &= G(x^T\beta) \end{aligned}$$

where the correct form of G depends on the distribution F of ε.

The censored Tobit model lends itself immediately to the least absolute deviations approach discussed earlier, and this approach yields a procedure that achieves a $\sqrt{N}$ rate of convergence in a substantively nonlinear context. In particular, if ε has conditional median 0 in model (2.3), then the median of y is $(\alpha + x^T\beta)1[0 > -\alpha - x^T\beta]$, and estimation can be based on minimizing the absolute distance of y from its median, as

$$(\hat{\beta}, \hat{\alpha}) = \arg\min_{\tilde{\alpha},\tilde{\beta}} \sum \left| y_i - (\tilde{\alpha} + x_i^T\tilde{\beta})1[0 > -\tilde{\alpha} - x_i^T\tilde{\beta}] \right|.$$

Powell (1984) established general conditions under which these estimators are $\sqrt{N}$ consistent and asymptotically normal. Powell (1986b, 1991) has generalized the least absolute deviations approach to general models that are monotonic in unobserved disturbances, as well as for measuring quantiles other than the median. This is the quantile regression approach raised earlier, as advocated by Chamberlain (1991) as a general data analytic method.

The truncated regression, or Tobit model takes the form

$$y = \alpha + x^T\beta + \varepsilon; \quad \text{where } \varepsilon \text{ has c.d.f. } F(\varepsilon \mid \varepsilon > -\alpha - x^T\beta)$$

so that y is observed only when $\alpha + x^T\beta + \varepsilon > 0$. Again, the regression is in index form

$$\begin{aligned} E(y \mid x, y \text{ observed}) &= \alpha + x^T\beta + E[\varepsilon \mid \varepsilon > -\alpha - x^T\beta] \\ &= G(x^T\beta). \end{aligned}$$

An alternative approach to Tobit model estimation can be based on symmetry of the distribution of ε. As evident above, it is clear that the problems caused by censoring and truncation amount to the deletion of one tail of the distribution of ε, namely where $\varepsilon \leq -\alpha - x^T\beta$. If ε is symmetrically distributed, then we could consider chopping off the other tail in a symmetric fashion and performing least squares. Powell (1986a) noticed how this other tail, defined by $\varepsilon \geq \alpha + x^T\beta$, is equivalently determined by $y \geq 2(\alpha + x^T\beta)$, and so estimation can be based on solving the first order condition[5]

$$\sum 1[0 < y_i < 2(\tilde{\alpha} + x^T\tilde{\beta})](1, x_i^T)^T(y_i - \tilde{\alpha} - x_i^T\tilde{\beta}) \cong 0.$$

[5]Here "$\cong 0$" accounts for the fact that the left hand expression is discontinuous for finite N (wherein it may only be possible to find $\tilde{\alpha}$ and $\tilde{\beta}$ values so that the left hand side is $o_p(1)$).

Powell showed how the resulting estimators are $\sqrt{N}$ consistent and asymptotically normal. The same idea can be applied to censored Tobit models; symmetric zeroing; and Powell shows similar limiting properties for the estimators in this case. Recently, Honore (1991) has applied trimming and least absolute deviations in a novel method of estimating censored models with panel data.

D. Models of Selected Samples

In Tobit models, the bounding of the response y is modeled on the basis of the index defining the position of y when it is not subject to constraint. More broadly, one may model the censoring or bounding process as dependent on the value of a different index variable. With regard to our "work-don't work" example, we might consider an enhanced setting where because of fixed costs or other overriding concerns, the decision to work involves different considerations than how much to work, after working has been chosen. In this case, one latent index could be used to model the choice to work, with a different index indicating how many hours are worked. Suppose as before that the main structural equation of interest is modeled via a latent variable $y^* = \alpha_0 + x_0^T\beta_0 + \varepsilon$. Suppose further that the response y hits the bound 0 when another latent variable $\psi(x_1) + \nu$ is negative, or that $y = 0$ when $d = 1[\nu > -\psi(x_1)] = 0$. Therefore, the model for the censored response y and the discrete variable d is

$$\begin{aligned} y = y^*d &= [\alpha_0 + x_0^T\beta_0 + \varepsilon]1[\nu > -\psi(x_1))] \\ d &= 1[\nu > -\psi(x_1))] \end{aligned}$$

together with the joint distribution of the unobserved terms ε, ν. Note further that the regression for responses that are positive is

$$E(y \mid x, y > 0) = \alpha_0 + x_0^T\beta_0 + E[\varepsilon \mid \nu > -\psi(x_1)]$$

so that the departures from the basic model result from how the selection truncates the distribution of the structural disturbance ε. If the distribution of (ε, ν) is independent of x, then the final term of the regression depends only on the selection index $\psi(x_1)$. This setting gives rise to a "tiered" index structure for the regression model, namely

$$\begin{aligned} E(y \mid x) &= G_0[x_0^T\beta_0, \psi(x_1)] \\ E(d \mid x) &= G_1[\psi(x_1)]. \end{aligned}$$

This formulation suggests a two-stage approach: characterize the selection index $\psi(x_1)$ directly using observations on the selection variable d, and then measure β_0 from the observations with positive y values. The regression for positive values of y is written in partially linear form as

$$E(y \mid x, y > 0) = \alpha_0 + x_0^T \beta_0 + G_2[\psi(x_1)].$$

The structure of censoring and the selection index $\psi(x_1)$ can be developed more generally using the notion of "propensity score,"[6] since the selection variable d is observed directly. In particular, the probability $P(x_1) =$ Prob $(d = 1 \mid x_1) =$ Prob $(\nu > -\psi(x_1))$, or the "propensity score," can be studied directly, using methods appropriate for binary response models. When ν is distributed independently of x, then the probability P depends only on the selection index $\psi(x)$, and it is invertible in this index (over regions of positive density of ν). Consequently, we can write $\psi(x_1) = \eta[P(x_1)]$, and write the regression for positive y values as

$$\begin{aligned} E(y \mid x, y > 0) &= \alpha_0 + x_0^T \beta_0 + E[\varepsilon \mid \nu > -\eta[P(x_1)]] \\ &= \alpha_0 + x_0^T \beta_0 + \tau[P(x_1)] \end{aligned}$$

Therefore, the propensity score $P(x_1)$ can be used in place of the selection index $\psi(x_1)$ above. If the selection index $\psi(x_1)$ were not of interest, then one can dispense with it, using the propensity score directly.

This formulation also suggests studying the structural index using variations exhibited by individuals with the same "propensity score." A second stage estimation could be based on difference regressions on positive y values: since $E(y \mid P) = \alpha_0 + E(x_0 \mid P)^T \beta_0 + \tau(P)$, we have

$$y - E(y \mid P) = [x_0 - E(x_0 \mid P)]^T \beta_0 + \upsilon$$

where $E(\upsilon \mid x_0, x_1, y > 0) = 0$. Therefore, a regression of the differences $y - E(y \mid P)$ on $x_0 - E(x_0 \mid P)$ would measure β_0. Estimates of these regressions could be formed using estimates of the regressions $E(y \mid P)$ and $E(x_0 \mid P)$.

[6] See Heckman and Hotz (1989) for references to the literature on "propensity score" and selected samples.

Semiparametric implementation of these ideas focuses on estimation of β_0 with partial restrictions on the selection mechanism. For instance, when both index variables are linear, the regression for positive y values takes the form

$$E(y \mid x, y > 0) = \alpha_0 + x_0^T \beta_0 + \tau^*[x_1^T \beta_1].$$

Powell (1987b) proposes using an average derivative approach to estimate β_1 of $P = G_1(x_1^T \beta_1)$ up to scale, and a differencing approach to estimate β_0. Ahn and Powell (1990) propose a method based on direct nonparametric estimation of the propensity score $P(x_1)$, using the second stage "differenced" equation as above. Both of these methods are shown to yield coefficient estimators that are $\sqrt{N}$ asymptotically normal. Robinson (1988b), Stock (1989), Chamberlain (1986b) and Newey (1988) have developed a general implementation of this differencing method when the selection equation depends on different variables than the structural index. In particular, if x_0 and x_1 have no common components, then β_0 can be isolated as in

$$y - E(y \mid x_1) = [x_0 - E(x_0 \mid x_1)]^T \beta_0 + \upsilon_1$$

where $E(\upsilon_1 \mid x_0, x_1, y > 0) = 0$. Robinson (1988b) demonstrates how using nonparametric kernel estimators of $E(y \mid x_1)$ and $E(x_0 \mid x_1)$ can be used to produce $\sqrt{N}$ asymptotically normal estimates of β_0 . These ideas also clarify how valuable the direct observation of the selection mechanism is for estimating the structural equation. For instance, if x_0 and x_1 contained a component in common, the associated component of the departure $x_0 - E(x_0 \mid x_1)$ would be identically zero, so the associated coefficient in β_0 could not be measured from the difference regression above. The equation for the selection variable d does structure the impact of x_1, and permits separation of that impact from the impact of x_0 in the structural index.

When the selection process is not observed (say with a truncated sample), then this separation is a much more difficult problem, which has an immediate solution only when x_0 and x_1 have no common components. For instance, if $y = \alpha_0 + x^T \beta_0 + \varepsilon$ is observed only when $\nu > -\psi(x_1)$, then the regression of y on x is

$$E(y \mid x_0, x_1, y \text{ observed}) = \alpha_0 + x_0^T \beta_0 + G_2(x_1)$$

where we suppress $\psi(x_1)$ since there is no selection equation to infer $\psi(x_1)$ from. If x_0 and x_1 have no components in common, this is a partially linear model,

that can be approached via difference regressions as above (or with more general average derivative techniques). If x_0 and x_1 do overlap, then one faces the difficult obstacle of measuring the structural index $x_0^T\beta_0$ separately from the nonlinear term $G_2(x_1)$, while only observing their sum y (up to random error).

E. Transformation Models

The above examples are tied to concrete practical problems such as accounting for discreteness, or modelling bounds on response variables. It is clear that the index model restriction

$$E(y \mid x) = G(x^T\beta) \tag{2.1}$$

is quite weak, stating only that the mean of y is some function of $x^T\beta$. In particular, if the distribution of y were determined by $x^T\beta$, then this restriction would always hold (provided the mean of y exists). This would obtain, for instance, if the model dictated that there was a function ξ such that

$$\xi(y, x^T\beta) = \varepsilon$$

where ε was distributed independently of x.

This structure arises in many contexts, such as those discussed above, as well as general transformation models. We raise this specifically to also point out some weaknesses in estimation based on the index restriction. Suppose that there exist functions ζ_1 and ζ_2 such that the data is generated from

$$\zeta_1(y) = \zeta_2(x^T\beta) + \varepsilon$$

where ζ_1 is invertible and ε is distributed independently of x. In practice, parametric models of ζ_1 and ζ_2 include the popular Box–Cox (1964) power transformation encompassing linear and logarithmic transformations. The index restriction is spelled out explicitly as

$$E(y \mid x) = E(\zeta_1^{-1}[\zeta_2(x^T\beta) + \varepsilon)] \mid x) = G(x^T\beta).$$

The fact that such transformation models are captured is a strength, although they also highlight deficiencies. In particular, in some contexts (such as duration models below), the transformation ζ_1 is of intrinsic interest, and it cannot obviously be learned from an estimate of G. Even if ζ_2 is known (for instance as the identity), ζ_1 differs from G because of the smoothing over the random term ε.

It is also worth remarking that the popular framework of generalized linear (GLIM) models are a variation of this transformation set-up.[7] In particular, such models assume that the regression of y on x can be transformed to a linear model as in

$$H[E(y \mid x)] = \alpha + x^T\beta,$$

where $H(\cdot)$ is invertible. This specializes of the basic index model restriction by requiring G to be invertible, with $G^{-1}(\cdot) = H(\cdot - \alpha)$.

F. Duration Models

Our final example concerns how index restrictions arise in the context of duration models. Duration models are designed to explain the distribution of (continuous) duration T in a state, conditional on observed variables x, such as how long one remains unemployed. While regression models can be appropriate for such data, it is more standard to model the distribution of durations through the use of an (instantaneous) hazard function. Here we discuss enough of this framework to motivate regression based, index model approaches. We will later specify the response y as the duration, or a known transformation such as log-duration. For this discussion we assume durations are not censored; if partial durations are observed then they can be modeled in standard fashion using the truncation ideas above.

The statistical structure for such data is captured by the cumulative distribution

$$\tilde{F}(t \mid x, \theta) = \text{Prob}\,\{T < t \mid x, \theta\}$$

where for simplicity, we suppose the predictors x are constant over the duration period studied. Instead of modeling this distribution directly, it is more common to model the hazard function, or (instantaneous) probability of changing states, or ending a duration spell:

$$\lambda(t; x, \theta) = \lim_{h \to 0} \text{Prob}\,\{t \le T < t + h \mid T \ge t; x, \theta\}/h.$$

[7] See McCullagh and Nelder (1983).

This approach is equivalent to modeling the distribution: if $\tilde{f} = \partial \tilde{F}/\partial t$ is the density, and $S = 1 - \tilde{F}$ is the survivor function, then

$$\lambda(t, x; \theta) = f(t, x; \theta)/S(t, x; \theta).$$

If we define the integrated hazard as

$$\Lambda(t, x; \theta) = \int_0^t \lambda(u, x; \theta) du$$

then the survivor function is determined as

$$S(t, x; \theta) = \exp[-\Lambda(t, x; \theta)].$$

For later reference, duration dependence refers to how the hazard changes with duration: no duration dependence means $\partial\lambda/\partial t = 0$, and positive or negative duration dependence corresponds with whether $\partial\lambda/\partial t$ is positive or negative.

Taking the integrated hazard $\Lambda(t, x; \theta)$ to be invertible in duration t, it is easy to see how a regression model arises from the above framework. In view of the survivor function, it is easy to see that if we define $\varepsilon = -\ln \Lambda(t, x; \theta)$, then the distribution of ε is independent of x and the structure of Λ, namely

$$\begin{aligned} \text{Prob } \{\varepsilon \leq E \mid x; \theta\} &= \text{Prob } \{\Lambda(y, x; \theta) > \exp(-E)\} \\ &= \text{Prob } \{y > \Lambda^{-1}(\exp(-E), x; \theta)\} \\ &= S(\Lambda^{-1}(\exp(-E), x; \theta)) = \exp(-\exp(-E)) \end{aligned}$$

so that ε has a type 1 extreme value distribution.

Consequently, if the dependence of the hazard function on x is modeled through an index $x^T\beta$, then the "model"

$$-\ln \Lambda(t; x^T\beta) = \varepsilon$$

fits the tenets of the transformation models already discussed, and leads to the index model

$$E(y \mid x) = G(x^T\beta)$$

where y is any known transformation of t. We take $y = \ln t$ for the remainder of this example.

However, as noted above, the function G of the index model is not sufficient to identify the original function Λ, and consequently will not immediately permit one to assess duration dependence issues. To see this clearly, consider a traditional proportional hazard model, where

$$\lambda(t, x) = \phi(x^T\beta)\lambda_0(t)$$

with $\lambda_0(t)$ the baseline hazard, that is scaled up or down by x as above. In this case, the regression model is

$$\ln \Lambda_0(t) = -\ln \phi(x^T\beta) - \varepsilon.$$

No duration dependence occurs when $\ln \Lambda_0(t) = \ln t = y$, although this form is not discernible from the function G, without further restrictions. Some prospects exist if $\phi(x^T\beta)$ is known exactly, in which case "no dependence" translates to a linear G function with known slope. But if β is measured up to unknown scale, then the scale is inherently intwined with the duration dependence of the process. An alternative way of seeing this problem is to note simply that the index model restriction does not make use of the form of the distribution of ε, namely the extreme value distribution.

The most familiar example of a semiparametric estimation method for duration models is Cox's (1972) partial likelihood estimator. Applicable to proportional hazards models without additional heterogeneity, this estimator is shown to be $\sqrt{N}$ consistent and asymptotically normal by Tsiatis (1981).

Semiparametric approaches to the duration problem typically deal with further complications due to unobserved individual heterogeneity. In the above formulation, ε arises from randomness of durations, where individuals are different only through the observed variables x. If one permits unobserved differences in individual characteristics, say represented by τ, then there is an additional level of smearing to complicate the duration dependence structure. If $\tilde{F}(t \mid x, \tau, \theta) = \text{Prob } \{T < t \mid x, \tau, \theta\}$ denotes the distribution of durations given x, τ, and τ has c.d.f. $\tilde{F}^*(\tau \mid x)$, then the observed distribution of durations given x is

$$\mathcal{F}(t \mid x) = \int \tilde{F}(t \mid x, \tau) d\tilde{F}^*(\tau \mid x).$$

Duration dependence is a property of $\tilde{F}(t \mid x, \tau)$, of which the observed distribution $\mathcal{F}$ is a smear. Moreover, it is well known how this smearing can create apparent

dependence when there is none in the basic process: if there are two types of individuals with constant but differing hazard rates, then the early completion of spells for the high hazard group relative to the low hazard group will induce apparent dependence into $\mathcal{F}$. In terms of our discussion above, if $\tilde{F}$ and $\tilde{F}^*$ depend only on an index $x^T\beta$, then an index model $E(y \mid x) = G(x^T\beta)$ is appropriate, but G then involves conditional integration over both ε and τ, and thus is less connected to a clear summary of duration dependence. This setting would arise if the heterogeneity were modeled by an additive term, namely as $\varphi(\tau + x^T\beta)$ in the above, where τ is independent of x.

Semiparametric approaches to duration models have focused on these latter issues, by treating only one aspect of the above as flexible, and the others through fully parametric models. Heckman and Singer (1984) implement a nonparametric maximum likelihood approach, by using a parametric model for $\tilde{F}(t \mid x, \tau)$ and approximating $\tilde{F}^*$ as a discrete distribution; estimating both the number of different values of τ as well as the percentage of individuals with each value. Other estimation methods that treat the hazard and/or heterogeneity distribution in a flexible manner are given in Kalbfleisch and Prentice (1980), Meyer (1987) and Hausman and Han (1990), among others.

II. Various Approaches to Estimating Index Models

The next two lectures cover approaches that use average derivative estimators of index model coefficients. To round out our coverage, we now discuss various general ideas directly connected to the index restriction (2.1).

A. Least Squares Estimation

The fact that the restriction (2.1) is a regression restriction immediately suggests considering least squares estimation. In particular, the defining restriction for a regression function $m(x) = E(y \mid x)$ is

$$m(x) = \arg\min_{\Gamma} E[(y - \Gamma) \mid x].$$

This suggests estimating G and β of $E(y \mid x) = G(x^T\beta)$ by the sample analog

$$(\hat{G}, \hat{\beta}) = \arg\min_{|\tilde{\beta}|=1} \sum [y_i - \tilde{G}(x_i^T\tilde{\beta})]^2.$$

This approach has a fairly long history, as the first step on projection pursuit regression. A computationally practical approach based on using nonparametric estimators in the first order conditions for this minimization is given by Ichimura (1986), and enhanced to include automatic smoothing parameter choice in Hardle, Hall and Ichimura (1991).

B. Unconditional Estimation

We now consider estimation approaches that are (surprisingly) based on using information about the distribution of the regressors x. Work on these procedures was originally stimulated by the following remarkable fact:

> *If the model is in index form $E(y \mid x) = G(x^T\beta)$, and x has a joint normal distribution, then the least squares coefficients of y regressed on x (and a constant) consistently estimate β up to scale.*

For notation, let $\Sigma_{xz} = \text{cov}(x, z) = E[(x - E(x))(z - E(z))^T]$, and S_{xz} denote the corresponding sample covariance; for univariate random variables we may occasionally use σ_{xz} and s_{xz}. Consequently, the OLS coefficients of y regressed on x are $\widehat{b} = (S_{xx})^{-1}S_{xy}$, with plim $\widehat{b} = (\Sigma_{xx})^{-1}\Sigma_{xy}$.

We will give two separate derivations of the above fact about OLS coefficients, because each motivates separate approaches. The first assumes that $y = \tilde{g}(x^T\beta + \varepsilon)$, where ε is independent of x, as well as the linearity condition

$$E(x \mid x^T\beta) = A + B[x^T\beta]$$

which holds for x's whose distribution is spherically symmetric, such as with the multivariate normal distribution.

For simplicity, denote the index as $\mathrm{x} = x^T\beta$. The (large sample) OLS coefficients of x regressed on x are $\beta = \Sigma_{xx}^{-1}\Sigma_{x\mathrm{x}}$, and the coefficients of x regressed on x are $B = \sigma_{\mathrm{xx}}^{-1}\sum_{\mathrm{x}x}$. We therefore have

$$\begin{aligned}\text{plim } \widehat{b} &= \Sigma_{xx}^{-1}\Sigma_{xy} = \Sigma_{xx}^{-1}E[(x - E(x))y] = \Sigma_{xx}^{-1}E_{\mathrm{x}}E[(x - E(x))y \mid \mathrm{x}] \\ &= \Sigma_{xx}^{-1}E[B(\mathrm{x} - E(\mathrm{x}))G(\mathrm{x})] = \Sigma_{xx}^{-1}B\sigma_{\mathrm{x}y} = \Sigma_{xx}^{-1}\Sigma_{\mathrm{x}x}\sigma_{\mathrm{xx}}^{-1}\sigma_{\mathrm{x}y} \\ &= [\sigma_{\mathrm{xx}}^{-1}\sigma_{\mathrm{x}y}]\beta\end{aligned}$$

so that $\widehat{b}$ consistently estimates β up to scale, where the scale factor is the large sample regression coefficient of y on $\mathrm{x} = x^T\beta$. See Ruud (1983) and Chung and Goldberger (1984) for similar formulations.

The second derivation is based on the concept of average derivatives, as follows. Assume that the index model form $m(x) = E(y \mid x) = G(x^T\beta)$ has G a.e. differentiable, and x is a continuously distributed vector with density $f(x)$, that vanishes on the boundary of x values. The index model formulation is equivalent to the restriction

$$m'(x) \equiv \frac{\partial E(y \mid x)}{\partial x} = \frac{dG}{d(x^T\beta)}\beta = \gamma(x^T\beta)\beta;$$

that the derivatives of the regression function are proportional to β, although the factor $\gamma(x^T\beta)$ may vary with x. Define the (unweighted) average derivative as

$$\delta = E(m')$$

and recall that for an index model, we have

$$\delta = E(m') = E[\gamma(x^T\beta)]\beta = \gamma\beta$$

so that the vector of average derivatives are proportional to β, given $\gamma \neq 0$.

The connection to OLS coefficients is seen by the following derivation, that is of some independent interest. Let $\ell(x) = -(\partial f/\partial x)/f(x)$, or the (translation) score vector of $f(x)$. Then the average derivative is

$$\begin{aligned} E(m') &= \int m' f dx = -\int m f' dx \\ &= -\int m\left(\frac{f'}{f}\right) f dx = E[\ell(x)m(x)] = E(\ell y) = \mathrm{cov}(\ell, y) \end{aligned}$$

where the second inequality follows from integration by parts, and the final equality follows from $E(\ell) = 0$. Now, if $x \sim \mathcal{N}(\mu_x, \Sigma_{xx})$, then it is easy to verify that $\ell(x) = \Sigma_{xx}^{-1}(x - \mu_x)$, so that

$$E(m') = \mathrm{cov}(\ell, y) = \Sigma_{xx}^{-1}\Sigma_{xy} = \mathrm{plim}\ \widehat{b}.$$

Stoker (1986) gives the derivation above, and Brillinger (1983) uses a similar argument. Use of information on the density $f(x)$ of x to measure average derivatives will be discussed in the next lecture.

The above fact about normal regressors also suggests estimation methods based on reweighting the data sample. The essence of this idea is captured in the following. Suppose $\widehat{b}_w$ is the weighted least squares estimator

$$\widehat{b}_w = \left[\sum w(x_i)(x_i - \overline{x})(x_i - \overline{x})\right]^T \left[\sum w(x_i)(x_i - \overline{x})(y_i - \overline{y})\right].$$

Suppose that $\varphi(x)$ denotes the density of any normal distribution. If $w(x) = \varphi(x)/f(x)$, then in a large sample $\widehat{b}_w$ will behave like the OLS coefficients of y regressed on x, where x has the normal density $\varphi(x)$. Therefore, with appropriate reweighting, $\widehat{b}_w$ will consistently estimate β up to scale.

Clearly, if $f(x)$ were known, $w(x)$ could be formulated. If $f(x)$ were estimated up to a finite parameterization, $w(x)$ could be estimated, and provided some regularity conditions, $\widehat{b}_w$ would be $\sqrt{N}$ consistent and asymptotically normal. Ruud (1986) proposes a nonparametric reweighting procedure, showing its consistency, and Newey and Ruud (1991) demonstrate $\sqrt{N}$ consistency for coefficients from a reweighting procedure of this type. Ruud (1986) also extends these ideas to a quasi-maximum likelihood approach.

Li and Duan (1989) derive many results on M estimation when the true model is of the form $y = \tilde{g}(x^T\beta, \epsilon)$, where ϵ is independent of x, and the design of the regressors obeys the linearity condition $E(x \mid x^T\beta) = A + B[x^T\beta]$. Li (1991) exploits this kind of linear structure for multiple index problems in a different way. Taking the case of one index, Li shows how the inverse regression $E(x \mid y) - E(x)$ lies in a subspace spanned by $\beta\Sigma_{xx}$, proposes estimating $E(x \mid y)$ by a sliced regression method, and shows how a $\sqrt{N}$ consistent estimate of β (normalized as a direction) can be obtained from principal component analysis of the inverse regression estimates. The multiple index case (for the true model and the linearity condition) provides estimates of all sets of normalized directions along the same lines.

LECTURE 3
AVERAGE DERIVATIVE ESTIMATION

The first two lectures have been devoted to motivating semiparametric estimation as a field of interest, and showing how index models can provide a useful benchmark approach. We now focus on a specific method of estimation, namely the average derivative approach. In this lecture, we will embellish the motivation somewhat, propose specific estimators, and show how their approximate distribution theory works.

I. The Average Derivative Approach

As before, we denote the mean regression of y on x as $m(x) = E(y \mid x)$, which we assume is a.e. first differentiable in x. We take x as continuously distributed with density $f(x)$, where $f(x)$ vanishes on the boundary of x values, and is also first differentiable.[1] The local effects of changing x on y are given as the derivative $m'(x) = \partial m/\partial x$, and the (unweighted) average derivative is the mean of these effects over the population:

$$\delta = E(m')$$

where expectation is taken with respect to x.

Our discussion of index models in the first two lectures was partly a prelude for the motivation of the average derivative δ as a concept of practical interest. In particular, when we have an index model

$$m(x) = G(x^T\beta)$$

then $m' = (dG/d(x^T\beta))\beta$ is proportional to β for all x, and likewise

$$\delta = E[dG/d(x^T\beta)]\beta = \gamma\beta$$

is proportional to β. We can equivalently replace β by δ (provided $\gamma \neq 0$), redefining G so that

$$m(x) = G(x^T\delta)$$

which has G obeying the normalization $E[dG/d(x^T\delta)] = 1$, or that δ is scaled so that y and $x^T\delta$ are, on average, related in a 1-1 manner. This says that the values of δ are interpretable in units of "y changes"/"x changes," and is related to the fact that for linear models, δ equals the linear coefficients.

The "ADE Method" of Härdle and Stoker (1989) suggests studying data on $\{y_i, x_i\}$ by first estimating δ, say with $\widehat{\delta}$, forming $z_i = x_i^T\widehat{\delta}$, and then estimating a one-dimensional regression G between y and $x^T\widehat{\delta}$. The resulting regression $\widehat{G}(x^T\widehat{\delta})$ will in general estimate $E(y \mid x^T\delta)$ at the convergence rate appropriate for one dimensional problems, and when the true model is in index form, this method completely characterizes the true regression.

[1] We consider issues raised by diserete predictor variables in Lecture 5.

We now turn to a simple example, that illustrates how semiparametric estimators can be used for parsimonious data summary, in combination with the binary response model introduced in Lecture 2. In particular, this example involves variable choice and functional specification in the single index framework.

A. An Empirical Example: Collision Data

We now consider estimates from the collision data of Kallieris, Mattern and Härdle (1989), which is given in Härdle and Stoker (1989). The data consists of 58 observations from simulated automobile collisions, where the response $y = 1$ if the accident resulted in a fatality and $y = 0$ if not. The predictor variables are the age of the driver (AGE, x_1), the velocity at impact (VEL, x_2) and the acceleration at impact (ACL, x_3), which are standardized for the analysis (each variable is centered by its sample mean and divided by its standard deviation). The regression $E(y \mid x)$ is the probability of a fatality given the predictors x. The ADE method implements the index regression $E(y \mid x) = G(x^T\delta)$; we will decide whether each of the predictor variables has a significant impact by examining the estimates of δ, and further characterize the probability by examining the estimate of G.

This data is analyzed in Härdle and Stoker (1989), and the method here is altered in only two respects. First, the bandwidth values are chosen by generalized cross validation (GCV).[2] Second, we use a different average derivative estimator, the "indirect slope" estimator discussed below. As noted in Härdle and Stoker (1989), the original estimates ("indirect") exhibited problems with the scaling normalization $E(G') = 1$. Subsequent analysis, to be given in Lecture 5, suggests the use of a slope estimator to correct this problem. Estimates and standard errors are computed in line with the statistical theory discussed later in this lecture. The Appendix contains detailed formulae for these calculations.

The average derivative estimates are contained in Table 3.1, where we also include the OLS (ordinary least squares) estimates of y regressed on x for comparison. We see that AGE has the largest impact, followed by VEL and ACL. An initial examination of the standard errors suggests that given AGE and VEL, ACL may may have no impact ($\delta_3 = 0$), and so we proceed to test this formally using a Wald statistic. In particular, to test the restriction $R\delta = r_0$, the Wald

[2] See Craven and Wahba (1979).

statistic

$$W = N(R\widehat{\delta} - r_0)^T(R\widehat{\Sigma}_\delta R^T)^{-1}(R\widehat{\delta} - r_0)$$

is compared to a χ^2 (rank R) critical value, where $\widehat{\delta}$ is the average derivative estimate and $\widehat{\Sigma}_\delta$ is a consistent estimate of its asymptotic variance-covariance matrix. On the basis of the test results given in Table 3.1, we drop ACL, and reestimate the average derivatives for AGE and VEL.

TABLE 3.1:
AVERAGE DERIVATIVE ESTIMATES FOR COLLISION DATA
$(N = 58; h = 1.6)$

		ADE	ADE (ACL omitted)	OLS
AGE	(δ_1)	.318 (.082)	.330 (.031)	.321 (.095)
VEL	(δ_2)	.139 (.097)	.157 (.049)	.103 (.144)
ACL	(δ_3)	.062 (.087)		.084 (.145)

(ADE are Indirect Slope Estimators, Standard Errors in Parentheses)

Wald Statistics

Restriction	Value W	d.f.	Prob$[\chi^2$(d.f.) $> W]$
$\delta_1 = \delta_2 = \delta_3 = 0$	15.79	3	.0012
$\delta_2 = \delta_3 = 0$	5.983	2	.047
$\delta_3 = 0$	.928	1	.335

Further analysis is based on forming the index $x^T\delta = x_1\widehat{\delta}_1 + x_2\widehat{\delta}_2$, and estimating G by a kernel regression of y on $x^T\widehat{\delta}$. Choosing the bandwidth by GCV gives $h = .19$, and we display the resulting curve $\widehat{G}$ in Figure 3.1. The estimates $\widehat{G}$ and $\widehat{\delta}$ give the index model representation of the probability of a fatality from the collision data. The probability is increasing in the index, save for a nonmonotonic range for the high index values.

It is natural to ask whether this data is consistent with a standard binary response model of the form

$$y = \begin{cases} 1 & \text{if} \quad \varepsilon \le x^T\beta \\ 0 & \text{otherwise} \end{cases}$$

where the c.d.f. of ε is G^* , for instance of normal or logistic form (corresponding to a probit or logit model). Following the methods discussed in the next lecture, we could test any specific parametric model formally. However, at this juncture it is useful to see what our estimates imply about such a model.

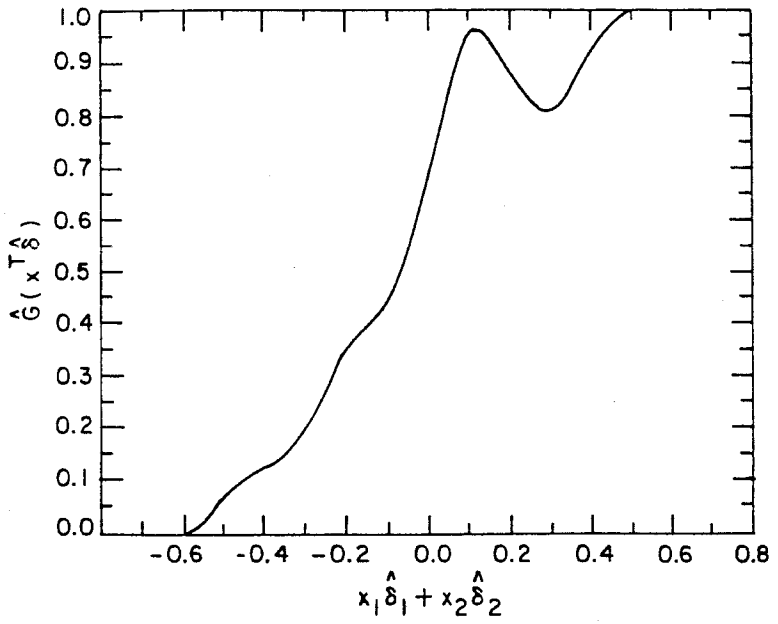


Figure 3.1
Probability of a Fatality $(y = 1)$ from Collision Data $(h = .19)$

First, the binary response model is an index model, and so we replace β by δ, absorbing the appropriate scaling. The regression implied by the binary response model is $E(y \mid x) = \text{Prob}\,\{\varepsilon \le x^T\delta\} = G^*(x^T\delta)$, which implies that the probability of a fatality is monotonically increasing in the index $x^T\delta$. Moreover, it is easy to see that the derivative of the c.d.f. G^* is the density f^* of ε.

We study this possibility with our ADE estimates by increasing the bandwidth used for fitting $\widehat{G}$ to the point where it is monotonic in the index $x^T\widehat{\delta}$, thereby implementing "critical" smoothing. This bandwidth value is $h^* = .34$, and the resulting estimator $\widehat{G}^*$, given in Figure 3.2. Finally, the derivative of $\widehat{G}^*$, namely $\widehat{f}^*$, is depicted in Figure 3.3. The net impact of these pictures is that a simple binary response model such as a probit or logit model does not capture all the nuances of the estimated probability function. In particular, there are three distinct modes, suggesting that the distribution of ε would best be modeled as a mixture, such as a mixture of normals. The upper mode is particularly pronounced, suggesting the greatest departure from a probit or logit model would occur in that range. In this way, the semiparametric approach can be more informative than reliance on a standard parametric model.

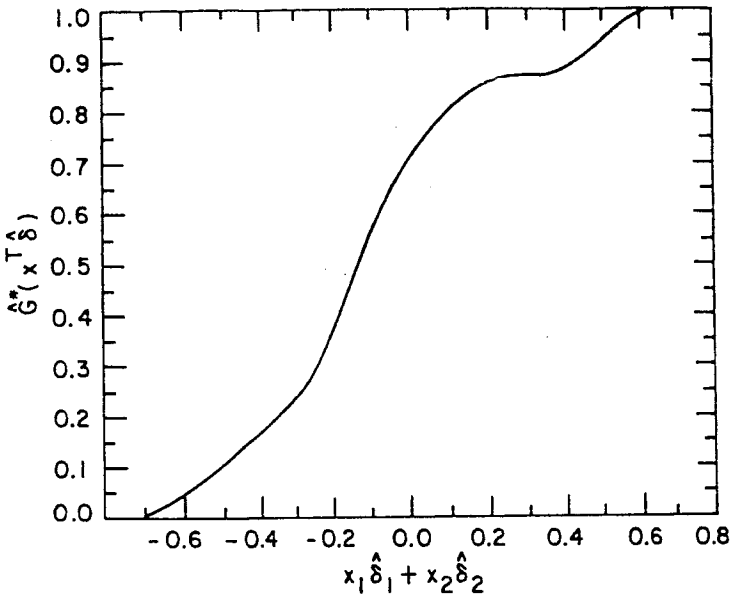


Figure 3.2
Smoothed Probability ($h = .34$)

Table 3.2 gives the (maximum likelihood) coefficient estimates from probit specifications, where the c.d.f. G^* of ε is normal with mean α and variance 1 (α is the estimated "Constant"). While recognizing that the probit coefficients are subject to a different scale normalization than the average derivatives, we see a broadly similar pattern of effects depicted by the probit estimates. Some more pointed differences emerge from the relative coefficient values (which should be comparable if the probit specification is correct). For instance, we see that the probit estimates dictate less relative impact of VEL to AGE when ACL is included, and more relative impact when ACL is omitted, than the average derivative estimates. Finally, while not reported, the probit specification is rejected against the single index specification using regression test statistics (discussed in the next lecture), as would be expected from Figures 3.2 and 3.3.

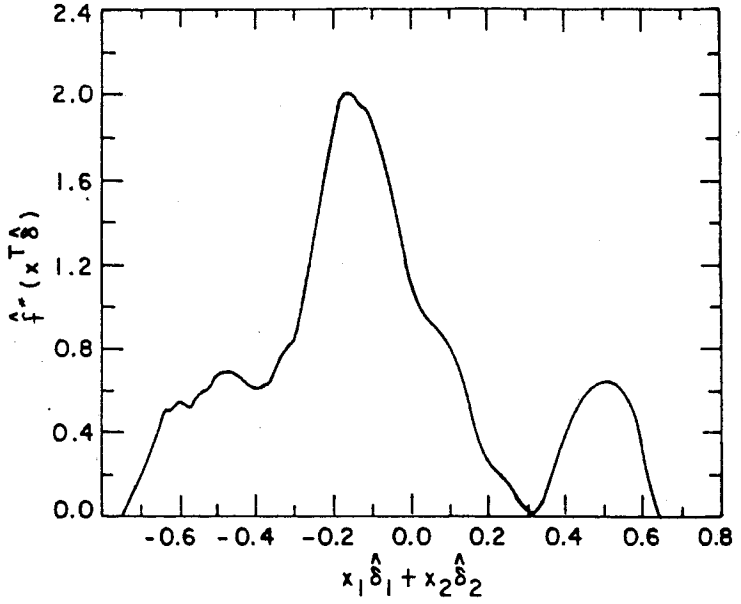


Figure 3.3
Derivatives of Smoothed Probability ($h = .34$)

TABLE 3.2: PROBIT ESTIMATES FOR COLLISION DATA
($N = 58$)

		Probit	Probit (ACL omitted)
Constant		.395 (.220)	.381 (.218)
AGE	(δ_1)	1.290 (.289)	1.251 (.286)
VEL	(δ_2)	.448 (.335)	.739 (.265)
ACL	(δ_3)	.342 (.338)	

(Standard Errors in Parentheses)

Relative Coefficient Estimates, ADE and Probit

	ADE	ADE (ACL omitted)	Probit	Probit (ACL omitted)
δ_2/δ_1	.437	.475	.378	.591
δ_3/δ_1	.194		.265	

B. Further Motivation

Average derivatives are of interest to a greater range of models than pure index models, as well as to some specific economic problems. We now outline some of these cases. The first, partial index models, are used in the testing format to be introduced in Lecture 4.

B.1 Partial Index Models

Given that many of the limited dependent variable models discussed earlier involve multiple index, or partially linear structure, it behooves us to consider how average derivatives could be used in these contexts. For this, partition x as (x_1, x_2) into a $k - k_2$ vector x_1 and an k_2 vector x_2 , and partition δ analogously as (δ_1, δ_2). Average derivatives will measure coefficients when the true regression

obeys partial index structure (c.f. Newey and Stoker (1989)): if

$$m(x) = G(x_1^T \beta_1, x_2)$$

then δ_1 equals β_1 up to scale. With an estimator $\widehat{\delta}_1$ of δ_1, we can extend the ADE Method to fitting a $k_2 + 1$ dimensional regression in the second stage, as $\widehat{G}(x_1^T \widehat{\delta}, x_2)$. If the model is in multiple index form, as

$$m(x) = G(x_1^T \beta_1, x_2^T \beta_2)$$

then δ_2 is likewise proportional to β_2 , but with a different scale factor (namely $E[\partial G/\partial (x_2^T \beta_2)]$ than that which connects δ_1 and β_1) (namely $E[\partial G/\partial (x_1^T \beta_1)]$). Again, the ADE method is easily extended. We discuss methods of empirically characterizing index structure using models of this type in the next lecture. If the model suggests that a component of x is in both indices, then the average derivative of that component will be a weighted average of the respective β_1 and β_2 components; this gives one setting where a single summary coefficient will not adequately represent many effects of a single variable, and where one would naturally expect to need more modeling structure to sort out such effects.

The scaling issues mollify with linearity. Clearly, if the true regression is linear

$$m(x) = \alpha + x^T \beta$$

then $\delta = \beta$. If the model is partially linear, as in

$$m(x) = \alpha + x_1^T \beta_1 + g(x_2)$$

then $\delta_1 = \beta_1$.[3] Moreover, if the additive function is in index form,

$$m(x) = \alpha + x_1^T \beta_1 + g(x_2^T \beta_2)$$

then $\delta_1 = \beta_1$ and δ_2 is proportional to β_2. The earlier remarks apply regarding predictor variables that appear in both index variables — the average derivative is a smear of the two index effects.

[3] From our discussion of models of selected samples of the last lecture, we could apply this framework with x_2 set to the propensity score P. It is natural to conjecture that average derivative estimators (computed using an estimate of P) would consistently measure the linear coefficients β_1 as above.

Finally, note that proportionality to index coefficients would hold for any weighted average derivative, say $\delta_w = E[w(x)m'(x)]$. If the weighting function has mean $[w(x)] = 1$, then δ_w would match linear model coefficients as above.

B.2 Economic Applications Unrelated to Index Models

Further motivation for average derivative estimators can be found in specific measurement problems in economics. A primary example is that of Härdle, Hildenbrand and Jerison (1991) on measuring the aggregate income effects matrix for assessing the "Law of Demand". Under various assumptions that motivate studying demand Y_j for good j as a function of income x, they show how the symmetrized income effects matrix of consumer demand theory takes the form $A = [\delta_{jj'}]$, where

$$\delta_{jj'} = E[\partial E(Y_j Y_{j'} \mid x)/\partial x].$$

These authors establish general demand properties, including that demand curves slope downward, using average derivative estimators of this matrix. This work represented the first large scale data analysis based on nonparametric average derivative estimators.

Further applications are suggested by the central role played by derivatives in economic modeling, in the form of marginal reactions or elasticities. For instance, suppose the data consists of log-production y, the log-input vector x_1 and other technology variables x_2 , and the regression $m(x_1, x_2)$ is a log production function. In this case $m_1' = \partial m/\partial x_1$ is the vector of output elasticities, and $\delta_1 = E(m_1')$ the mean output elasticity vector over the sample. The "scale elasticity" for each firm is $i^T m_1'$, and the restriction of "constant returns to scale" coincides with $i^T m_1' = 1$. Thus $i^T \delta_1$ is the mean scale elasticity, and a test of $i^T \delta_1 = 1$ tests whether returns are constant on average. For regression style average derivative estimators, this approach generalizes the method of fitting a Cobb–Douglas production function and testing whether the sum of the log input coefficients is one. See Stoker (1989) for a development of tests of this type involving average first and second derivatives (in indirect form), with the latter applicable to testing the additivity of the basic regression model . This approach to functional structure is generalized by Lewbel (1991) and Samarov (1990), with the latter focusing on empirical determination of the number of approximants in projection pursuit regression.

II. Kernel Estimation of Average Derivatives

With this motivation in hand, we now turn to the specifics of measuring average derivatives, as well as characterizing their distribution structure. This development illuminates some standard distributional characteristics of semiparametric "plug in" estimators; or procedures that use nonparametric estimators as ingredients.

A. Various Average Derivative Estimators

The defining condition $\delta = E(m')$ naturally suggests measuring δ by averaging nonparametric derivative estimators. We now introduce several types of estimators of this kind, each of which employ kernel density and regression estimators. In the following, denote the (Rosenblatt–Parzen) kernel density estimator as

$$\widehat{f}(x) = \frac{1}{Nh^k} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right),$$

where K is a kernel function that gives the weights for local averaging. The (Nadaraya–Watson) kernel regression estimator is denoted as

$$\widehat{m}(x) = \widehat{c}(x)/\widehat{f}(x)$$

where

$$\widehat{c}(x) = \frac{1}{Nh^k} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) y_i.$$

We also have need for the associated density derivative estimator,

$$\widehat{f}'(x) = \frac{1}{Nh^{k+1}} \sum_{i=1}^{N} K'\left(\frac{x - x_i}{h}\right),$$

the associated estimator of the (translation) score $\ell(x) = -\partial \ln f/\partial x$,

$$\widehat{\ell}(x) = -\widehat{f}'(x)/\widehat{f}(x)$$

and the regression derivative estimator

$$\widehat{m}'(x) = \frac{\widehat{c}'(x)}{\widehat{f}(x)} + \widehat{m}(x)\widehat{\ell}(x).$$

Under well-known bandwidth and regularity conditions, the estimators $\widehat{f}(x)$, $\widehat{m}(x)$, $\widehat{f}'(x)$, $\widehat{\ell}(x)$ and $\widehat{m}'(x)$ are consistent nonparametric estimators of $f(x)$, $m(x)$, $f'(x)$, $\ell(x)$ and $m'(x)$ respectively (c.f. Härdle (1991)). It should be noted that other nonparametric estimators could be used below; although the statistical properties cited would need to be established for those procedures.

Define the "direct" average derivative estimator as the (trimmed) sample analog of $\delta = E(m')$:

$$\widehat{\delta}_{\text{dir}} = N^{-1} \sum_{i=1}^{N} \widehat{m}'(x_i) \widehat{I}_i$$

where $\widehat{I}_i = 1[\widehat{f}(x_i) \geq b]$ is an indicator that drops observations with small estimated density b. This indicator is suggested to control for erratic behavior induced by division by the density in the above formulae; and is required for technical reasons in the distributional analysis of this estimator.

On the basis of applying integration by parts, as in Lecture 2, we have $E(m') = E(\ell y)$. Härdle and Stoker (1989) proposed an "indirect" estimator, the sample analog of the latter expectation.

$$\widehat{\delta}_{\text{ind}} = N^{-1} \sum_{i=1}^{N} \widehat{\ell}(x_i) y_i \widehat{I}_i.$$

The terminology is due the the somewhat indirect use of density estimates to measure average regression derivatives.

We consider two "slope" estimators, or estimators that would measure coefficients of a linear equation fit between y and x:

$$y_i = \widehat{c} + x_i^T \widehat{d} + \widehat{u}_i \qquad i = 1, \ldots, N$$

that are computed using appropriate instrumental variables. These are motivated by the following: Since $E(\ell) = 0$, $E(\ell y) = \text{cov}(\ell, y) = \Sigma_{\ell y}$. Now, the average derivative of x^T is $E(\partial x^T / \partial x) = Id$, the $k \times k$ identity matrix. Moreover, we have $E(\partial x^T / \partial x) = Id = E(\ell x^T) = \text{cov}(\ell, x) \equiv \Sigma_{\ell x}$. Consequently,

$$\delta = E(m') = \left[E\left(\frac{\partial x^T}{\partial x} \right) \right]^{-1} E(m')$$

$$= [\text{cov}(\ell, x)]^{-1} \text{cov}(\ell, y) = \Sigma_{\ell x}^{-1} \Sigma_{\ell y}.$$

The latter takes the familiar form of the limit of a linear instrumental variables estimator. Define the "indirect slope" estimator as the linear coefficients obtained by using $(1, \widehat{\ell}(i)\widehat{I}_i)$ as the instrumental variable, or

$$\widehat{d}_{\text{ind}} = S_{\ell x}^{-1} S_{\ell y}$$

where $S_{\ell y} = N^{-1} \sum \ell(x_i) \widehat{I}_i (y_i - \overline{y})$, etc. The "direct slope" estimator implements the first line above, by applying the direct estimator formulae for y set to each component of x; obtaining an estimator $\widehat{\delta}_x$ of $E(\partial x^T / \partial x)$; and then forming

$$\widehat{d}_{\text{dir}} = \widehat{\delta}_x^{-1} \widehat{\delta}_{\text{dir}}.$$

This is a "slope" estimator because it can be rewritten in instrumental variables form; $\widehat{d}_{\text{dir}} = S_{\omega x}^{-1} S_{\omega y}$ for a certain instrument $\widehat{\omega}(x_i)$; although these details need not concern us here (c.f. Stoker (1991a)).

We discuss the distribution theory for these estimators below. However, to simplify understanding the basic issues, we also introduce two other estimators, for which the technical details are held to a minimum. The "density weighted" average derivative estimator of Powell, Stock and Stoker (1989) is the sample analog of $\delta_f = E(fm') = -2E(f'y)$, or

$$\widehat{\delta}_f = -2N^{-1} \sum_{i=1}^{N} \widehat{f}'(x_i) y_i.$$

The technical advantages arise from the fact that $\widehat{\delta}_f$ is a simple linear average of kernel estimators $\widehat{f}'(x_i)$, and there is no division by the estimated density, so that trimming for low estimated density is not necessary for the technical analysis. As indicated above, the "density weighted" estimator retains the proportionality to index model coefficients, although we have $E(f) \neq 1$. This scaling is corrected in the "density weighted slope" estimator

$$\widehat{d}_f = S_{f'x}^{-1} S_{f'y}$$

that uses $\widehat{f}'(x_i)$ as an instrumental variable. $\widehat{d}_f$ estimates $E(fm')/E(f)$, and so has the proper scaling for linear model coefficients.

B. Asymptotic Distribution Theory

We now point out how each of the above estimators is a $\sqrt{N}$ consistent, asymptotically normal estimator of the appropriate average derivative. The distribution theory is based on the same general ideas; although we will focus on the "density weighted" estimator for simplicity. Recall at the outset the main issues raised by the distributional theory. Each estimator has nonparametric ingredients that converge at rates slower than $\sqrt{N}$ to their limits. The $\sqrt{N}$ consistency alluded to above implies that a faster rate is possible when the nonparametric estimators are combined as above. The precision properties implied by this rate are comparable to those available for estimators when the model is specified correctly and fully parameterized. Our development below is focused on explaining how the "curse of dimensionality" is deflected.

The relevant conditions for the $\sqrt{N}$ asymptotic normality of the "density weighted" average derivative estimator are given in Powell, Stock and Stoker (1989), as is a formal proof. These conditions include that i) the density $f(x)$ vanishes on the boundary of its support, ii) $f(x)$ is P-th order differentiable, where P in an integer, $P \geq (k+3)/2$ and iii) the kernel $K(u)$ symmetric function of order P. Then under the bandwidth conditions $Nh^{2P} \to 0$ and $Nh^{k+2} \to \infty$, the the asymptotic distribution of the density weighted estimator

$$\widehat{\delta}_f = -2N^{-1} \sum_{i=1}^{N} \widehat{f}'(x_i) y_i$$

is given from the influence representation

$$\sqrt{N}(\widehat{\delta}_f - \delta_f) = N^{-1/2} \sum_{i=1}^{N} r_f(y_i, x_i) + o_p(1)$$

where

$$r_f(y,x) \equiv 2f(x)m'(x) - \delta_f - [y - m(x)][2f'(x)].$$

Thus, since $E(r_f) = 0$, if the variance of r_f exists, we have $\sqrt{N}(\widehat{\delta}_f - \delta_f) \to \mathcal{N}(0, \Sigma_f)$, where $\Sigma_f = E(r_f r_f^T) = \text{cov}(r_f)$. Thus, while each of the nonparametric components $\widehat{f}'(x)$ is incapable of converging to $f'(x)$ at rate $\sqrt{N}$, the estimator $\widehat{\delta}_f$ does achieve the rate typical for parametric problems.

We begin the demonstration by separating the issues into asymptotic bias and variance, which work somewhat differently;

$$\sqrt{N}(\widehat{\delta}_f - \delta_f) = \sqrt{N}[E(\widehat{\delta}_f) - \delta_f] + \sqrt{N}[\widehat{\delta}_f - E(\delta_f)]$$

where the first term represents bias and the second term represents sampling variation. The bias term is analyzed first, as

$$\begin{aligned} E(\widehat{\delta}_f) - \delta_f &= -2N^{-1} \sum E(\widehat{f}'(x_i) y_i) - 2E(f'y) \\ &= -2N^{-1} \sum E[\widehat{f}'(x_i) y_i - f'(x_i) y_i] \\ &= -2E\{[\widehat{f}'(x_i) - f'(x_i)] y_i\} \end{aligned}$$

or -2 times the average pointwise bias of $\widehat{f}'(x_i)y_i$. For the nonparametric estimator $\widehat{f}'(x)$, we have discussed how the bias of the estimator is $O(h^P)$, when the kernel function K is of order P. By a Taylor series expansion analogous to that for the pointwise estimator $\widehat{f}'(x)$, we can show that the average bias is likewise $O(h^P)$. Consequently,

$$\sqrt{N}[E(\widehat{\delta}_f) - \delta_f] = O(\sqrt{N} h^P).$$

This asymptotic bias will vanish if $Nh^{2P} \to 0$, which is our bias condition above. This is associated with the bias of $\widehat{f}'(x)$ vanishing more rapidly than for optimal pointwise approximation of $f'(x)$. In this context, "asymptotic undersmoothing" is required for the convergence of $\widehat{\sigma}_f$ to δ_f at rate $\sqrt{N}$. Moreover, since optimal pointwise approximation balances pointwise bias and variance, the pointwise variance must shrink at a suboptimal rate. Consequently, if the result above is true, something fundamentally different must be happening with sampling variation of $\widehat{\delta}_f$ relative to that of $\widehat{f}'(x)$.

Consider Figure 3.4, which similar to Figure 1.1 with an additional point added. As the bandwidth h is shrunk, the number of terms in the average comprising $\widehat{f}'(x)$ is $O(Nh^k)$, and since local differencing is involve, the variance of $\widehat{f}'(x_i)$ declines with sample size N and bandwidth h as $O[(Nh^{k+2})^{-1}]$. However, for $\widehat{\delta}_f$, when the sample size is increased, we also add terms $\widehat{f}'(x_j)y_j$ for the new sample points. So while the individual approximation regions are shrinking, the number of terms in $\widehat{\delta}_f$ is a increasing, and their approximation regions overlap. Consequently, an analysis of the sampling variation of $\widehat{\delta}_f$ requires a method for accounting for the overlaps explicitly, or take account of the relative positions of (x_i, x_j) pairs. Doing this yields the theorem, which gives a case when averaging nonparametric estimators speeds up the rate of convergence to $\sqrt{N}$.

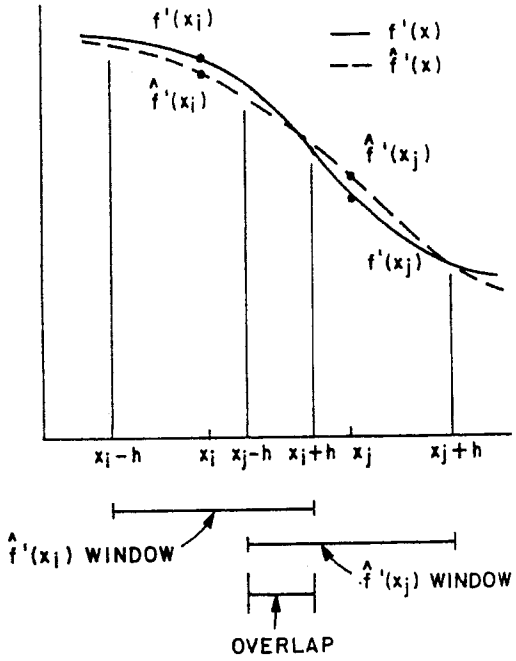


Figure 3.4
Overlaps in Double Averaging

We develop this theory in some detail, through the additive approximation structure of U-statistics. First, write $\widehat{\delta}_f$ explicitly as

$$\begin{aligned}\widehat{\delta}_f &= -2N^{-1}\sum_i \widehat{f}'(x_i)y_i \\ &= -2N^{-2}\sum_i\sum_j h^{-k-1}\mathcal{K}'\left[\frac{x_i - x_j}{h}\right]y_i \quad i \neq j \\ &= N^{-2}\sum_i\sum_j p_N[z_i, z_j] \qquad i \neq j\end{aligned}$$

where we denote $z_i = (x_i, y_i)$, and $p_N[z_i, z_j] = -2h^{-k-1}\mathcal{K}'[(x_i - x_j)/h]y_i$, and the dependence of p on h is depicted through the N subscript. Note that symmetry of $\mathcal{K}$ implies $\mathcal{K}'(0) = 0$, which yields the "$i \neq j$" indexing on the double sum.

As such, $\widehat{\delta}_f$ is a fairly complicated function of pairs of data points; summarized by $p_N(z_i, z_j)$. The influence representation of the theorem gives a (strictly

additive) sample average, to which the central limit theory is directly applicable. Consequently, we need to establish an additive approximation of $\widehat{\delta}_f$. To get a clue on how to do this, recall the following property of additive regressions:

Lemma (Lehmann(1975)): *Let $z_1, \ldots, z_N$ be independent random variables, with F_i the c.d.f. of z_i . Let $S(z_1, \ldots, z_N)$ be such that $E(S) = 0$. Then the additive function*

$$S_a = R_1(z_1) + \cdots + R_N(z_N)$$

that minimizes $E(S - S_a)^2$ has

$$R_i(z) = E(S \mid Z = z_i).$$

If the z's are identically distributed, with $F_i \equiv F$, then $R_i \equiv R = E(S \mid Z = z)$.

We apply this lemma by noting that $\widehat{\delta}_f$ has U-statistic structure (c.f. Serfling (1980), Lehmann (1975) among others). This type of structure is analyzed as follows.

A (bivariate) U-statistic is a double sum of the form

$$U = N^{-1}(N-1)^{-1}\sum_{i \neq j} p(z_i, z_j).$$

Such a statistic is "symmetrized" by replacing $p(z_i, z_j)$ by

$$p^*(z_i, z_j) = (1/2)[p(z_i, z_j) + p(z_j, z_i)],$$

and simplifying as

$$\begin{aligned}U &= N^{-1}(N-1)^{-1}\sum_{i<j}[p(z_i, z_j) + p(z_i, z_j)] \\ &= 2N^{-1}(N-1)^{-1}\sum_{i<j} p^*(z_i, z_j) \\ &= \binom{N}{2}^{-1}\sum_{i<j} p^*(z_i, z_j).\end{aligned}$$

Denote $E(U) = E(p) = \theta$, and note that

$$\begin{aligned}U - \theta &= \binom{N}{2}^{-1}\sum_{i<j}[p^*(z_i, z_j) - \theta] \\ &= \binom{N}{2}^{-1}\sum_{i<j} s(z_i, z_j)\end{aligned}$$

where $E(s) = 0$.

The "projection" $\tilde{U}$ of U is the additive approximation of $U - \theta$ in line with the above lemma, found from the additive approximation of $s(z_i, z_j)$. In particular, define

$$\tilde{r}(z_i) = E[s(z_i, z_j) \mid z_i] = \{E[p^*(z_i, z_j) \mid z_i] - \theta\},$$

so that the best approximation of $s(z_i, z_j)$ is $\tilde{r}(z_i) + \tilde{r}(z_j)$, and the best additive approximation of $U - \theta$ is

$$\begin{aligned}\tilde{U} &= \binom{N}{2}^{-1} \sum_{i<j} [\tilde{r}(z_j)] \\ &= N^{-1} \sum_i 2\tilde{r}(z_i).\end{aligned}$$

Central limit theory can be applied to yield the asymptotic normality of $\tilde{U}$.

To consider the quality of the additive approximation, consider the squared expectation of $\sqrt{N}(U - \theta - \tilde{U})$, or $NE|U - \theta - \tilde{U}|^2$. By construction,

$$U - \theta - \tilde{U} = \binom{N}{2}^{-1} \sum_{i<j} q(z_i, z_j)$$

where $q(z_i, z_j) = p^*(z_i, z_j) - \tilde{r}(z_i) - \tilde{r}(z_j) - \theta$ is a "within" deviation; obeying $E(q) = 0$, $E[q(z_i, z_j)q(z_{i'}, z_{j'})] = 0$ unless $(i, j) = (i', j')$, and variance of q is $E(|q|^2) \leq E|p^*|^2 \leq E|p|^2$. Thus

$$\begin{aligned}NE|U - \theta - \tilde{U}|^2 &= N\binom{N}{2}^{-2} \sum_{i<j} E(|q(z_i, z_j)|^2) \\ &= N\binom{N}{2}^{-2} \binom{N}{2} E(|q(z_i, z_j)|^2) \\ &= 2N/N(N-1)E(|q|^2) \\ &\leq 2/(N-1)E|p|^2\end{aligned}$$

As such, if the variance of p is exists, then we have that

$$\sqrt{N}(U - \theta - \tilde{U}) = o_p(1)$$

so that if $\tilde{U}$ is asymptotically normal, then so is U.

This demonstration of the asymptotic normality of U statistics is due to Hoeffding (1948). For our purposes, we note some variations. If $p(z_i, z_j)$ varies with N, as in $p_N(z_i, z_j)$, then we need only add N subscripts to θ, $\tilde{r}$, $\tilde{U}$, s, and q, and the equivalence follows if $E|p_N|^2$ is bounded. Moreover, from the last line of the derivation, we see that the equivalence holds if $E[|p_N|^2]$ grows with N, so long as $E[|p_N|^2] = o(N)$. This modification is Lemma 3.2 of Powell, Stock and Stoker (1989).

We can now relate this back to $\widehat{\delta}_f$. We have that

$$\widehat{\delta}_f = \frac{N-1}{N} N^{-1}(N-1)^{-1} \sum_{i \neq j} p_N(z_i, z_j)$$

where $p_N(z_i, z_j) = -2h^{-k-1}\mathcal{K}'[(x_i - x_j)/h]y_i$. If we denote $M(x) = E(y^2 \mid x)$, then the second moment $E|p_N|^2$ is

$$\begin{aligned}E|p_N|^2 &= 4h^{-2k-2} \int \left|\mathcal{K}'\left[\frac{(x_i - x_j)}{h}\right]\right|^2 M(x_i)f(x_i)f(x_j)dx_i dx_j \\ &= 4h^{-k-2} \int |\mathcal{K}'[u]|^2 M(x_i)f(x_i)f(x_i + hu)dx_i du \\ &= O(h^{-k-2}) = O(NN^{-1}h^{-k-2}) = o(N)\end{aligned}$$

provided $Nh^{k+2} \to \infty$, or when the pointwise variance of $\widehat{f}'(x_i)$ vanishes. Thus, we have shown that

$$\sqrt{N}[\widehat{\delta}_f - E(\widehat{\delta}_f)] = N^{-1/2} \sum r_N(z_i) + o_p(1)$$

where $r_N(z_i) = E(p_N(z_i, z_j) + p_N(z_j, z_i) \mid z_i) - 2E(p_N)$. The analysis of $\widehat{\delta}_f$ is completed by showing that $r_N(z)$ approximates $r_f(z)$ given above, by directly showing that $E(r_N - r_f)^2 = o(1)$.

Our intention was to elaborate on how $\sqrt{N}$consistency depends on undersmoothing to control bias, and double averaging to deliver the variance properties. We further have characterized the variance of $\widehat{\delta}_f$ in a way that permits simple consistent estimation, for the purposes of statistical inference. In particular, we need to measure the variance of $r_N(z_i)$, where $r_N(z_i) = E(p_N(z_i, z_j) + p_N(z_j, z_i) \mid z_i) - 2E(p_N)$. Since p_N is the (known and computable) function above, we can estimate the leading term of r_N by averaging p_N , holding one argument constant, as in

$$\tilde{r}_N(z_i) = N^{-1} \sum_j [p_N(z_i, z_j) + p_N(z_j, z_i)].$$

We estimate the variance Σ_f of $\sqrt{N}(\widehat{\delta}_f - \delta_f)$ by the sample covariance of $\widehat{\Sigma}_f$ of $\widehat{r}_N(z_i)$, and the variance of $\widehat{\delta}_f$ by $N^{-1}\widehat{\Sigma}_f$. This method of variance estimation is directly connected to the U-statistic structure of $\widehat{\delta}_f$, and thus may more accurately reflect the variation than using an alterative or unrelated measure of the influence term r_f .

The analysis of the remaining (unweighted) average derivative estimators is complicated by two features, and main technical material in Härdle and Stoker (1989) is devoted to dealing with them. Consider the indirect estimator $\widehat{\delta}_{\text{ind}} = N^{-1}\sum[-\widehat{f}'(x_i)/\widehat{f}(x_i)]y_i\widehat{I}_i$. First, it is an average of nonlinear combinations of the nonparametric estimators $\widehat{f}(x)$ and $\widehat{f}'(x)$ and second, it uses estimated trimming to control for erratic behavior induced by division by $\widehat{f}(x)$. The first feature requires "linearizing" $\widehat{\delta}_{\text{ind}}$; showing that $\widehat{\delta}_{\text{ind}}$ is equivalent to separate weighted averages of $\widehat{f}'(x_i)$ and $\widehat{f}(x_i)$; and the second feature must be examined directly.[4] Härdle and Stoker (1989) carry this out in showing that under various conditions, including those on bandwidth h and trimming bound b, then

$$\sqrt{N}(\widehat{\delta}_{\text{ind}} - \delta) = N^{-1/2}\sum_{i=1}^{N} r(y_i, x_i) + o_p(1)$$

where

$$r(y, x) \equiv m'(x) - \delta + [y - m(x)]\ell(x),$$

so that as before, $\sqrt{N}(\widehat{\delta}_{\text{ind}} - \delta) \to \mathcal{N}(0, \Sigma)$, where $\Sigma = E(rr^T) = \text{cov}(r)$.

In Stoker (1991a), the same influence representation was shown for the direct, indirect slope, and direct slope estimators. Consequently, $\widehat{\delta}_{\text{ind}}$, $\widehat{\delta}_{\text{dir}}$, $\widehat{d}_{\text{ind}}$ and $\widehat{d}_{\text{dir}}$ are first-order ($\sqrt{N}$)equivalent estimators of $\delta = E(m')$. The demonstration for the direct estimator is tedious, but based broadly on the same principles as above. The equivalence of the indirect slope estimator rests on the fact that $S_{\ell x} - Id = o_p(1/\sqrt{N})$, or that the leading matrix of the slope estimators converge to their limits faster than rate $\sqrt{N}$. See the above reference for the derivations.

Finally, for all of these estimators, simple estimators of asymptotic variance can be derived from mimicking their U-statistic structure. These estimators reflect the variation in the "linearized" versions of the basic statistics, in an analogous manner to that described above.

[4] An error in the trimming step of the proof in Härdle and Stoker (1989) was pointed out to me by J. Powell, with a corrected argument given by his student K. Jeong.

On the basis of first-order distribution theory, we are left with little reason to choose between the direct, indirect or the slope estimators. This situation is somewhat surprising, as the various procedures are based on approximating different functions; namely density or regression functions. In Lecture 4 we discuss some recent work that explains this phenomena, by pointing out how the method of nonparametric approximation can be irrelevant for the asymptotic theory of semiparametric estimators.

On more practical grounds, two further concerns favor the slope estimators, which I recommend for all empirical applications, including those discussed in these lectures. First, when the true model is linear, the slope estimators are unbiased for the linear model coefficients conditional on the x values, and the same cannot be said for the direct or indirect estimators. Second, on more basic grounds, the slope estimators require less precision from the nonparametric components, because of their ratio form. In particular, if the estimated score $\widehat{\ell}(x)$ were too small over the sample, then its level would affect the indirect estimator but not the indirect slope estimator. We explore this further in the Lecture 5, in our discussion of small sample smoothing bias problems.

LECTURE 4

ECONOMETRIC TOOLS: MODEL AND ESTIMATOR ASSESSMENT

To this point we have discussed many aspects of semiparametric estimation, including a brief example of how semiparametric methods can be used to guide predictor variable and model selection. We have not considered more formal aspects of testing for model specification, aside from carrying out inference on the values of average derivatives. In this lecture, we address the specification testing issues directly, by assessing partial index structure in an empirical example. We utilize a natural "regression style" testing approach, which likely has wide application beyond assessing index structure. Further, the distribution theory of our tests illustrates a standard singularity problem of semiparametric specification tests. We close this lecture by discussing some recent theoretical work on semiparametric variance issues, that sheds light on various nonparametric approximation issues raised here and in Lecture 3.

I. Semiparametric Specification Testing

A requirement of good empirical work is the application of methods to check the statistical adequacy of an estimated model. A large variety of such specification tests have been developed, ranging from methods to ask very specific questions (is the coefficient of a variable equal to 1) to checking broad model implications (is there any evidence of endogeneity of the predictor variables).[1] Semiparametric methods should be particularly well designed for model testing, in terms of parameter values as well as the specifications of basic functions. In particular, suppose that some aspect of a model, say a parameter value or overall goodness-of-fit, were inordinately affected by assuming that the model is linear, or by assuming a normal distribution for unobserved elements. A semiparametric method that permits a more general regression, or a more general distribution of unobserved elements, should reveal important differences in such sensitive aspects of the model. In reflection, it seems difficult to conceive of a setting where flexible measurement methods would not be helpful in checking model specification. In other words, the range of diagnostic applications of semiparametric methods is enormous, and should represent an important role for them in future empirical research.

Instead of cataloging the possibilities for specification tests based on semiparametric methods, we will just consider a simple empirical application. This example concerns choosing an adequate functional form among various semiparametric (reduced dimension) alternatives. More specifically, we compare partial index models to general regression, in search of the simplest (lowest dimension) model that is statistically adequate. This contrasts with the example of the last lecture where we maintained the single index model, and looked for additional refinements (dropping a variable, and considering the probit specification).

We now turn to some salient features of an analysis of data on house prices in Boston described in Rodriguez and Stoker (1992). This data, used to ascribe dollar values of impacts of changes in air pollution levels on housing prices, was first analyzed by Harrison and Rubinfeld (1978a,b), and the housing price equation was studied at length in the book *Regression Diagnostics*, by Belsley, Kuh and Welsch. Our application studies the use of index models to characterize the housing price

[1] Ruud (1984) is a good survey of traditional specification tests in econometrics. Pagan and Vella (1989) give a recent survey, including some uses of nonparametric and semiparametric methods.

data, as well as testing the adequacy of the linear equations used previously. Our initial expectation was that this testing endeavor would confirm the well studied linear equation, but we find substantive nonlinearity, that is naturally summarized in a partial index framework. This finding is interesting in and of itself, because it states that an extremely careful analysis limited to linear regression can miss substantive features of multivariate data.

A. Measuring Hedonic Price Equations with the Boston Housing Data

The basic approach regards each house as a bundle of characteristics (with the level of air pollution as one characteristic), and the price of each house as reflective of the value of its characteristics. Denote the vector of housing attributes as (a, h), where a is the air pollution level, and denote the (required annual expenditure) price of a house with attributes (a, h) as $p(a, h)$. The standard approach as adopted by Harrison–Rubinfeld is to regard the supply of houses as fixed, and the hedonic function $p(a, h)$ as exogenously determined, and we adopt that approach in the empirical analysis that follows. See Tinbergen (1956) for an ingenious depiction of how hedonic price functions can be determined endogenously, and Rosen (1974) for an analysis of the issues raised by a competitive supply of new houses. Harrison–Rubinfeld analyze the price function using linear regression on the logarithm of house prices; we study these data using semiparametric index model methods. We reorder the characteristics (a, h) as (x, ε), into observed characteristics x and unobserved characteristics ε, and our basic statistical model is based on housing prices determined as $p = p(x, \varepsilon)$.

This application is a useful test case for index models because of the previous work on characterizing log-linear price equations with this data, as well as similar types of data. Because of the studies of Harrison and Rubinfeld, and Belsley, Kuh and Welsch, we define the forms of the predictor variables in the same fashion. In particular, y_i will denote the log of price of house i, and x_i denotes the vector of nine predictor variables that Harrison and Rubinfeld found to be statistically significant in their analysis. The data consists of 506 observations on the variables depicted in Table 4.1. As mentioned above, the earlier work produced a linear equation between y and x, with house prices in log form and the predictors transformed as indicated above.

Before discussing our results explicitly, it is useful to consider the implications of standard forms of the hedonic price equation $p = p(a, h) = p(x, \varepsilon)$. Consider a linear equation in price levels

$$p = \alpha + x_i^T \beta + \varepsilon_i.$$

TABLE 4.1:
VARIABLE SPECIFICATION IN THE BOSTON HOUSING DATA

$y = \ln p$	LMV	log of home value
x_1	NOXSQ	nitrogen oxide concentration
x_2	CRIM	crime rate
x_3	RMSQ	number of rooms squared
x_4	DIS	distance to employment centers
x_5	RAD	accessibility to radial highways
x_6	TAX	tax rate
x_7	PTRATIO	pupil teacher ratio
x_8	B	proportion of black residents in neighborhood
x_9	LSTAT	log of proportion of residents of lower status

In this case, β is interpreted as a vector of hedonic prices, as the change in overall housing price given a change in the level of the characteristic vector x. This model is dictated if arbitrage exists under competition and if houses can be easily repackaged (or their characteristics unbundled, with an effective market for each characteristic). As such, the relative impact of characteristics x_j and x_k is measured by β_j/β_k, or the ratio of their (competitively determined) prices. Here $\alpha + \varepsilon_i$ represents the value of all unobserved characteristics.

But clearly for houses, it is not natural to expect this kind of unbundling for all conceivable characteristics. In particular, different locations of houses are associated with persistent differences in prices, even after controlling for the few variables above, and the specific location is hard to unbundle from the house that

sits on it. As such, a linear equation such as that above implies a very strong restriction, namely that for the observed data, the impact of specific locations ε_i is additive in housing prices. Changing from one location value to another only alters the level of prices, with the spread between big and small houses held constant.

Traditional empirical analysis using hedonic price equations, as in previous analysis of our data, finds empirical support for linear equations in log prices, or

$$y \equiv \ln p = \alpha + x^T \beta + \varepsilon.$$

Here, the coefficients are interpreted as the proportional changes in prices associated with changes in characteristics, holding specific location features constant. The proportional impact of x_j relative to x_k is summarized compactly as β_j/β_k. Here the model dictates that when the location value ε_i is changed, the same proportional configuration of housing prices is observed along the lines of the observed characteristics.

A fully nonparametric analysis of the hedonic price equation dispenses with all these restrictions. In particular, the basic "model" in that case is

$$y = \ln p(x, \varepsilon)$$

and a nonparametric regression analysis utilizes estimators of the mean log price given the characteristics x, or

$$E(y \mid x) = E[\ln p(x, \varepsilon) \mid x] \equiv m(x).$$

For a given configuration of characteristics x, the percentage marginal effects are then $\partial m/\partial x$. We will use a kernel estimator $\widehat{m}(x)$ to estimate this regression function (choosing smoothing parameters again by generalized cross validation). Because $\widehat{m}(x)$ is a flexibly estimated nine-dimensional function, graphical interpretation of the estimated effects $\partial \widehat{m}/\partial x$ is somewhat difficult. Aside from their use in the estimation of average derivatives, we essentially retain $\widehat{m}(x)$ only as the base case for our specification tests.

We structure our analysis by using index models, that retain the summary of relative effects through coefficients, as exhibited by log-linear equations. The single index model is an implication of a log housing price equation of the form

$$y = g(x^T \beta, \varepsilon)$$

where the density of ε could vary with $x^T\beta$. This model states that similar proportional configurations of housing values are observed for areas with the same specific location value ε, and that given ε, the proportional effect of x_j relative to x_k is measured as $[\partial y/\partial x_j]/[\partial y/\partial x_k] = \beta_j/\beta_k$. We implement this model empirically in its single index form

$$E(y \mid x) = G_1(x^T\delta)$$

where we have chosen the scaling to set $\beta = \delta = E(m')$, the average derivative vector. The (log) linear model coincides with a linear G_1 function.

We also consider partial index generalizations of the housing price equation. For instance, suppose that the proportionality of marginal effects was deemed unreasonable for variations in x_1, the pollution variable. This feature is relaxed for x_1 in the partial index formulation

$$E(y \mid x) = G_2(x_1, x_{-1}^T\delta_{-1})$$

where $x_{-1} = (x_2, \ldots, x_9)$ is the vector of all characteristics except for x_1. This is a natural generalization in view of our specific interest in the marginal pollution effects. However, we consider each variable in turn as a candidate for flexible treatment as with x_1 above. Likewise, we can generalize further in a couple ways — we could allow two or more variables a flexible role, or consider a function based on more than one linear index (thereby guaranteeing proportionality of effects among specific subsets of variables), for instance. The partial index framework provides us with a sequence of less restrictive semiparametric regression models for carrying out the specification tests.

B. Index Model Estimates, Testing Results and Graphical Analysis

First, we compare the average derivative estimates (ADE) to the OLS coefficient estimates of Belsley, Kuh and Welsch. These estimates are contained in Table 4.2. As before, details on estimation of the ADE estimates are discussed in the Appendix.

The average derivative estimates are quite reasonable, with only the "Black Proportion" coefficient (B) differing substantially. The OLS coefficient of B is somewhat counter intuitive, with the ADE estimate more reasonable. This says that forcing the log housing price equation to be linear induces a positive "Black

effect,"[2] that in reality could be some kind of nonlinearity. While the bandwidth value for the average derivatives was chosen by GCV, none of the values were sensitive to the bandwidth, save that of x_1 = NOXSQ, which varied considerably over different bandwidth values. While certain of the average derivative estimates are not significant, we retain all the variables for the analysis of functional form below.

TABLE 4.2:
COEFFICIENT ESTIMATES
FOR THE HOUSING PRICE EQUATION

$y = \ln p$	LMV	ADE	OLS
x_1	NOXSQ	−.0034	−.0060
		(.0035)	(.0011)
x_2	CRIM	−.0256	−.0120
		(.0056)	(.0012)
x_3	RMSQ	.0106	.0068
		(.0025)	(.0012)
x_4	DIS	−.0746	−.1995
		(.0504)	(.0265)
x_5	RAD	.0669	.0977
		(.0468)	(.0183)
x_6	TAX	−.0009	−.00045
		(.0003)	(.00011)
x_7	PTRATIO	−.0175	−.0320
		(.0152)	(.0047)
x_8	B	−.0526	.3770
		(7.514)	(.1033)
x_9	LSTAT	−.2583	−.3650
		(.0370)	(.0225)

(Standard Errors in Parentheses)

[2] While not reported here, robust estimators of the linear model coefficients likewise show a substantial positive effect for B. For instance, Belsley, Kuh and Welsch report estimates based on a method of Huber (method 2) that exhibit this positive effect, even more strongly than the OLS estimates.

Nonlinearity in the relationship is evidenced by our estimates of the index regression $E(y \mid x) = G_1(x^T\delta)$. To see this, we plot the estimate $\widehat{G}_1$ of G_1 in Figure 4.1 (over-smoothing slightly for interpretability), as well as the density of the index $x^T\delta$. It shows clearly that housing prices move roughly linearly with the index except for in the mid-range of index values, where they do not vary appreciably with the index values. Our interpretation of the odd "Black effect" evidenced by OLS is that it forced together the two lines for low and high index values.

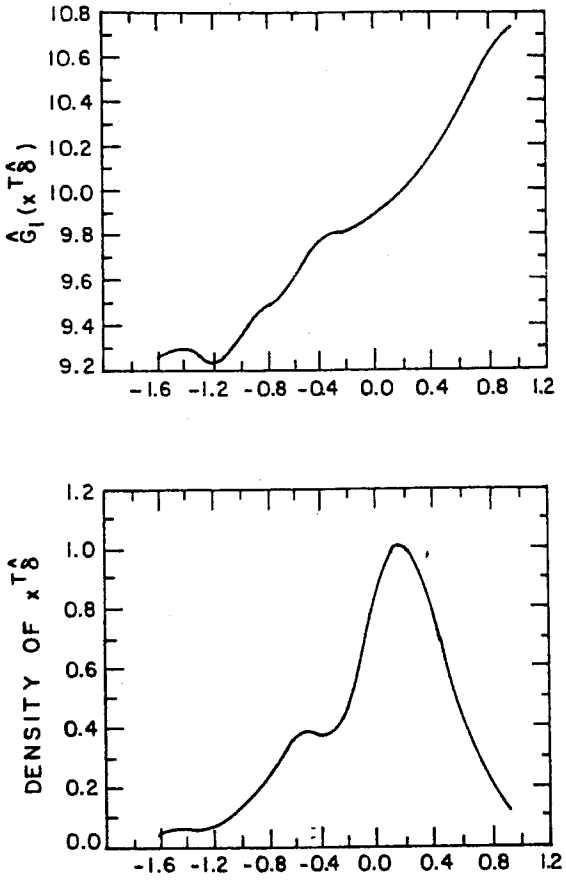


Figure 4.1
Single Index Function for Housing Data

In order to carry out tests of functional structure, we employ a simple regression test. We compute the test statistics as follows: let $\widehat{y}_{ri}$ be the fitted value from the restricted model, and $\widehat{y}_{gi}$ be the fitted value from the general model. Estimate the coefficient $\widehat{\gamma}$ of the linear regression

$$y_i - \widehat{y}_{ri} = \widehat{\alpha} + \widehat{\gamma}\,\widehat{y}_{gi} + \widehat{u} \tag{4.1}$$

by ordinary least squares. Rejection of the restricted model versus the general model is based on the t-value of this coefficient; if that t value is large, there is evidence that the specific model fails to account for as much systematic structure as the general model. For instance, to test the index model $E(y \mid x) = G(x^T\delta)$ versus a general regression $E(y \mid x) = m(x)$, we form the restricted fitted value as $\widehat{y}_{ri} = \widehat{G}(x_i^T\widehat{\delta})$ and the residual as $y_i - \widehat{y}_{ri}$, which is regressed against the general fitted value $\widehat{y}_{gi} = \widehat{m}(x_i)$ to estimate the coefficient $\widehat{\gamma}$. In the next section we show how the true coefficient value is 0 when the restricted model is correct, and positive when the restricted model is false, for our applications of the tests. We also discuss the asymptotic distribution theory of $\widehat{\gamma}$, and how we estimate its standard error for use in the "t-value" above.

As indicated above, our testing strategy consists of considering partial index models of increasing generality, until a model is found that is not rejected against the general nonparametric regression.[3] In particular, we first consider the linear model, then the linear index model, then partial index models with one variable included freely, and so forth. It is useful to note the simplicity of estimation that this testing strategy permits; namely only one set of average derivatives need be computed, with the partial index variables constructed using the appropriate ADE coefficient values from Table 4.2. Therefore, to fit the partial index models with one variable treated freely involves a two dimensional nonparametric regression (the omitted variable and the partial index), and those with two variables considered freely a three dimensional regression, and so forth.

A summary of the findings is as follows. The linear model is not rejected against the single index model, so that they represent statistically equivalent depictions of the data. However, both the linear model and the single index model are

[3] We base our tests solely on the values of the t-statistics, without giving attention to the issues of setting the overall confidence level for these tests. It would be valuable to develop an appropriate Scheffe S method or use of Bonferroni intervals for these tests.

strongly rejected against the general nonparametric regression. The best partial index model with one variable treated freely is written as

$$E(y \mid x) = G_2(x_1, x_{-1}^T \delta_{-1}),$$

which treats the pollution variable $x_1 =$ NOXSQ flexibly. This model is referred to as PARTIAL1 in subsequent tables and figures. The single index model is rejected against this model, but it too is rejected against the general regression. Finally, the best partial index model that treated two variables freely is written as

$$E(y \mid x) = G_3(y \mid x_1, x_9, x_{-19}^T \delta_{-19}),$$

which permits flexible effects of the pollution variable $x_1 =$ NOXSQ and the "lower status" variable $x_9 =$ LSTAT. This model is referred to as PARTIAL2 in tables and figures. The model PARTIAL1 with x_1 treated freely is rejected against this model, and this model is not rejected against the general regression (at least at a level less than 3%). The test statistics are summarized in Table 4.3.

TABLE 4.3:
REGRESSION TESTS OF FUNCTIONAL FORM

Tests against general regression				
Restricted	Unrestricted	$\widehat{\gamma}$	t-value	Prob $[\chi^2(1) > t^2]$
LINEAR	GENERAL	.1712	3.41	.0006
INDEX	GENERAL	.2314	5.96	0.0
PARTIAL1	GENERAL	.0718	4.52	0.0
PARTIAL2	GENERAL	.0116	2.19	.0291
Partial index model tests				
Restricted	Unrestricted	$\widehat{\gamma}$	t-value	Prob $[\chi^2(1) > t^2]$
LINEAR	INDEX	.0276	.52	.602
LINEAR	PARTIAL2	.1862	4.51	0.0
INDEX	PARTIAL1	.1975	4.59	0.0
PARTIAL1	PARTIAL2	.0893	3.72	.0002

To see how the model PARTIAL1 (where $E(y \mid x) = G_2(x_1, x_{-1}^T \delta)$) improves on the index model $E(y \mid x) = G_1(x^T \delta)$, consider the plot of $\widehat{G}_2$ in Figure 4.2. Recall that if the single index model held, then parallel cross-sectional slices of the

G_2 surface should have the same shape. We do see the general downward trend of housing prices with pollution values, although for high values of the partial index $x_{-1}\widehat{\delta}_{-1}$, the marginal effect of pollution differences is somewhat erratic in the data. This may be connected to the sensitivity of the x_1 coefficient to bandwidth values noted earlier.

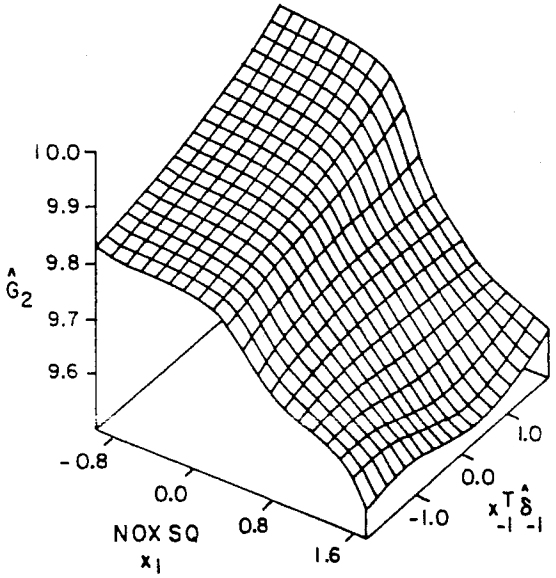


Figure 4.2
Effect of NOXSQ and Index Variable; Model PARTIAL1.

The model PARTIAL2 (where $E(y \mid x) = G_3(x_1, x_9, x_{-19}^T \delta_{-19})$) is somewhat more difficult to interpret. We include Figure 4.3a,b to summarize the separate effects of x_1, x_9 , and illustrate how they are inadequately captured by an index variable. Figure 4.3a plots the fitted value $\widehat{G}_3$ against the pollution variable and the remaining index, with a similar shape to that in Figure 4.2. Figure 4.3b plots the fitted value $\widehat{G}_3$ against the pollution and lower status variables. We see that the marginal pollution effect is flat or slightly positive for low "lower status" values, and strongly negative for high "lower status" values. Consequently, the pollution effect varies quite nonlinearly over ranges of the "lower status" variable.

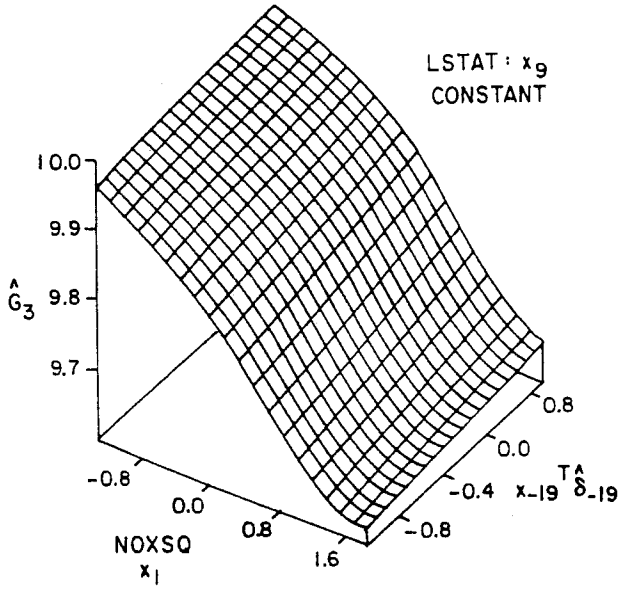


Figure 4.3a
Effect of NOXSQ and Index Variable; Model PARTIAL2

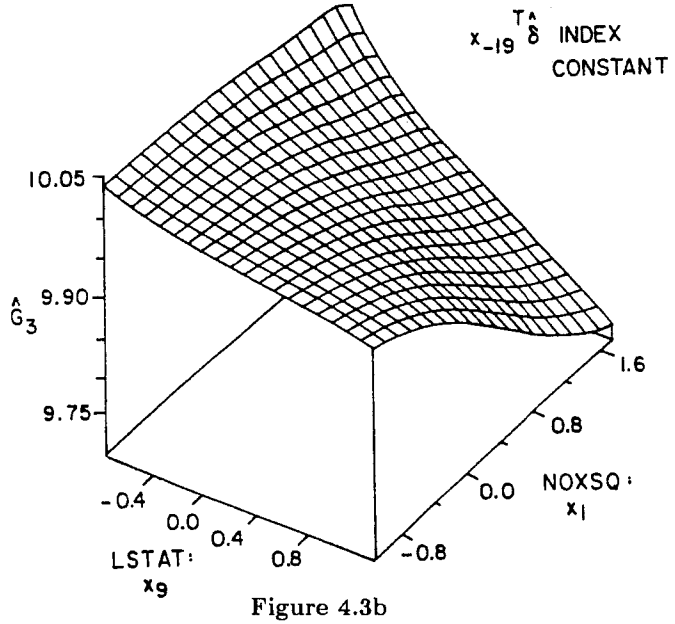


Figure 4.3b
Effect of NOXSQ and LSTAT; Model PARTIAL2.

C. "Regression" Specification Tests: Distributional Theory

It is quite natural to consider specification tests with semiparametric and nonparametric methods, because such methods permit a much wider range of structure to be captured than parametric models with a few parameters, no matter how cleverly they are designed.

Just as there are many kinds of specification tests for parametric models, semiparametric and nonparametric methods could be used in an enormous variety of ways. For instance, suppose that a finite vector of parameters θ were the object of estimation, that $\widehat{\theta}_r$ were estimates from a parametric model, and $\widehat{\theta}_g$ were estimates from a more general, semiparametric model. An immediate idea is to assess the adequacy of the parametric measure $\widehat{\theta}_r$ relative to $\widehat{\theta}_g$ by testing whether the difference $\widehat{\theta}_g - \widehat{\theta}_r$ is small. Provided that the asymptotic covariance matrix of the difference $\widehat{\theta}_g - \widehat{\theta}_r$ can be consistently estimated by $\widehat{V}_{gr}$, then an appropriate Wald statistic is

$$W = N(\widehat{\theta}_g - \widehat{\theta}_r)^T \widehat{V}_{gr}^{-1}(\widehat{\theta}_g - \widehat{\theta}_r).$$

If $\hat{\theta}_g$ and $\hat{\theta}_r$ are (jointly) $\sqrt{N}$ asymptotically normal estimators of θ, then the distribution of W can be approximated as a $\chi^2(k)$, where k is the number of components of θ. If, as would be typical, $\widehat{\theta}_r$ is an efficient estimator under the parametric model, then standard (Hausman–Wu) logic indicates that V_{gr} can be estimated by the difference $\widehat{V}_g - \widehat{V}_r \equiv \widehat{V}_{gr}$, where $\widehat{V}_g$ estimates the asymptotic variance of $\widehat{\theta}_g$ and $\widehat{V}_r$ estimates the asymptotic variance of $\widehat{\theta}_r$.[4] Many other tests based on parameter values could be proposed; for instance, $\widehat{\theta}_r$ is an M estimator, with $\mathcal{M}(\widehat{\theta}_r) = 0$ the first-order condition defining $\widehat{\theta}_r$, then a LM-type specification test is based on the size of $\mathcal{M}(\widehat{\theta}_g)$, judged by a statistic such as

$$LM = N\mathcal{M}(\widehat{\theta}_g)\left[\left(\frac{\partial \mathcal{M}}{\partial \theta}\right)^T \widehat{V}_{gr}\left(\frac{\partial \mathcal{M}}{\partial \theta}\right)\right]^{-1} \mathcal{M}(\widehat{\theta}_g)$$

or any other that utilizes $\widehat{\theta}_g$ to check a condition that must be obeyed under the parametric specification.[5]

[4] This differencing method of measuring variance can lead to practical problems in finite samples, namely negative variances of the coefficient differences. Consequently, other estimators of V_{gr} may be preferable, such as those based on direct estimation of the influence representation of $\widehat{\theta}_g - \widehat{\theta}_r$.

[5] See Ruud (1984) and Newey (1986) for many parametric tests of this kind.

Different styles of tests arise out of using nonparametric estimators explicitly in the role of alternatives, giving rise to effectively general "goodness of fit" criterion. For instance, Bierens (1990) and Lewbel (1991) use nonparametric formulations to expand the types of alternative directions in conditional moment tests, in line with the logic discussed in Newey (1986), among many others. Stoker (1989) and Samarov (1990) have proposed using average derivative estimators to assess constraints on regression derivatives; with Samarov (1990) applying this idea to the specification of "projection pursuit" style statistical models.

Above we have used a different type of specification test, which is a variation on the testing theme proposed by Wooldridge (1990). It is based on regressing restricted model residuals on general model fitted values, and works as follows. For our discussion suppose that the restricted model is an index model $E(y \mid x) = G(x^T\delta)$, and the general model is a general smooth regression $E(y \mid x) = m(x)$. All arguments are based on nested conditional expectations, and extend naturally to each of the hypothesis tests reported above. The test is based on the coefficient $\widehat{\gamma}$ of the OLS regression

$$y_i - \widehat{G}(x_i^T\widehat{\delta}) = \widehat{\alpha} + \widehat{\gamma}\widehat{m}(x_i) + \widehat{u}_i. \tag{4.2}$$

Equation (4.2) amounts to fitting an analog of the population regression equation

$$y - G(x^T\delta) = \alpha + \gamma m(x) + u \tag{4.3}$$

where the parameter γ is defined via OLS projection, as

$$\gamma = \frac{E\{[m(x) - E(m)][y - G(x^T\delta)]\}}{E[m(x) - E(m)]^2}. \tag{4.4}$$

As such, if $E(y \mid x) = G(x^T\delta)$ is the true regression model, then $y - G(x^T\delta)$ is not correlated with any function of x, or in particular, with $m(x)$, so that $\gamma = 0$. If the regression is not in single index form, then the estimate $\widehat{G}(x^T\widehat{\delta})$ estimates $E(y \mid x^T\delta)$. Consequently, by the law of iterated expectation, we have $G(x^T\delta) = E(y \mid x^T\delta) = E[m(x) \mid x^T\delta]$, so that

$$\begin{aligned} m(x) &= E(m(x) \mid x^T\delta) + [m(x) - E(m(x) \mid x^T\delta)] \\ &= G(x^T\delta) + U(x) \end{aligned}$$

where $U(x) = m(x) - E(m(x) \mid x^T\delta)$ has mean 0 conditional on $x^T\delta$. This implies that

$$\gamma = \frac{E[U(x)^2]}{E[m(x) - E(m)]^2} > 0 \tag{4.5}$$

when $m(x)$ differs from $G(x^T\delta)$ on a set of positive probability. Thus, an estimate of γ measures the percentage of (structural) variance not accounted for by the restricted model, and therefore has the ability to detect any nonparametric departures in $E(y \mid x)$ from the index model $G(x^T\delta)$. It is easy to verify the analogous properties for tests of linear or partial index models against more general alternatives.

The distribution theory for our test is somewhat complicated, and contained in Rodriguez and Stoker (1992). The basic verifications involve linearization and U-statistic structure as before, and utilize trimming (or omitting data points with low estimated density). Here we discuss the basic ideas for estimating the variance, as well as the result, for later purposes. We give the computational formulae for the test of a single index model versus general regression in the Appendix.

Our approach involves comparing $\widehat{\gamma}$ to the hypothetical estimate $\tilde{\gamma}$ that would be obtained if the parameters and functions were known. In particular, $\tilde{\gamma}$ is the OLS coefficient from computing the regression

$$y_i - G(x_i^T\delta) = \tilde{\alpha} + \tilde{\gamma}m(x_i) + u_i, \quad i = 1, \ldots, N. \tag{4.6}$$

Standard reasoning implies that

$$\sqrt{N}(\tilde{\gamma} - \gamma) = \sigma_m^{-1}N^{-1/2}\sum[m_i - E(m)]u_i + o_p(1) \tag{4.7}$$

where $\sigma_m = E[m - E(m)]^2$, and $u = y - G - \gamma[m - E(m)]$. With known functions, this relation would permit estimation of the variance of $\tilde{\gamma}$ using the standard (White) heteroskedasticity-corrected variance formula. Our approach for estimating the variance of $\widehat{\gamma}$ is to establish that

$$\sqrt{N}(\widehat{\gamma} - \tilde{\gamma}) = RA_N - LA_N + o_p(1) \tag{4.8}$$

where RA_N and LA_N are adjustments for the estimation of the unknown functions:

$$\begin{aligned} RA_N &= N^{-1/2}\sum[\widehat{m}(x_i) - m(x)][y_i - G(x_i^T\delta)] \\ LA_N &= N^{-1/2}\sum[\widehat{G}(x_i^T\widehat{\delta}) - G(x_i^T\delta)][m(x_i) - E(m)]. \end{aligned}$$

Thus, we approximate the influence term of $\widehat{\gamma}$ by adding estimates of the terms on the right-hand sides of (4.7) and (4.8). The adjustments in (4.8) can be linearized to U-statistics, and we use estimators that mimic that structure, as in our discussion of average derivative estimators of the last lecture. At any rate, if our estimate of the asymptotic variance of $\widehat{\gamma}$ is denoted $\widehat{\sigma}_\gamma$, then the "t-value" is found as

$$t = \frac{\sqrt{N}\widehat{\gamma}}{\sqrt{\widehat{\sigma}_\gamma}}.$$

The square of this statistic is compared to critical values from the $\chi^2(1)$ distribution as above. As before, we feel that the method of mimicking the U-statistic structure could more accurately reflect the variation in g than an alternative estimator of its asymptotic variance. We have spelled out this posture because for these tests, there is a problem with the asymptotic precision theory, that seems endemic to tests of this kind. In particular, under various conditions, we can show that the influence representation of $\widehat{\gamma}$ is

$$\sqrt{N}(\widehat{\gamma} - \gamma) = \frac{1}{\sigma_m} N^{-1/2} \sum r_{\gamma i} + o_p(1)$$

where

$$\begin{aligned} r_{\gamma i} &= [m(x_i) - E(m)]u_i + [m(x_i) - G(x_i^T\delta)][y_i - m(x_i)] \\ &\quad - [G(x_i^T\delta) - E(y)][y_i - G(x_i^T\delta)] - B\, r(y_i, x_i) \\ B &= E\{G'[E[(y+m)x \mid x^T\delta] - 2G\ E(x \mid x^T\delta)]\} \end{aligned}$$

and $r(y, x)$ is the influence function of $\widehat{\delta}$ given in the last lecture (with the matrix B giving the correction for using the estimated coefficients $\widehat{\delta}$). As before, this representation would typically imply that $\sqrt{N}(\widehat{\gamma} - \gamma) \to \mathcal{N}(0, \sigma_\gamma)$, where σ_γ is the variance of the influence term r_γ.

The problem is that under the null hypothesis that $m(x) = G(x^T\gamma)$, it is easy to verify that $B = 0$ and $r_{\gamma_i} = 0$ for all i, but that these terms are nonzero when $m(x) \neq G(x^T\delta)$. Therefore, the asymptotic distribution exhibits a singularity under the null hypothesis, which could throw into question our use of the normal distribution for the test statistics. Notice how peculiar this finding is; if the functions were known exactly, then $\tilde{\gamma}$ of (4.6, 7) has asymptotic variance under the null, but when the functions are estimated, the adjustment terms serve to cancel out this variance.

It is difficult to describe the reasons for this phenomena, but some intuition may be available from considering the original problem attacked by Wooldridge (1990). There the restricted model is linear; $E(y \mid x) = \alpha + x^T\beta$, and the general model is a general regression $E(y \mid x) = m(x)$. Suppose that estimates $\widehat{\alpha}$ and $\widehat{\beta}$ are computed by OLS regression of y_i on x_i, and that the general function $m(x)$ is estimated by the fitted values $\widehat{m}(x_i)$ of regressing y_i on x_i and powers of x_i of several orders. Under the null hypothesis of linearity, we certainly have that $y - \alpha - x^T\beta$ is uncorrelated with $m(x)$, but least squares estimation invokes this lack of correlation in stronger form. In particular, $\widehat{\alpha}$ and $\widehat{\beta}$ are computed so that $\sum x_i(y_i - \widehat{\alpha} - x_i^T\widehat{\beta}) = 0$, or that the sample correlation between x_i and the linear model residuals is exactly 0. Consequently, it is natural to suspect that the sample correlation between $y_i - \widehat{\alpha} - x_i^T\widehat{\beta}$ and $\widehat{m}(x_i)$ will likewise be close to zero, converging at a faster rate than $\sqrt{N}$ when the true model is linear. For comparing (4.2) and (4.6), it is clear that when $y = G(x^T\delta) + \varepsilon$ is the true model, then $\varepsilon = y - G(x^T\delta)$ induces asymptotic variance in the coefficient $\tilde{\gamma}$. However, the logic above suggests that the estimation of G and δ by $\widehat{G}$ and $\widehat{\delta}$ may absorb much of the variation of ε, so that asymptotically $y - \widehat{G}(x^T\delta)$ varies less than $y - G(x^T\delta)$; with no asymptotic variance arising for $\widehat{\gamma}$.

Our approach to the estimation of variance is to mimic the U-statistic structure directly, relying on the idea that smoothing will involve some errors in small samples that do not arise in the asymptotic distribution above. Further, various artificial methods can be used to introduce asymptotic variance and assure asymptotic normality of the coefficient estimator $\widehat{\gamma}$. One method would be to add independent random noise to the restricted model residuals; namely draw η_i for each i, independently of x_i , and perform the regression above with $y_i - \widehat{G}(x_i^T\widehat{\delta}) + \eta_i$ as the dependent variable. Our method of measuring the variance of $\widehat{\gamma}$ would be consistent in this case as well. However, we have refrained from this since the variance of η_i could be chosen to be extremely small, and therefore one would not expect that this method would make any difference to the testing results. A theoretical justification for our method may be available along the lines of Wooldridge (1990), who argues that using an asymptotically poor estimate $\widehat{m}(x)$ of $m(x)$ can create asymptotic variance in test statistics of this type. Such an estimate is theoretically "poor," in the sense that it must converge very slowly to the true function. Other methods could include splitting the sample, say estimating with half the data and testing specification with the other half, which might be useful

with a huge number of observations. Nevertheless, the best way to carry out tests of this type is a topic that merits future research.

II. Asymptotic Variance Issues

For semiparametric problems, there are many features of asymptotic distribution approximations that at first glance appear quite counter intuitive. One type is discussed above, where the use of semiparametric and nonparametric estimators causes singularities in the asymptotic distribution of regression test statistics. Another concerns the finding discussed in Lecture 3, that the same asymptotic distribution arises for different average derivative estimators, regardless of whether a regression function or a density function is measured in their formulation. Other examples where the method of nonparametric estimation appears irrelevant include Robinson's (1988b) use of kernel estimators and Newey's (1988) use of polynomial estimators in partially linear models.

A further issue of this kind concerns situations where the variance of a "plug in" semiparametric estimator is not affected at all by the estimation of unknown functions. To motivate this, recall the problem of estimating the coefficients of a linear model when the disturbances are heteroskedastic. The efficient weighted least squares estimator results from inverse weighting by the disturbance variances, and the same asymptotic distribution is obtained if consistent estimates of the disturbance variances are used. Thus, the precision of the variance estimates is a secondary concern (to first order), and Carroll (1982) and Robinson (1987) point out how nonparametric variance estimators yield the same conclusion. But this is not limited to variance estimates with a linear model. In particular, Andrews (1989) discovers an orthogonality condition that assures that the precision of the nonparametric estimators used is irrelevant; namely using a consistent estimator gives the same asymptotic distribution as using the true functions in the problem; and this orthogonality condition holds in a wide range of important examples. We note in passing that our average derivative estimators do not fit this paradigm; for instance the direct estimator $\widehat{\delta}_{\text{dir}} = N^{-1}\sum \widehat{m}'(x_i)\widehat{l}_i$ would have influence function $m'(x) - \delta$ if estimation of $m(x)$ made no difference, but its influence function is $r(y,x) = m'(x) - \delta + \ell(x)(y - m(x))$. Moreover, Newey and Stoker (1989) note how $r(y,x)$ is the efficient influence function, so that it is not possible to construct a regular estimator where the estimation of $m(x)$ (or $f(x)$) is irrelevant in the above fashion.

These sorts of issues are just now becoming understood, as part of the movement to "unify" asymptotic theory for semiparametric estimators. To date this attempt at unification is not complete, with regularity and other primitive conditions left for verification in individual problems. However, this work does reveal some of the basic structure of semiparametric methods. In this spirit, we introduce some recent theoretical work of Newey (1991), on the asymptotic variance of semiparametric estimators. This work, in my view, comes the closest to getting to the heart of the variance issues raised above.

A. Asymptotic Variances through Functional Derivatives

Newey's method of computing asymptotic variance is based on the correspondence between functional derivatives and variation, or influence representations.[6] For this introduction, consider the estimator $\overline{y} = N^{-1}\sum y_i$ of the mean $\mu = E(y)$. Further, suppose that the density of y is parameterized as $f^+(y \mid \theta)$, with the mean of y written as $\mu(\theta) = \int y f^+(y \mid \theta)dy$ for various values of θ. As before, $\overline{y}$ has the trivial influence representation

$$\sqrt{N}[\overline{y} - \mu(\theta)] = N^{-1/2}\sum [y_i - \mu(\theta)]$$

to which the Central Limit theory is applicable. We note further how the sensitivity of $\mu(\theta)$ to θ is expressed in this trivial case:

$$\frac{\partial \mu}{\partial \theta} = \int y \frac{\partial f^+}{\partial \theta} dy = \int y \mathbf{1}_\theta f^+(y \mid \theta)dy \tag{4.9}$$

where $\mathbf{1}_\theta = \partial \ln f^+ / \partial \theta$, which we could equivalently write as $\partial \mu / \partial \theta = E[(y - \mu(\theta))\mathbf{1}_\theta]$, since $E(\mathbf{1}_\theta) = 0$. As such, $y - \mu(\theta)$ is a gauge to how $\mu(\theta)$ varies with θ, through its interaction with the score $\mathbf{1}_\theta$. This formulation holds for any (differentiable) variations in the density of y, provided of course that we can differentiate with regard to θ under the integral above.

Alternatively, suppose that we have another estimator $\widehat{Y}$, that estimates $E(y) = \mu(\theta)$ when the density of y is $f^+(y \mid \theta)$. Since the position of the distribution of $\widehat{Y}$ varies with $\mu(\theta)$ in the same fashion as $\overline{y}$, one might conjecture

[6] In particular, Newey (1991) establishes standard connections for Gâteaux derivatives of Von Mises functionals, for general semiparametric problems.

that

$$\sqrt{N}[\widehat{Y} - \mu(\theta)] = N^{-1/2}\sum[y_i - \mu(\theta)] + o_p(1) \tag{4.10}$$

under suitable regularity conditions on $\widehat{Y}$. Moreover one might conjecture that the influence representation was connected to the "gauge" $y - \mu(\theta)$ of positional shifts $\partial\mu/\partial\theta$.

In this way, consider the estimation problem by starting from a representation such as (4.9). In particular, suppose that a magnitude a of interest could be solved for under variations of the density given by $f^*(y, x \mid \theta)$, namely as $\alpha(\theta)$, and for every (suitably regular) family we could write

$$\frac{\partial\alpha}{\partial\theta} = E(d\mathbf{l}_\theta) \tag{4.11}$$

where $\mathbf{l}_\theta = \partial \ln f^*/\partial\theta$. In this case, d is the "derivative" of (the functional) $\alpha(\theta) \equiv \alpha[f^*(y, x \mid \theta)]$. For parametric estimation problems, it is well known[7] that under certain regularity conditions, the derivative d determines the influence function for estimators of $\alpha(\theta)$. In particular, suppose that $\widehat{\alpha}$ is an estimator that consistently estimates $\alpha(\theta)$ for any density variation $f^*(y, x \mid \theta)$, with influence representation

$$\sqrt{N}(\widehat{\alpha} - \alpha(\theta)) = N^{-1/2}\sum\psi(y_i, x_i) + o_p(1); \tag{4.12}$$

so that $\sqrt{N}(\widehat{\alpha} - \alpha)$ has a limiting normal distribution with mean 0 and variance $E(\psi\psi^T)$. Then we have that $\psi = d - E(d)$, so that the asymptotic variance of $\widehat{\alpha}$ is the variance of the derivative d.

Newey's (1991) work gives general conditions under which the "influence-derivative" connection applies in semiparametric problems. He also indicates how this connection can be used as a device for characterizing the asymptotic variance of semiparametric estimators. Two main steps are required, as well as the verification of regularity conditions. Consider an estimator $\widehat{\beta}$. The first step is to characterize the limit of the estimator when the population density $f^*(y, x \mid \theta)$ takes on any "suitably regular" parametric form, including forms that do not necessarily obey the semiparametric restrictions on which $\widehat{\beta}$ is originally motivated. The second step is the computation of the derivative

$$\frac{\partial\beta}{\partial\theta} = E(d\mathbf{l}_\theta) \tag{4.13}$$

[7] See Serfling (1980) among others.

where one must be certain that the expectation factors into the derivative times the score for any regular parametric family. The conclusion is that the influence function of $\widehat{\beta}$ is $d - E(d)$, and the asymptotic variance of $\widehat{\beta}$ is the covariance matrix of d.

This device does require some regularity conditions to be verified — two important nonprimitive conditions are listed below. The first step may appear difficult or tedious — computing the limit of $\widehat{\beta}$ under arbitrary directions of misspecification — but is actually a valuable step for understanding the estimator. In particular, this step makes precise exactly what aspect of the data $\widehat{\beta}$ measures, regardless of the true model. For example, consider Powell's (1984) LAD estimator for censored Tobit models. While the Tobit model parameters are a tight characterization of the limit of this estimator, more valuable for empirical interpretation is the fact that the LAD procedure matches the median of y given x to a linear form in x. This fact summarizes in words what the consistent limit $\beta(\theta)$ is for this model.

Derivative calculations of this type can illustrate the sources of variation of a semiparametric procedure in a clear way. Newey carries this out to indicate a number of basic aspects of "suitably regular" semiparametric estimation problems.

For instance, the influence representation of the average derivative estimators can be rationalized, with an adjustment term identified for the estimation of the unknown functions. Consider the average derivative functional $\delta = E(m')$, where $m(x) = E(y \mid x)$ and $m'(x) = \partial m/\partial x$. Assume that the true density is an element of a parametric family $f^*(y, x \mid \theta) = q(y \mid x, \theta)f(x \mid \theta)$, where q is the conditional density of y given x and f is the marginal density of x. The density score $\mathbf{l}_\theta = \partial \ln q(y \mid x, \theta)/\partial\theta + \partial \ln f(x \mid \theta)/\partial\theta = \mathbf{l}_{c\theta} + \mathbf{l}_{m\theta}$, or the sum of conditional and marginal scores. Consequently,

$$\frac{\partial\delta}{\partial\theta} = \frac{\partial}{\partial\theta}\int m'(x;\theta)f(x;\theta)dx = E_\theta\left(\frac{\partial m'(x,\theta)}{\partial\theta}\right) + E[m'\mathbf{l}_{m\theta}]. \tag{4.14}$$

Note further that

$$E(m'\mathbf{l}_{m\theta}) = E[m'(\mathbf{l}_{c\theta} + \mathbf{l}_{m\theta})] \tag{4.15}$$

and that

$$
\begin{aligned}
E_\theta\left(\frac{\partial m'(x,\theta)}{\partial\theta}\right) &= \int \frac{\partial}{\partial x}\left(\int (y-m(x))\frac{\partial q}{\partial\theta}dy \mid x\right) f(x;\theta)dx \\
&= -\int [y-m(x)]\mathbf{1}_{c\theta}\frac{f'(x;\theta)}{f(x;\theta)}q(y \mid x;\theta)f(x;\theta)dydx \\
&= -\int [y-m(x)]\frac{f'(x;\theta)}{f(x;\theta)}[\mathbf{1}_{c\theta}+\mathbf{1}_{m\theta}]qfdydx \\
&= E[(y-m(x))\ell(x)\mathbf{1}_\theta].
\end{aligned} \tag{4.16}
$$

Thus, we have that

$$
\frac{\partial\delta}{\partial\theta} = E[(m'(x)+[y-m(x)]\ell(x))\mathbf{1}_\theta] \tag{4.17}
$$

so that the derivative is $d = m'(x)+[y-m(x)]\ell(x)$, and the associated influence function is $\psi = d - E(d) = m'(x) - \delta + [y-m(x)]\ell(x) = r(y,x)$, as given in Lecture 3. Therefore, the term $m'(x)-\delta$ is associated with variation in the density when the function m' is known, and $[y-m(x)]\ell(x)$ is associated with variation in the function m'. This interpretation is further confirmed by considering ratios of the components of the average derivative estimators. When a single index model is valid, these ratios correspond to the ratios of index model coefficients, which are solely a function of the regression $m(x)$, and therefore should not be affected by the variation represented by the first influence term $m'(x)-\delta$. This is easily verified; namely the variance of a ratio of average derivative estimators is determined by the second term $[y-m(x)]\ell(x)$, when the true model is an index model. This derivation and interpretations of it are given in Newey and Stoker (1989). Analogous arguments can be used to rationalize the form of the influence function for the density weighted average derivative, from examining the derivative of the functional $E(fm')$.

Given our discussion of U-statistics, it is useful to consider the derivative in this case.[8] Consider a U statistic of the form

$$
U = N^{-1}(N-1)^{-1}\sum_{i\neq j} p(z_i,z_j).
$$

The functional here is $E[p(z_i,z_j)]$, and a parametric model is represented by

[8] See Serfling (1980) among others.

$f(z \mid \theta)$. In this case, the derivative is easily seen to be

$$
\begin{aligned}
\frac{\partial E(p)}{\partial\theta} &= \frac{\partial}{\partial\theta}\int p(z_i,z_j)f(z_i \mid \theta)f(z_j \mid \theta)dx \\
&= \int p(z_i,z_j)[\mathbf{1}_\theta(z_i)+\mathbf{1}_\theta(z_j)]f(z_i \mid \theta)f(z_j \mid \theta)dz_idz_j \\
&= \int \{E[p(z_i,z_j) \mid z_i]+E[p(z_j,z_i) \mid z_i]\}\mathbf{1}_\theta(z_i)f(z_i \mid \theta)dz_i \\
&= E(d\mathbf{1}_\theta)
\end{aligned} \tag{4.18}
$$

where $d - E(d) = E[p(z_i,z_j) \mid z_i] - E(p) + E[p(z_j,z_j \mid z_i)] - E(p)$ is the influence function of the projection, introduced earlier.

Newey's study contains numerous examples, as well as theorems that allow checking of existing results on asymptotic variance. These include results on general "method of moments" estimation, as well results for checking the variance structure of the "regression" test statistic given above.

B. Asymptotic Irrelevance of the Nonparametric Estimation Technique In Semiparametric Estimation

Aside from these examples, consider the further implications of the "derivative-influence" connection here. In particular, the derivative d is derived from the functional of interest, which is a feature of the basic model, and not of the method of estimation. Consequently, the variance implied by the "derivative-influence" method cannot be affected by a particular method of nonparametric estimation. For instance, this is consistent with the notion that the same asymptotic behavior is exhibited by average derivative estimators based on kernel regression, and kernel density approximation. Alternatively, if polynomial expansions were used to estimate average derivatives, the resulting estimator has the same asymptotic variance as the kernel based estimators.[9]

The other main insights apply to the myriad of problems where the same asymptotic distribution is given for a semiparametric method based on nonparametric estimators, as for the same method based on the true functions (were they known). For instance, Newey demonstrates this feature for problems where the nonparametric estimators are chosen by optimizing the same objective function

[9] This follows from the verification of the regularity conditions, which is also done for polynomials in Newey (1991).

as that used to estimate the parameters, given the estimate of the function. This helps place Andrews (1989) results in a clear context.

What are the required regularity conditions? See Newey (1991) for a detailed exposition of them. However, two conditions which either must hold or be demonstrated are worthy of separate mention, because of their primary role in both Andrew's and Newey's analysis. Suppose that we consider a general setting where the econometric model dictates a moment condition

$$E[\mu(z, \beta, g(z))] = 0$$

for the true value of the parameter β and function $g(z)$. Suppose that the parameter β is estimated by solving the sample analog of this moment condition,[10] namely

$$\widehat{\beta} = \arg \underset{\tilde{\beta}}{\text{solve}}\{N^{-1} \sum \mu(z_i, \tilde{\beta}, \widehat{g}(z_i)) = 0\}.$$

Further, denote $M(z) = \partial \mu(z, \beta, g)/\partial g$, evaluated at the true values of β and g. Then two conditions that are required for the asymptotic analysis of the precision of $\widehat{\beta}$ are

Linearization:

$$N^{-1} \sum \mu[z_i, \beta, \widehat{g}(z_i)] = N^{-1} \sum \mu[z_i, \beta, g(z_i)] + N^{-1} \sum M(z_i)[\widehat{g}(z_i) - g(z_i)] + o_p\left(\frac{1}{\sqrt{N}}\right).$$

Stochastic Equicontinuity:

$$N^{-1} \sum M(z_i)[\widehat{g}(z_i) - g(z_i)] = E\{M(z)[\widehat{g}(z_i) - g(z_i)]\} + o_p\left(\frac{1}{\sqrt{N}}\right).$$

These main conditions require suitable regularity in the moment function μ with respect to the function argument g, as well as uniformity in the approximation of g by the nonparametric estimator $\widehat{g}$. These conditions permit an interesting set of tools to be applied to semiparametric problems, namely those of empirical

[10] More generally, $\hat{\beta}$ could be chosen to minimize the "size" $|N^{-1} \sum \mu(z_i, \tilde{\beta}, \hat{g}(z_i))|^T W |N^{-1} \sum \mu(z_i, \tilde{\beta}, \hat{g}(z_i))|$, where W is positive definite.

process theory, as demonstrated by Andrews (1989). Moreover, violations of these conditions would naturally be expected to cause serious problems for a "smooth" precision theory. However, demonstrating these conditions can be complicated; such as requiring uniform nonparametric approximation and trimming, as with the average derivative estimators (c.f. Härdle and Stoker (1989)).

Nevertheless, to the extent that these unifying connections are representative of many problems, this work does suggest broad answers for how to assess the precision impacts of using nonparametric estimators in semiparametric methods. Since, subject to regularity, the method of nonparametric estimation is secondary from the vantage point of asymptotic precision theory, one might naturally inquire as to where a difference could be found. In the next lecture we consider some troubling small sample problems, where the type of nonparametric estimation that is used can make a big difference to the results.

LECTURE 5

OUTLOOK: SUGGESTIONS AND CAUTIONARY NOTES

The earlier lectures presented a broad argument in favor of semiparametric methods as sensible empirical tools. We now turn to some prescriptions for future directions, as well as some practical problems.

I. General Outlook: Applications, Applications, Applications

These lectures have discussed the motivation for the semiparametric approach to econometric modeling, and given some empirical illustrations. We have argued for the attractiveness of these methods in general empirical research. However, it would be quite another thing to look back on a decade long empirical research program, and conclude that semiparametric methods have had real value in understanding economic relationships. The number of applications of semiparametric methods is still quite small, and experience with the methods is still quite limited. It is natural that the most important conclusion to a discussion of a new approach is to say that applications and empirical experience are the highest immediate priority.

It is also important to stress that applications should use the tools fully. In particular, one role of semiparametric methods is to assess the parameter values estimated from parametric models. As such, semiparametric methods could appear as a minor adjunct to standard econometric modeling, as professed in the recent survey by Pagan and Wickens (1989). However, just getting the same parameter estimates ignores an large part of the semiparametric approach, namely the characterizations of the unknown functions. For instance, consider Newey, Powell and Walker's (1990) study that confirms of Mroz's (1987) estimates of labor supply parameters. While the estimates in this study are not overly precise, graphs of semiparametric estimates of the selection probability and the selection term in the labor supply equation (which the authors have computed) would greatly aid the overall assessment of Mroz's model. For instance, how confident should one be in using a normal selected sample model with other data sets, if the estimated functions from the Mroz data bore no relation to the forms implied by assumptions of normal distributions?

The overall development of asymptotic theory for semiparametric estimators in the last five years has been nothing short of spectacular, and currently continues at the same pace. In fact, the efforts directed toward general theory have met considerable success in clarifying large sample sources of variation of semiparametric estimators. As such, it is natural to call for some reorientation toward studying questions of practical implementation. One immediate question concerns the "best" methods for setting the approximation parameters of nonparametric

estimators used in semiparametric methods, for moderate sample sizes. For nonparametric estimation, some progress has been made on these issues in the statistical literature (cross validation and the like). But since semiparametric problems involve different bias-variance tradeoffs, entirely different rules will be called for. Our applications of kernel estimators above have employed some benchmark suggestions (generalized cross validation and positive kernel functions), but these suggestions do not mesh completely with the asymptotic conditions, nor do they obviously arise from an optimality theory for the estimation of average derivatives.[1] Likewise, there exists no clear optimality theory for choosing the degree of a truncated polynomial or Fourier series, when the function is used as an ingredient to semiparametric analysis.

Other issues of practical implementation are associated with specific kinds of data problems, which are revealed only through application. For instance, when the data configuration has large gaps, it is not clear how to best implement quantile regression methods. This "coarseness" problem is discussed in Powell's (1987a) discussion of the Horowitz and Neumann (1987) study, as well as in Chamberlain (1991). Moreover, procedures that use smoothing will "smooth" over big gaps, but it is natural to suspect that substantive mismeasurement could arise in these cases. At any rate, a myriad of implementation problems are likely to arise with more empirical experience, and they should have high priority in the study of these methods. This is not to say that the interesting theoretical questions have been exhausted; quite the contrary is true; but rather just an appeal for balance in a methodological research program.

The final suggestion concerns examining the quality of the (now standard) asymptotic theory as an approximation with small or moderate sample sizes. This concern is not new, and has hardly been addressed in situations where complicated nonlinear parametric models have been applied. But for semiparametric and nonparametric methods, there are additional reasons for examining the quality of asymptotic approximations.

For semiparametric methods that use nonparametric estimators, results that assert asymptotic normality involve more than central limit theory, or that sample averages are approximately normally distributed in moderate samples. Such

[1] Härdle, Hart, Marron and Tsybakov (1990) study the optimal bandwidth problem for the "indirect" estimator in the one-dimensional case.

results also involve "finer and finer" functional approximation as the sample size is increased, such as the bandwidth vanishing for kernel estimators, or the degree increasing for truncated series expansions. This is an essential feature of the semiparametric approach, but leaves open the question of what kinds of mismeasurement are possible in small or moderate samples. In particular, what structure will kernel estimators miss or mismeasure when a substantive bandwidth value is set, or likewise, what structure will a series expansion miss when the truncation rule only includes a small number of leading terms? Obviously silent on this issue are results from asymptotic theory that say that the method of nonparametric approximation makes no difference.

In the empirical examples in the lecture, we have paid homage to this issue by using variance estimators that mimic U-statistic structure, using the loose justification that such estimators may more adequately proxy variation than alternative, unrelated measures. Later in this lecture, we discuss the smoothing bias problem, that asserts that particularly large differences can arise with different methods of nonparametric approximation, and underlies our earlier use of the "slope" version of average derivative estimators. At any rate, at this juncture we just note that the extensive asymptotic theory now available does not rule out surprising differences in empirical results from "equivalent" estimation methods. Applications, analysis and simulation studies are definitely necessary to confirm the substantial promise of semiparametric methods.

The remaining sections of the lecture take some editorial license, in describing problems and work that are recent and not at all well developed, with the intention of stimulating interest in them. The first group concerns two issues with index models and average derivative estimation; namely the accommodation of discrete predictors and nonlinear index variables. The second group concerns smoothing bias as it impacts on the measurement of derivatives.

II. Further Issues of Index Model Estimation

Our discussion of index model estimation and average derivatives has involved two specializations that can potentially limit the verification of results obtained from parametric models. In particular, the basic index model

$$E(y \mid x) = G(x^T\beta) \tag{5.1}$$

has taken x to be a continuous random vector, where no component of x is functionally determined by other components of x. We now discuss issues and solutions that arise from relaxing these characteristics of the predictor vector x. We begin with some remarks on the discrete variable problem, and then discuss an approach to the estimation of models with nonlinear index variables.

A. The Discrete Variable Problem in Index Models

We now give some speculative remarks on a difficult problem, namely estimating coefficients of discrete variables in index models. Discrete predictor variables whose impact is not linear often lead to problems, because there is no obvious way to exploit smoothness of the connection between the response and the discrete variable. When the discrete variable takes on many values, such as age in years, it may be acceptable to smooth over its values if the age impact is regarded as smooth. But for qualitative 0 - 1 variables, such smoothing is not likely to give an accurate depiction. In addition, this problem becomes quite extreme when all predictors are discrete, wherein the coefficients of an index model are not identified separately from the unknown function of the index (Chamberlain (1986a)), even up to scale.

Because of our focus on average derivatives, we discuss the difficulties in defining an analogous notion of average derivative for discrete variables. First, however, it is useful to note that the presence of discrete variables does not prohibit the estimation of continuous variable coefficients up to scale. In particular, the connection of index model coefficients to derivatives is valid given the values of discrete variables, so that average derivatives can be constructed and estimated from derivative estimates that vary with the values of all continuous and discrete variables. In this way, the discrete predictor variables alter the data requirements, by requiring nonparametric estimation of score or regression derivatives for each value of the discrete variables. While this necessarily implies that the nonparametric derivative estimates are less precise (than if the discrete predictors were absent), it is not clear whether less precision is implied for the averaged derivatives, because the advantages of "double averaging" are still applicable.

Moreover, given identification of discrete variable coefficients, the estimation problem is not difficult theoretically, but rather one of finding a direct, computationally simple coefficient estimator. For instance, it is quite natural to propose

estimators of discrete variable coefficients using semiparametric least squares, with simplified versions of the method of Ichimura (1986) and Härdle, Hall and Ichimura (1991). For instance, suppose the predictor variables include a single qualitative variable D and a vector of continuous variables x, and the true regression is an index model

$$E(y \mid x, D) = G(D\Delta + x^T\delta). \tag{5.2}$$

Suppose δ is estimated by an average derivative estimator $\widehat{\delta}$, that accounts for the presence of D as above. Suppose also, that by using the observed data points with $D = 0$, the function G is estimated as $\widehat{G}$ (since $E(y \mid x) = G(x^T\delta)$ for that data). Then Δ could be estimated by least squares, as in

$$\widehat{\Delta} = \arg\min_{\tilde{\Delta}} \sum [y_i - \widehat{G}(D_i\tilde{\Delta} + x_i^T\widehat{\delta})]^2. \tag{5.3}$$

It is natural to conjecture that the estimator $(\widehat{\Delta}, \widehat{\delta})$ so constructed would be $\sqrt{N}$ asymptotically normal. The estimator $\widehat{G}$ could then be reestimated over the entire sample, and the process iterated, for potentially more efficient results.

It is intriguing to consider the possibility of a more direct estimator of the discrete variable coefficients, in the same spirit as the (continuous variable) average derivative estimators. One natural candidate is to estimate the regression $E(y \mid D, x) = m(D, x)$ with a nonparametric estimator $\widehat{m}(D, x)$, and form the average difference

$$\tilde{\Delta} = N^{-1} \sum [\widehat{m}(1, x_i) - \widehat{m}(0, x_i)] \tag{5.4}$$

as an estimator of the mean difference

$$E[m(1, x) - m(0, x)]. \tag{5.5}$$

When the regression is in index form, $m(D, x) = G(D\Delta + x^T\delta)$, then the average difference is nonzero only when Δ is nonzero. The problem is that the mean difference cannot be proved to obey the same scaling as the continuous coefficients δ, as a Taylor expansion shows. It is worthwhile noting that if the model is partially linear, as in

$$E(y \mid x, D) = \alpha + D\Delta + G(x^T\delta) \tag{5.6}$$

then the average difference measures Δ, the mean difference in this case. This underscores the point that the discrete variable "problem" is only a problem if one has a strong theoretical reason to model the discrete variable impact through the index variable. The difference between models (5.2) and (5.6) is that in (5.2), the variable D alters the level of the index $D\Delta + x^T\delta$, but in model (5.6), the variable D alters the level of the mean response $E(y \mid x, D)$, given the continuous index $x^T\delta$. Model (5.6) may be perfectly adequate for giving an accurate (dimension reduced) depiction of the data, and therefore may be recommendable for ease of estimation.

It is interesting to note that the delicacy of this problem extends to the use of proxies. For instance, suppose that a (threshold crossing) binary response model applies to D,

$$D = 1[\psi(z, x) < \varepsilon] \tag{5.7}$$

where z is a vector of additional continuous variables. Suppose that the continuously distributed probability $E(D = 1 \mid z, x) = P(z, x)$ were known, setting aside the issues of estimating it. What happens if we replace D by P for estimation of the index model $E(y \mid D, x) = G(D\Delta + x^T\delta)$, computing the average derivative of y with respect to (P, x)?

It is easy to see that the average derivative of y with respect to P estimates the mean difference above, and thus the proxy approach measures the same effect as the average difference. In particular,

$$E(y \mid z, x) = E(y \mid P, x) = Pm(1, x) + (1 - P)m(0, x). \tag{5.8}$$

Therefore,

$$E\left[\frac{\partial E(y \mid P, x)}{\partial P}\right] = E[m(1, x) - m(0, x)] \tag{5.9}$$

so that this method has the same features as the average difference. Of course, if the application suggests that P is the proper predictor in the index (and if estimating P raised no complications), then the index model coefficients could be measured as usual with average derivative estimators.

B. Estimation of Nonlinear Index Coefficients

When the function G of an index model $E(y \mid x) = G(x^T\beta)$ is specified parametrically, one could employ nonlinear variations of the basic index $x^T\beta$ to describe the data. In particular, one could include powers or other nonlinear transformations of the basic predictor variables, fitting a separate coefficient for each transformation. This kind of modeling is occasionally practiced in econometrics, when the index variable has a strong interpretation, and that interpretation suggests certain kinds of nonlinearity in the index. For instance, consider a discrete choice model where the index represents a difference in utilities. If the empirical problem suggests that the utility difference is nonlinear in certain predictor variables, then one might want to use a nonlinear index model, while simultaneously taking a strong stand on the form of the function $G(\cdot)$, the distribution of unobserved elements of the utility. For the following discussion, we assume that the empirical problem of interest gives a strong reason for incorporating nonlinearity in the index variable relative to the function G.

This issue becomes quite delicate when the function G is to be measured nonparametrically. For instance, suppose there is a single predictor x_1 in the model

$$E(y \mid x_1) = G(x_1 + \beta x_1^2). \tag{5.10}$$

If G is specified, then β can be identified and estimated. However, if G is estimated nonparametrically, then β is not identified, as the regression equation is indistinguishable from $E(y \mid x_1) = m(x_1)$, where m is the univariate function defined by $m(x_1) \equiv G(x_1 + \beta x_1^2)$.

When there is more than one predictor variable, the coefficients can often be identified separately from the function G. A setting where this would be of interest is where one is checking the specification of an estimated parametric model with a nonlinear index. At any rate, consider a model with two continuous predictor variables x_1 and x_2 , in the following form.

$$E(y \mid x_1, x_2) \equiv m(x_1, x_2) = G(\beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \delta_2 x_2). \tag{5.11}$$

For simplicity, denote the partial derivatives as $m_1' = \partial m/\partial x_1$, $m_2' = \partial m/\partial x_2$, so that $m' = (m_1', m_2')^T$. For this model, G can be defined to assure the scaling

$\delta_2 = E(m_2')$, so that δ_2 is the average derivative of y with respect to x_2 . The question is how to measure β_1, β_2 and β_3 subject to this scaling. One method would be to apply least squares as in (5.3), but there exists a more direct method, based on the same nonparametric estimators used to estimate average derivatives.

The method is based on a simple insight by Ai (1991b) that a nonlinear index gives rise to more involved derivative restrictions that a model with a linear index. Recall how a regression based on a linear index is characterized by proportionality restrictions between derivatives for different predictors. Ai's method is to implement such restrictions directly, in the nonlinear index case. For model (5.11), write out the regression derivatives as

$$\begin{aligned} m_1' &= G'(\beta_1 + 2\beta_2 x_1 + 3\beta_3 x_1^2) \\ m_2' &= G'\delta_2. \end{aligned} \tag{5.12}$$

Use the second equation so eliminate G' from the first equation, as

$$\begin{aligned} m_1' &= \left(\frac{\beta_1}{\delta_2}\right) m_2' + \left(\frac{\beta_2}{\delta_2}\right) 2x_1 m_2' + \left(\frac{\beta_3}{\delta_2}\right) 3x_1^2 m_2' \\ &\equiv \alpha_1 m_2' + \alpha_2 2x_1 m_2' + \alpha_3 3x_1^2 m_2' \end{aligned} \tag{5.13}$$

where $\alpha_j \equiv \beta_j/\delta_2$, , $j = 1, 2, 3$. Equation (5.13) is the restriction among regression derivatives implied by the model (5.11). Moreover, it is easy to propose estimation methods that implement this restriction. For instance, one could estimate the regression $m(x_1, x_2)$, compute the derivatives $\widehat{m}_1'(x_1, x_2)$ and $\widehat{m}_2'(x_1, x_2)$ for each observation, and estimate $\alpha_j = \beta_j/\delta_2$, $j = 1, 2, 3$ by ordinary least squares applied to (5.13) (using the estimated derivatives). The coefficient β_j is then estimated from $\beta_j = \alpha_j\delta_2$. See Ai (1991b) for a discussion of the generality of the method, as well as the statistical properties of the coefficient estimators. The practical attraction is how the method depends only on the direct nonparametric estimates of the regression $m(x_1, x_2)$, and not on an iterative scheme of semiparametric least squares. Moreover, with reference to the next section, estimation of equation (5.13) is robust to proportional smoothing bias in the estimated values $\widehat{m}'(x_1, x_2)$.

III. Smoothing Bias in Derivative Estimates

A. Motivation

In Lecture 3, we discussed how the different estimators for average derivatives were asymptotically equivalent, namely the direct, indirect and slope formulations. We also used this fact for motivation of Newey's work on asymptotic variances, as well as a potential reason for looking beyond first order asymptotic theory for estimator precision in small or moderate samples. In our empirical examples, we advocate the use of the "slope" formulations, and we now turn to the reason for this recommendation. We can begin by noting that the Monte Carlo simulations of "density weighted" estimators in Powell, Stock and Stoker (1989) gave some preference for slope estimators using positive kernels. But those results used estimators that were normalized to have the same scale for comparison, which is fine when one is only concerned with estimation of index coefficients up to scale. However, when one simulates the average derivative estimators of Section 3 without normalizing, a large difference is noted between the scale of the estimators based on the levels of derivatives, and the slope estimators in ratio form.

For motivation, consider the following results from a limited Monte Carlo simulation of the average derivative estimators.[2] Table 5.1a,b contains means and standard deviations of the various average derivative estimators over 20 Monte Carlo samples. The design for each table has $k = 4$ normally distributed x's, distributed with zero mean and the identity as covariance matrix. The response y is simulated from a linear model in Table 5.1a, and for a probit model in Table 5.1b. A normal kernel is used and the bandwidth is set to $h = 1$, which matches the standard deviation of each component of x. The linear model of Table 5.1a has $R^2 = .80$, and we have included OLS estimators for comparison. We also include OLS estimators for the probit model of Table 5.1b, as motivated by the normal design of the x's (c.f. Lecture 2).

In both cases, the means of the "direct" and "indirect" estimators are roughly half the true values; whereas the means of the "slope" estimators are roughly equal to the true values. Since these are asymptotically equivalent estimators, it seems entirely possible that there is something wrong with the standard asymptotic approximation. Of course, the bandwidth $h = 1$ was not chosen automatically, and therefore could be an awful choice. However, my experience has been that smaller bandwidth values produce more erratic estimator behavior and larger bandwidth values produce a bigger difference between the "slope" and the other kinds of estimators. Likewise, the sample size of $N = 100$ may be tiny for a four-dimensional estimation problem, so that it may be unreasonable to think that the asymptotic theory should have relevance. At any rate, I would invite others to repeat simulations such as these, because the observed differences in the estimators seem the rule, not the exception.

In particular, two features seem common from simulations like these. First, the estimators based on levels of estimated derivatives are too small. Second, estimators in ratio form, or the "slope" estimators, are surprisingly accurate. Aside from recommending the "slope" estimators, how is one to explain this difference?

[2] Extended versions of these simulations for many designs will be reported in Stoker and Villas-Boas (1992).

Linear Model: $y_i = 1 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + \varepsilon_i$

TABLE 5.1a:
SIMULATION RESULTS — LINEAR MODEL
True Value: $\delta = (1, 1, 1, 1)$
$N = 100$, $h = 1$, K spherical normal, $(x, \varepsilon) \sim \mathcal{N}(0, I)$

	$\widehat{\delta}_1$	$\widehat{\delta}_2$	$\widehat{\delta}_3$	$\widehat{\delta}_4$
Indirect	.389	.390	.404	.385
	(.047)	(.078)	(.039)	(.062)
Direct	.447	.453	.477	.444
	(.063)	(.101)	(.059)	(.076)
Indirect Slope	.991	1.01	1.03	.984
	(.101)	(.154)	(.102)	(.101)
Direct Slope	1.01	1.02	1.03	.985
	(.098)	(.152)	(.115)	(.116)
OLS	1.01	1.01	1.02	.976
	(.078)	(.128)	(.111)	(.107)
PSS	1.02	1.02	1.03	.986
	(.130)	(.170)	(.131)	(.126)

Means and Standard Deviations from 20 Monte Carlo Samples

Probit Model: $y_i = 1[1 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + \varepsilon_i > 0]$

TABLE 5.1b:
SIMULATION RESULTS — BINARY RESPONSE MODEL

True Value: $\delta = (.161, .161, .161, .161)$

$N = 100$, $h = 1$, $\mathcal{K}$ spherical normal, $(x, \varepsilon) \sim \mathcal{N}(0, I)$

	$\widehat{\delta}_1$	$\widehat{\delta}_2$	$\widehat{\delta}_3$	$\widehat{\delta}_4$
Indirect	.063	.068	.070	.063
	(.021)	(.022)	(.015)	(.013)
Direct	.082	.083	.083	.076
	(.021)	(.023)	(.018)	(.014)
Indirect Slope	.177	.179	.171	.164
	(.033)	(.040)	(.036)	(.032)
Direct Slope	.186	.186	.178	.168
	(.039)	(.046)	(.036)	(.034)
OLS	.171	.171	.168	.160
	(.035)	(.033)	(.035)	(.028)
PSS	.194	.193	.182	.172
	(.047)	(.056)	(.038)	(.039)

Means and Standard Deviations from 20 Monte Carlo Samples

Addendum to Tables 5.1a,b

Estimators of Average Derivatives $[\delta = E(m')]$

Direct: $\widehat{\delta}_{\text{dir}} = N^{-1} \sum_{i=1}^{N} \widehat{m}'(x_i) \widehat{1}_i$

Indirect: $\widehat{\delta}_{\text{ind}} = N^{-1} \sum_{i=1}^{N} \widehat{\ell}(x_i) y_i \widehat{1}_i$

Indirect Slope: $\widehat{d}_{\text{ind}} = \left[N^{-1} \sum_{i=1}^{N} \widehat{\ell}(x_i)(x_i - \overline{x})^T \widehat{1}_i \right]^{-1} \left[N^{-1} \sum_{i=1}^{N} \widehat{\ell}(x_i)(y_i - \overline{y})^T \widehat{1}_i \right].$

Direct Slope: $\widehat{d}_{\text{dir}}$ (c.f. Lecture 3)

OLS: Ordinary Least Squares

PSS: $\widehat{d}_f$ (c.f. Lecture 3)

One thing that does differ from the limiting asymptotic conditions is the bandwidth value $h = 1$, which is certainly not minuscule. But if this is the culprit, there must be some sort of systematic problem with measuring derivatives from a smoothing estimator. This kind of explanation is advanced in my recent papers Stoker (1991b,1991c); namely that smoothing imparts a systematic, generic downward bias to estimated derivatives, whether in the form of density score vectors or regression derivatives. Moreover, in leading examples this downward bias can be seen to be approximately proportional across the sample. Consequently, the "direct" and "indirect" estimators are afflicted by downward bias, but the bias "cancels out" of the ratio form "slope" estimators.

Below we spell out the basic ideas of how derivatives can be mismeasured by smoothing. Before beginning, it is useful to note how this explanation is based on analysis that differs from the standard asymptotic theory. In particular, the bias is studied directly. This coincides with an approximation that is "asymptotic" in that the sample size is considered large, but has the bandwidth value remaining fixed. In this way we are able to focus on the impact of smoothing.

Recall our basic notation and setup. The data (y, x), $i = 1, \ldots, N$ is a random sample, where y is a response variable and x a k-vector of predictor variables. The marginal density of x is $f(x)$, with derivative $f'(x) = \partial f / \partial x$ and (translation) score $\ell(x) = -f'(x)/f(x)$. The mean regression of y on x is $m(x) = E(y \mid x)$, with derivative $m'(x) = \partial m / \partial x$. The next section discusses results of Stoker (1991b) on the bias problem in estimation of density functions, or downward bias in estimators of $f'(x)$ and $\ell(x)$. The following section discusses the results of Stoker (1991c) on the bias problem in estimation of regression functions, or downward bias in estimators of $m'(x)$.

B. Smoothing Bias in Density Derivative and Score Estimators

The standard kernel density estimator is

$$\widehat{f}(x) = N^{-1} h^{-k} \sum_{i=1}^{N} \mathcal{K}\left(\frac{x - x_i}{h} \right);$$

where we take the kernel $\mathcal{K}(\cdot)$ to be a positive density function. This estimator is easily seen to be the density of the random variable $X + hu$, where X is distributed with the empirical distribution of the data (x_i), and u is distributed with density

$\mathcal{K}$, independently of X. This convolution structure extends to the expected value of $\widehat{f}(x)$, namely:

$$E[\widehat{f}(x)] = \int \mathcal{K}(u) f(x - hu) du \equiv \phi_h(x),$$

where ϕ_h is the density of $x + hu$, where x is distributed with density $f(x)$, and u with density $\mathcal{K}(u)$. If $N \to \infty$ but h is held fixed, then we have

$$\text{plim} \widehat{f}(x) = \phi_h(x).$$

By an analogous argument, we have that $E[\widehat{f}'(x)] = \phi_h'(x)$, and

$$\text{plim } \widehat{f}'(x) = \phi_h'(x).$$

Therefore, under these guidelines the estimated score $\widehat{\ell}(x) = -\widehat{f}'(x)/\widehat{f}(x)$ can be shown to estimate

$$\text{plim } \widehat{\ell}(x) = -\frac{E[\widehat{f}'(x)]}{E[\widehat{f}(x)]} = -\frac{\phi_h'(x)}{\phi_h(x)} \equiv \lambda_h(x).$$

In sum, with the bandwidth fixed, the density estimator $\widehat{f}(x)$ estimates the convolution ϕ_h evaluated at x, and the score $\widehat{\ell}(x)$ estimates the score λ_h of the convolution ϕ_h, likewise evaluated at x. If $h \to 0$, then $\phi_h(x) \to f(x)$, $\phi_h'(x) \to f'(x)$ and $\lambda_h(x) \to \ell(x)$, but here we consider what happens if the bandwidth h is fixed.

How could this structure be associated with a downward bias in derivatives? For motivation, consider Figure 5.1. We have drawn a density $f(x)$, and and a dotted line representing a smoothed version ϕ_h of the density. The smoothed version in drawn in line with Jensen's inequality, and is noticeably flatter. But "flatter" means that derivatives from the smoothed version will typically be smaller than derivatives from the original, as portrayed in the graphs of the derivatives f' and ϕ_h'. This is the phenomena of smoothing bias in density derivative estimates. Figure 5.2 gives a picture of a bimodal density with the same conclusion. In other words, the operation of estimation by smoothing serves to dampen measured density derivatives.

This picture is not informative regarding the estimation of the score $\ell(x)$, but for this we present an example based on normal regressors and a normal kernel. In particular, suppose that $x \sim f = \mathcal{N}(\mu_x, \Sigma_x)$, $u \sim \mathcal{K} = \mathcal{N}(0, I)$. This implies

that ϕ_h , the density of $x + hu$, is the $\mathcal{N}(\mu_x, \Sigma_x + h^2 I)$ density. Computing scores, we have that

$$\ell(x) = -\frac{f'(x)}{f(x)} = \Sigma_x^{-1}(x - \mu_x)$$

and that the estimated score measures

$$\lambda_h(x) = -\frac{\phi_h'(x)}{\phi_h(x)} = (\Sigma_x + h^2 I)(x - \mu_x) = A_h \ell(x)$$

where $A_h = (\Sigma_x + h^2 I)^{-1} \Sigma_x$ is the matrix factor of proportionality. In particular, the structure of A_h indicates that λ_h smaller than $\ell(x)$, in the sense of being a matrix weighted average of $\ell(x)$ and 0. The downward bias is more evident if the x is distributed with covariance matrix $\Sigma_x = \sigma_x^2 I$; in which case

$$\lambda_h(x) = (1 - \nu)\ell(x) \quad \text{with} \quad \nu = \frac{h^2}{\sigma_x^2 + h^2}$$

so that the downward bias is is uniform across components of x, and determined by the relative size of σ_x^2 and h^2.

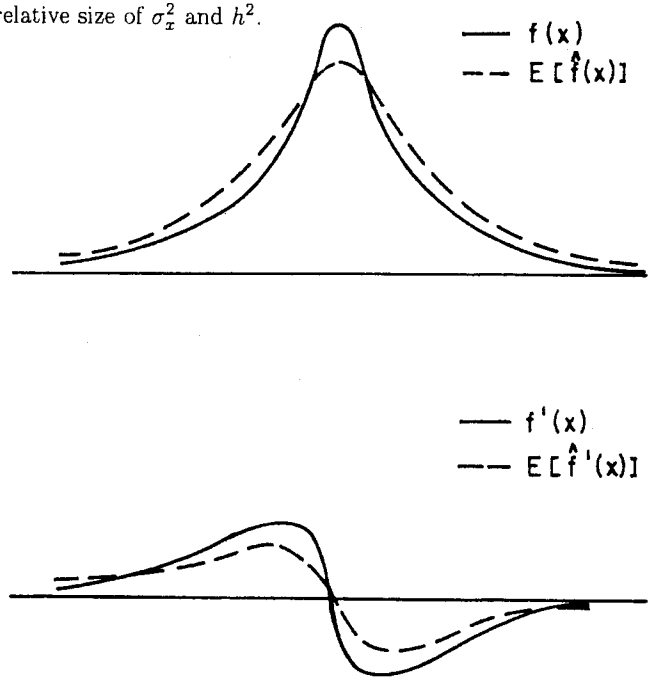


Figure 5.1
Smoothing Bias in Density Derivatives.

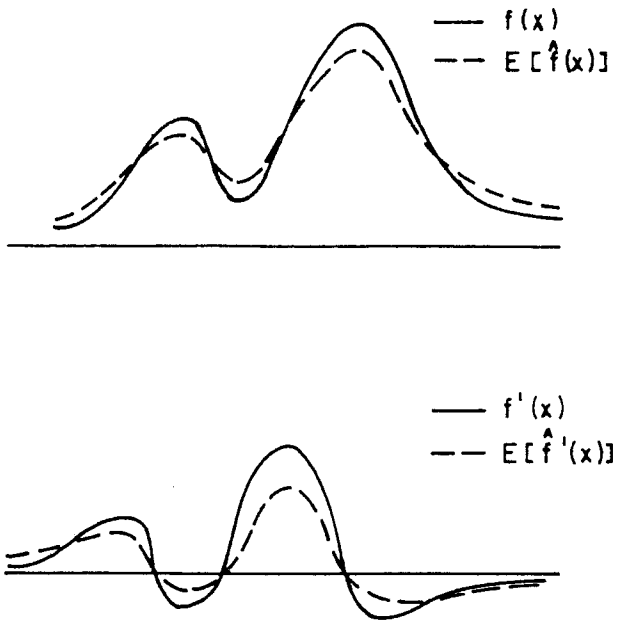


Figure 5.2

Smoothing Bias in Density Derivatives; Bimodal Case.

For the design of Tables 5.1a,b, $\sigma_x^2 = 1$ and $h^2 = 1$, so $\nu = 1/2$. This analysis implies that the estimated scores $\widehat{\ell}(x)$ will be roughly half the value of the true score $\ell(x)$, so that the "indirect" estimator should be roughly half the value of the true average derivative. While not exactly matching the means of the "indirect" estimator in Tables 5.1a,b, this smoothing effect is substantial. Moreover, the fact that the bias is uniformly proportional implies that it is cancelled out of the indirect "slope" estimator, which is consistent with the results of Tables 5.1a,b.

Stoker (1991b) gives a more precise development of these ideas, as well as some general results. Two further features of the analysis are worthy of mention. First, the score bias is seen to be approximately proportional in examples with normal mixture densities, so that proportional downward bias may be a decent practical approximation in nonnormal settings. Second, typical bias values are computed using various optimal bandwidth formulae, and are seen to be considerable. The derivative bias is especially large in problems of moderate dimension (say $k > 2$), even with large sample sizes.

C. Smoothing Bias in the Estimation of Regression Derivatives

We now consider nonparametric estimation of the regression $E(y \mid x) = m(x)$ by the kernel estimator

$$\widehat{m}(x) = \frac{\widehat{c}(x)}{\widehat{f}(x)}; \quad \text{where} \quad \widehat{c}(x) = N^{-1}h^{-k}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) y_i.$$

The analysis of bias in the derivatives of kernel regression estimators is similar to that for density estimators, but the bias arises for a somewhat different reason. For given h, as $N \to \infty$, we have that $\widehat{m}(x)$ estimates

$$\text{plim } \widehat{m}(x) = \frac{E[\widehat{c}(x)]}{E[\widehat{f}(x)]} = \frac{\int y\Phi_h(y,x)dy}{\phi_h(x)} \equiv \gamma_h(x).$$

Here Φ_h is the joint density of y, $x + hu$, where x, u are independent as before, and ϕ_h is the marginal density of $x + hu$. Therefore, the function γ_h is the conditional expectation of y given $z = x + hu$, not x. Consequently, $\widehat{m}(x)$ measures $\gamma_h(z) = E(y \mid z)$ evaluated at $z = x$. Likewise, $\widehat{m}'(x)$ measures

$$\text{plim } \widehat{m}'(x) = \gamma_h'(x)$$

or the derivative of the regression of y on z, evaluated at $z = x$.

With density estimation, the derivative bias arises from the "flattening" effect of convolution. With regression estimation, the comparison of $\gamma_h'(x)$ to the true derivative $m'(x)$ involves the impact of conditioning with respect to a variable distributed as a convolution. This kind of problem has a substantive history in econometrics, namely as the "errors-in-variables" problem. Here we note that the kernel estimator $\widehat{m}(x)$ does not estimate the regression m of y with respect to x, but rather the regression γ_h of y with respect to the "erroneous" variable $z = x + hu$. As above, if $h \to 0$, then $\gamma_h(x) \to m(x)$, but here we study the difference for a given value of the bandwidth h.

Although the structure is different from the case of estimating density, a downward bias can still arise. Consider the standard "errors-in-variables" bias problem for estimation of the coefficients of a linear model. Assume that the regressors are normal; $x \sim f = \mathcal{N}(\mu_x, \Sigma_x)$; a normal kernel is used; $u \sim K = \mathcal{N}(0, I)$; and that the true model is linear; $m(x) = \alpha + \beta^T x$, with $m'(x) = \beta$. Standard "errors in variables" calculations show that

$$\gamma_h(x) = [\alpha + (\beta - A_h\beta)\mu_x] + (A_h\beta)^T x$$
$$\gamma_h'(x) = A_h\beta;$$

where $A_h = (\Sigma_x + h^2 I)^{-1}\Sigma_x$ is the same factor that was found in the analysis of the density score. Again, if $\Sigma_x = \sigma_x^2 I$, the factor A_h is just $(1-\nu)I$, where $\nu \equiv h^2/(\sigma_x^2 + h^2)$ is that familiar "noise"/"total variance" ratio of "errors in variables" formulae. Also as before, the design of Table 5.1a implies that $A_h = (1/2)I$, so that the derivatives of the kernel regression should be roughly half the true values, and the "direct" estimator roughly half the true average derivative value.

The general analysis of the downward bias phenomena is difficult, as evidenced by the paucity of results on general nonlinear errors-in-variables problems. A pictorial analysis can be devised to show how the downward bias arises, as follows. Suppose the true model had $y = m(x) + \varepsilon$, with ε independent of x. Then if $z = x + hu$, we have

$$\begin{aligned}\gamma_h(z) &= E[m(z - hu) \mid z] \\ &= E\{m[z - hE(u \mid z) - e] \mid z\}\end{aligned}$$

where $e = h[u - E(u \mid z)]$. Smoothing alters m to γ_h in two ways. First, the argument at which m is evaluated is shifted from z to $z - hE(u \mid z)$. Second, the form of m is altered by integrating over e. The argument-shift effect can induce "flattening," or downward derivative bias, whereas the effect of the integration is not clear. For the case where $m(x)$ is linear, the latter effect is inconsequential. Figure 5.3 displays the case of the normal linear example given above, illustrating how the "argument-shift" induces downward bias. Figure 5.4 depicts nonlinear relations where the "argument-shift" induces downward bias.

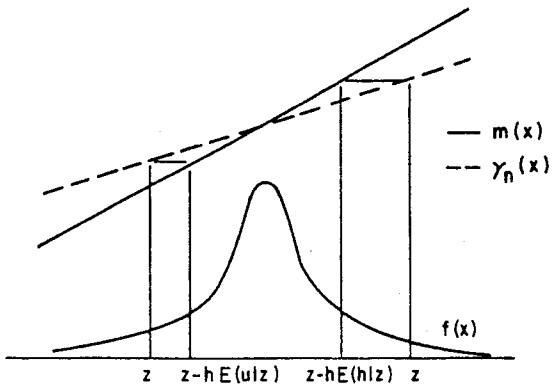


Figure 5.3
Downward Derivative Bias; Linear Model.

Stoker (1991c) gives a more precise development of these ideas, including results that connect the downward derivative bias to the shape of the regressor density, and numerous examples. A more complete result is given for the case of normal regressors and a normal kernel. Namely, suppose that $x \sim \mathcal{N}(\mu_x, \Sigma_x)$, $\mathcal{K}(u)$ is the spherical normal density, and that $y = m(x) + \varepsilon$, where ε is is distributed with mean 0, independently of x. Then the average of estimated derivatives can be written as

$$E[\gamma_h'(x)] = A_h E_w[m'(w)]$$

where $A_h = (\Sigma_x + h^2 I)^{-1}\Sigma_x$ as before, and $w \sim \mathcal{N}(\mu_x, \Sigma_x[I - A_h(I - A_h)])$. Therefore, the downweighting factor applicable to the linear case appears in general (in fact, because of the "argument shift"). This factor is applied to an average derivative value associated with a more compact normal distribution than that of the original regressors. The latter feature is inconsequential when the true model is linear or quadratic, wherein the average of the estimated derivatives is A_h times the average of the true regression derivatives.

D. Differences in Nonparametric Methods: Polynomial Approximations

One of the themes of the opening remarks of the lecture is that different nonparametric methods can make a difference to the results obtained. Consequently, we close this section by noting how the derivative bias problem does not exist to the same extent for other nonparametric estimators; in particular, for regression estimators fit by global fitting criteria.

In reflection, the relevance of the fitting criteria is quite natural. We have indicated how kernel estimators have biased derivatives with linear models. This goes hand-in-hand with the fact that computing a kernel estimator by setting $y_i = x_i$ does not reproduce the function $m(x) = x$. If an alternative (linear) method assures that "x" is reproduced in this fashion, it will not display bias for linear models in general. Such is the case if a formula is fit by global least squares, and the function "x" can be approximated by the formula. For example, such is the case with polynomial approximations (of degree 1 or greater). One might guess the the bias in derivatives would generally be smaller for such procedures.

A couple formulas can be derived to illustrate this point (c.f. Stoker (1991c)). In particular, suppose $\tilde{m}(x)$ is a nonparametric estimator of $m(x)$, but has limit

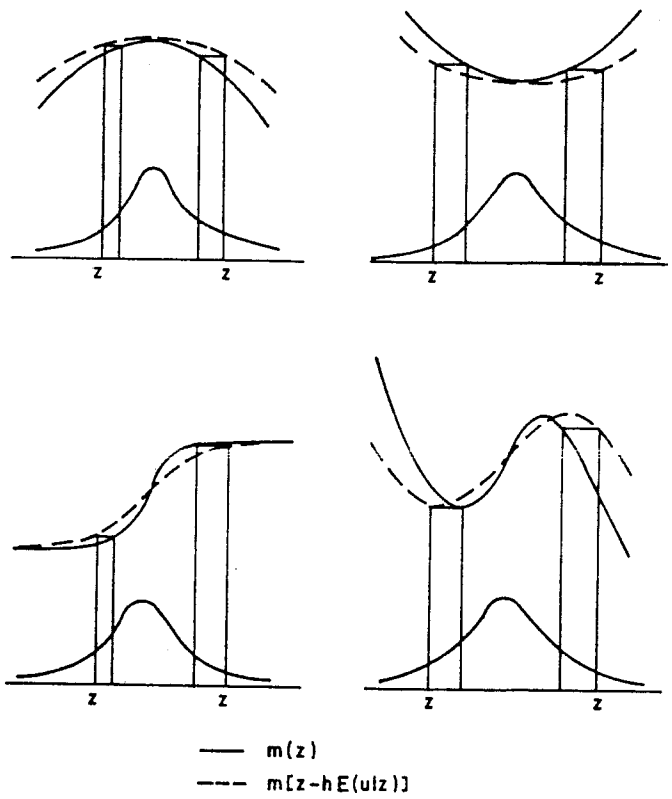


Figure 5.4
Flattening via "Argument Shift".

$\gamma(x)$ when the approximation is not made finer with sample size. For kernel estimators, γ is the limit γ_h for fixed bandwidth h, and for truncated polynomial expansions of degree q, γ is the polynomial of degree q that best approximates $m(x)$ in terms of integrated squared error. Regardless of the method, if the density f(x)

vanishes on the boundary of its support, the average bias in derivatives can be written as

$$E[\gamma'(x)] - E[m'(x)] = E\{\ell(x)[\gamma(x) - m(x)]\}$$

and so it depends on the interaction between the score $\ell(x)$ of the regressor density and the approximation error $\gamma(x) - m(x)$ of the regression function.

Now, suppose that $\tilde{m}(x)$ is restricted to be an element of a collection of functions $\mathcal{P}_q$, and is estimated by (global) least squares; so that the limit $\gamma(x)$ minimizes $E[m(x) - \tilde{\gamma}(x)]^2$ over $\mathcal{P}_q$. In this setting, the "residual" $m(x) - \gamma(x)$ will be uncorrelated with any element of $\mathcal{P}_q$. In particular, it will be uncorrelated with the element $\lambda(x)$ of $\mathcal{P}_q$ that best approximates the score $\ell(x)$, or minimizes $E[\ell(x) - \tilde{\lambda}(x)]^2$. Therefore, the average derivative bias can be refined as

$$E[\gamma'(x)] - E[m'(x)] = E\{[\ell(x) - \lambda(x)][\gamma(x) - m(x)]\}.$$

With least squares estimation, the derivative bias is the expected product of the biases from approximating $m(x)$ and $\ell(x)$ by the same method.[3]

Consider explicitly the case where $\mathcal{P}_q$ denotes polynomials of order q, with $q > 1$. For a linear model $m(x) = \alpha + \beta^T x$, we have $\gamma(x) = \alpha + \beta^T x$, so that the average derivative bias is zero. Alternatively, if the regressors are normal; $x \sim \mathcal{N}(\mu_x, \Sigma_x)$; then $\ell(x) = \Sigma_x^{-1}(x - \mu_x) = \lambda(x)$, so that there is no average derivative bias, regardless of the shape of the regression $m(x)$. Therefore, for the normal design of the simulation results in Table 5.1a,b, we should see less bias in average derivatives estimated using polynomial approximations.

With this in mind, Table 5.2a,b contains the results of simulating average derivatives estimated by polynomial regression. Specifically, this involves regressing y on a constant and all powers and cross products of x up to order 2 for a quadratic, and order 3 for a cubic. For $k = 4$ predictor variables, the quadratic polynomial has 15 coefficients to be estimated, and the cubic polynomial has 35 coefficients. The average derivative is then estimated by differentiating the fitted polynomial, and averaging the fitted derivative values across all data points. Consistent with our development above, these average derivative estimators are not

[3] See Stoker (1991c). This type of formula was used for polynomial expansions in Newey (1991), and the following results for polynomials were derived by Florens, Ivaldi and Larribeau-Nori (1991).

badly biased. However, they are more variable that the slope estimators based on kernels, given in Tables 5.1a,b.[4]

At any rate, these results give clear indications of how estimation results can vary with the choice of nonparametric method. Future research is definitely in order, to determine what methods work best for different categories of semiparametric applications.

Linear Model: $y_i = 1 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + \varepsilon_i; i = 1, \ldots, 100$

TABLE 5.2a:
POLYNOMIAL SIMULATION RESULTS — LINEAR MODEL
True Value: $\delta = (1, 1, 1, 1)$

	$\widehat{\delta}_1$	$\widehat{\delta}_2$	$\widehat{\delta}_3$	$\widehat{\delta}_4$
Quadratic	.969	.970	1.018	.991
	(.116)	(.112)	(.118)	(.114)
Cubic	1.008	.989	1.010	.990
	(.153)	(.172)	(.157)	(.173)

Means and Standard Deviations from 60 Monte Carlo Samples

Probit Model: $y_i = 1[1 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + \varepsilon_i > 0; i = 1, \ldots, 100$

TABLE 5.2b:
POLYNOMIAL SIMULATION RESULTS — PROBIT MODEL
True Value: $\delta == (.161, .161, .161, .161)$

	$\widehat{\delta}_1$	$\widehat{\delta}_2$	$\widehat{\delta}_3$	$\widehat{\delta}_4$
Quadratic	.167	.169	.170	.175
	(.036)	(.032)	(.036)	(.037)
Cubic	.158	.157	.155	.155
	(.055)	(.048)	(.054)	(.058)

Means and Standard Deviations from 60 Monte Carlo Samples

[4]The "direct slope" kernel estimator provides the proper comparison, because its leading term measures of $E(\partial x^T/\partial x)$, which is identically 1 when polynomial approximations are used.

IV. A Closing Note

One of the great advantages of giving a series of lectures on one subject is that one can cover material from many vantage points, applying one's own view as to what the interesting and important issues are. These lectures touched on many issues in semiparametric estimation in econometrics, from various theoretical results to some simple empirical examples. While the distribution theory for semiparametric methods is now quite extensive, I have completed this lecture with the work on small sample problems in derivative estimation, where different measurement methods give substantially different answers. The intention here was not to be overtly critical of existing first-order distributional theory, but rather to leave the correct impression that many interesting implementation questions remain open.

This brings me back to the main conclusion given above, that applications of semiparametric methods are the highest priority for future research. The enormous potential of these methods will naturally be associated with fascinating successes and surprising problems as a broad range of applications are undertaken. At any rate, these lectures will have served their purpose if they help stimulate work in *applied* semiparametric econometrics. The staying power and importance of the developments of the last decade will be determined by their success in empirical analysis.

APPENDIX:

PRACTICAL SPECIFICATIONS FOR THE EMPIRICAL ESTIMATION

This appendix gives specific details on how the estimators presented in Lectures 3 and 4 were computed. To begin with, all computations were performed using Gauss on a personal computer. Various other programs, such as XploRe (c.f. Härdle (1991)) could easily be used as well.

All nonparametric estimation uses the kernel estimators of the kind introduced in Lecture 3. All kernel functions are constructed as the product of biweight kernel functions: if the biweight kernel is given as

$$\kappa(u) = \left(\frac{15}{16}\right)(1-u^2)^2 1[|u|<1]$$

then the kernel $\mathcal{K}$ for a k-dimensional estimation is

$$\mathcal{K}(u_1,\ldots,u_k) = \Pi\kappa(u_i).$$

Positive kernels were used throughout, because of typically erratic behavior exhibited when higher order kernels (as prescribed by the theory of Lecture 3) are used in relatively small samples. Also, normal kernels were used in the simulations reported in Lecture 5, as consistent with the discussion there.

Bandwidths are chosen by Generalized Cross Validation (GCV) of Craven and Wahba (1979) as follows. For estimation of the regression $m(x)$, let Y denote the vector of observations $\{y_i\}$ and M_h denote the vector of values $\{\widehat{m}(x_i)\}$ computed with bandwidth h. Consider the weight matrix W_h defined from

$$M_h = W_h Y.$$

The GCV bandwidth is the value of h that minimizes

$$\frac{N^{-1}|(I-W_h)Y|^2}{[N^{-1}\mathrm{tr}\ (I-W_h)Y]^2}.$$

For estimating the density $f(x)$, the vector Y has the indicator $1[x = x_i]$ as i-th component, and redefines W_h to the local weights appropriate for density estimator. This method of choosing bandwidths was chosen for its simplicity. For adherence to the asymptotic theory, as N increases, the bandwidth would need to be shrunk in accordance with the conditions of the theorems, which would not be the same as performing GCV for all increased sample sizes. However, without an established theory of bandwidth selection for these estimators, we retain GCV because of ease of interpretation and simplicity of implementation.

For the average derivative estimators, trimming bounds were set to drop the observations with 5% lowest estimated density values for both applications, namely with the collision data and the Boston housing data. We have found that little difference occurs from setting trimming bounds in the 1-5% range. As above, a strict percentage trimming rule is not necessarily consistent with the asymptotic conditions in the technical analysis of the estimators, and we just use it here because it is simple to interpret in finite samples.

The predictor variables were scaled by their standard deviations prior to estimation for both of the empirical examples. For the Boston housing data analysis, we adjusted the average derivative estimators for the scaling, so that they would be directly comparable to the OLS estimates in Belsley, Kuh and Welsch (1980).

This sort of scaling and rescaling is consistent with the structure of average derivatives, as follows. In particular, suppose that the observations on the predictors x were transformed to $z = Ax$, where A is a nonsingular matrix. Note that the regression $E(y \mid x) = m(x) = m(A^{-1}z) = \tilde{m}(z) = E(y \mid z)$. If the average derivative of y with respect to z is denoted $\delta_z = E(\tilde{m}')$ and the average derivative of y with respect to x is $\delta = E(m')$, then by the chain rule, we have that

$$\delta = A\delta_z.$$

Scaling with regard to standard deviations corresponds to a diagonal A matrix, and an estimate of δ can be obtained by rescaling the estimate of the average derivatives from the scaled data, as above. Another scaling rule could be used to diagonalize the covariance structure of the predictors, say by using $A = S_{xx}^{-1/2}$,

where S_{xx} is the sample covariance matrix of x, but we have not used this here. The main practical arguments for scaling and/or diagonalizing the basic data are based on smoothing with data whose distribution is approximately elliptically symmetric. The arguments against such transformations are based on data with pronounced multimodal structure.

We have used indirect slope estimators for average derivatives. The variances are estimated by mimicking the U statistic structure of the statistics, as discussed before. In particular, the variance estimator proposed in Härdle and Stoker(1989) could be used, however we use a modification that is resilient to proportional smoothing bias in the same way as the slope estimator is. Apply the standard manipulations for deviations of linear coefficients to the indirect slope estimator

$$\widehat{d}_{\mathrm{ind}} = S_{\ell x}^{-1} S_{\ell y},$$

giving

$$\widehat{d}_{\mathrm{ind}} - \delta = S_{\ell x}^{-1}[N^{-1}\sum \widehat{\ell}(x_i)\widehat{1}_i u_i]$$

where $u_i = (y_i - \overline{y}) - (x_i - \overline{x})^T\delta$. The last term estimates the average derivative of u_i, and so apply the Härdle-Stoker variance formulation with $\widehat{u}_i = (y_i - \overline{y}) - (x_i - \overline{x})^T\widehat{d}_{\mathrm{ind}}$ used in place of y_i. Therefore, we compute

$$r_{ui} = \widehat{\ell}(x_i)\widehat{1}_i\widehat{u}_i + N^{-1}h^{-k}\sum_{j=1}^{N}\left[h^{-1}\mathcal{K}'\left(\frac{x_i - x_j}{h}\right) - \mathcal{K}\left(\frac{x_i - x_j}{h}\right)\widehat{\ell}(x_j)\right]\frac{\widehat{1}_j\widehat{u}_j}{\widehat{f}(x_j)}$$

and estimate the asymptotic covariance matrix of $\widehat{d}_{\mathrm{ind}}$ as

$$S_{\ell x}^{-1} S_{r_u r_u} S_{\ell x}^{-1}$$

where $S_{r_u r_u}$ is the sample covariance matrix of $\{r_{ui}\}$. Notice that if $\widehat{\ell}(x)$ is proportionately too small because of smoothing bias, $\widehat{d}_{\mathrm{ind}}$ and this variance estimate will not be affected by it.

The variance of the estimate $\widehat{\gamma}$ for the specification tests is likewise determined. The procedure for mimicking the U-statistic structure is to compute an estimate of the influence term of $\widehat{\gamma}$. The influence consists of two kinds of terms. The first term is appropriate if the functions m and G, and the parameters δ are known, associated with the (Eicher–White) heteroskedasticity consistent variance estimates. The second set of terms represents the adjustments for the estimation of m, G and δ, and are constructed from the U-statistic structure of the

adjustments. The "generic singularity" discussed in Lecture 4 is manifested by a particular relation between these terms under particular asymptotic conditions. In particular, under those conditions the adjustment terms exactly cancel out the first term corresponding to setting where the functions were known. Our posture is to base inference on the adjusted estimates, implicitly assuming that the U-statistic structure of the adjustments accurately reflects the variation of the regression coefficient $\widehat{\gamma}$ in small samples.

For testing an index model $E(y \mid x) = G(x^T\delta)$ against a general regression $E(y \mid x) = m(x)$, recall that the basic regression equation is denoted as

$$y - \widehat{G}(x_i^T\widehat{\delta}) = \widehat{\alpha} + \widehat{\gamma}\widehat{m}(x_i) + \widehat{u}_i,$$

where $\widehat{\gamma}$ denotes the slope coefficient, $\widehat{u}_i$ the residual, where we only include observations with $\widehat{1}_i = 1[\widehat{f}(x_i) > b] = 1$. We use the following notation:

$$\begin{array}{lll}
z_i & = x_i^T\widehat{\delta} & \\
m_i & = \widehat{m}(x_i) & ; \text{ Kernel regression of } y \text{ on } x \\
f_i & = \widehat{f}(x_i) & ; \text{ Kernel density of } x \\
\mathcal{K}_x & & ; \text{ Kernel function used to compute } \widehat{m}(x_i) \text{ and } \widehat{f}(x_i) \\
h_x & & ; \text{ Bandwidth used to compute } \widehat{m}(x_i) \text{ and } \widehat{f}(x_i) \\
\tilde{m}_i & = m_i - \overline{m};\ m = (\sum m_i)/(\sum \widehat{1}_i) & \\
S_m & = (\sum \tilde{m}_i^2)/(\sum \widehat{1}_i) & \\
G_i & = \widehat{G}(z_i) & ; \text{ Kernel regression of } y \text{ on } z \\
G_i' & = \widehat{G}'(z_i) & ; \text{ The derivative of } G \text{ evaluated at } z_i \\
f_{zi} & = \widehat{f}_{zi}(z_i) & ; \text{ Kernel density of } z \\
\mathcal{K}_z & & ; \text{ Kernel function used to compute } \widehat{G}(z_i) \text{ and } \widehat{f}_{zi}(zi) \\
h_z & & ; \text{ Bandwidth used to compute } \widehat{G}(z_i) \text{ and } \widehat{f}_{zi}(z_i).
\end{array}$$

We estimate the influence function of the coefficient $\widehat{\gamma}$ as

$$r_{\gamma i} = \frac{1}{S_m}\left(\tilde{m}_i\widehat{u}_i + \widehat{r}_{Ri} - \widehat{r}_{Li}\right)$$

where $\widehat{r}_{Ri}$ is the adjustment for the estimation of the regression m ("Right" side) and r_{Li} is the adjustment for the estimation of G and δ ("Left" side). The formulae

for these terms are

$$\begin{aligned}
\widehat{r}_{Ri} = {} & N^{-1}h_x^{-k}\sum_j \mathcal{K}_x\left(\frac{x_i - x_j}{h_x}\right)\left(\frac{y_j(y_i - G_i)}{f_i} + \frac{y_i(y_j - G_j)}{f_j}\right) \\
& - N^{-1}h_x^{-k}\sum_j \mathcal{K}_x\left(\frac{x_i - x_j}{h_x}\right)\left(\frac{m_i(y_i - G_i)}{f_i} + \frac{m_j(y_j - G_j)}{f_j}\right)
\end{aligned}$$

$$\begin{aligned}
\widehat{r}_{Li} = {} & N^{-1}h_z^{-1}\sum_j \mathcal{K}_z\left(\frac{z_i - z_j}{h_z}\right)\left(\frac{y_j\tilde{m}_i}{f_{zi}} + \frac{y_i\tilde{m}_j)}{f_{zj}}\right) \\
& - N^{-1}h_z^{-1}\sum_j \mathcal{K}_z\left(\frac{z_i - z_j}{h_z}\right)\left(\frac{G_i\tilde{m}_i}{f_{zi}} + s\frac{G_j\tilde{m}_j)}{f_{zj}}\right) \\
& + B_1 r_{ui}
\end{aligned}$$

where B_1 is

$$B_1 = N^{-1}\sum_i \widehat{D}_1\tilde{m}_i$$

and

$$\begin{aligned}
\widehat{D}_1 = {} & S_1(x_i)^{-1}\sum_{j=1}^{N}\frac{x_i - x_j}{h_z}\mathcal{K}_z'\left(\frac{z_i - z_j}{h_z}\right)y_j \\
& - G(z_i)S_1(x_i)^{-1}\sum_{j=1}^{N}\frac{x_i - x_j}{h_z}\mathcal{K}_z'\left(\frac{z_i - z_j}{h_z}\right) \\
S_1(x_i) = {} & \left[\sum_{j=1}^{N}\mathcal{K}_z\left(\frac{z_i - z_j}{h_z}\right)\right].
\end{aligned}$$

The variance of $\widehat{\gamma}$ is then estimated by $\widehat{\sigma}_\gamma/N$, where $\widehat{\sigma}_\gamma$ is the sample variance of $\widehat{r}_{\gamma i}$. Therefore, the "t-statistic" is formed as

$$t = \frac{\sqrt{N}\widehat{\gamma}}{\sqrt{\widehat{\sigma}_\gamma}}.$$

We base inference on the square of this statistic, as compared to a $\chi^2(1)$ critical value.

Test statistics for comparing partial index models are all computed in the same manner, namely by computing the proper adjustment terms for estimated functions and parameters. Rodriguez and Stoker (1992) give the general formulae and derivations for all the partial index models considered here.

REFERENCES

Abe, M. (1991), "A Marketing Mix Model Developed from Single Source Data: A Semiparametric Approach," doctoral dissertation, Department of Physics, MIT.

Ahn, H. and J.L. Powell (1990), "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," draft, Department of Economics, University of Wisconsin at Madison.

Ai, C. (1991a), "Efficient Semiparametric Estimations of Generalized Regression Models," doctoral dissertation, Department of Economics, MIT.

Ai, C. (1991b), "The Regression Based Estimation Method of Index Model," draft, Department of Economics, State University of New York at Stonybrook.

Amemiya, T. (1985), *Advanced Econometrics*, Cambridge, MA, Harvard University Press.

Andrews, D.W.K. (1989), "Asymptotics for Semiparametric Econometric Models: I. Estimation," Working Paper, Cowles Foundation, Yale University.

Bassett, G. and R. Koenker (1978), "Asymptotic Theory of Least Absolute Error Regression," *Journal of the American Statistical Association*, 73, 667–677.

Belsley, D.A., E. Kuh and R.E. Welsch (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York, Wiley.

Bierens, H.J. (1990), "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, 58, 1443–1458.

Bickel, P.J. and K.A. Doksum (1977), *Mathematical Statistics: Basic Ideas and Selected Topics*, San Francisco, Holden-Day.

Bierens, H.J. and H.A. Pott-Butler (1991), "Specification of Household Engel Curves by Nonparametric Regression," *Econometric Reviews*, 9, 123–184.

Box, G.E.P. and D.R. Cox (1964): "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Series B, 26, 211–252.

Brillinger, D.R. (1983), "A Generalized Linear Model with 'Gaussian' Regressor Variables," in P.J. Bickel, K.A. Doksum and J.l. Hodges, eds., *A Festschrift for Erich L. Lehmann*, Woodsworth International Group, Belmont, CA.

Buckley, J. and I. James (1979), "Linear Regression with Censored Data," *Biometrika*, 66, 429–436.

Carroll, R.J. (1982), "Adapting for Heteroskedasticity in Linear Models," *Annals of Statistics*, 10, 1224–1233.

Chamberlain, G. (1986a), "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics*, 32, 189–218.

Chamberlain, G. (1986b), "Notes on Semiparametric Regression," draft, Department of Economics, University of Wisconsin.

Chamberlain, G. (1991), "Quantile Regression, Censoring and the Structure of Wages," draft, Harvard University.

Chung, C-F and A.S. Goldberger (1984), "Proportional Projections in Limited Dependent Variables Models," *Econometrica*, 52, 531–534.

Cosslett, S.R. (1983), "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica*, 51, 765–782.

Cowing, T.G. and D.L. McFadden (1984), *Microeconomic Modeling and Policy Analysis*, New York, Academic Press.

Cox D.R. (1975), "Partial Likelihood," *Biometrika*, 62, 269–276

Craven, P. and G. Wahba (1979), "Smoothing Noisy Data With Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerische Mathematik*, 31, 377–403.

Deaton, A.S. (1989), "Rice Prices and Income Distribution in Thailand: A Nonparametric Analysis," *Economic Journal*, 99, 1-37.

Delgado, M.A. and P.M. Robinson (1991), "Nonparametric Methods and Semiparametric Methods for Economic Research," draft, London School of Economics.

Engle, R.F., C.W.J. Granger, J. Rice, and A. Weiss (1986), " Semiparametric Estimates of the Relation Between Weather and Electricity Sales." *Journal of the American Statistical Association*, 81, 310–320.

Florens, J-P., M. Ivaldi and S. Larribeau (1991), "Sobolev Estimation of Approximate Regressions," draft, July, Université des Sciences Sociales, Toulouse.

Friedman, J.H. and W. Stuetzle (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823.

Gallant, A.R., D. Hsieh and G. Tauchen (1991), "On Fitting a Recalcitrant Series: The Pound/Dollar Exchange Rate," in W. Barnett, J. Powell, G. Tauchen eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge, Cambridge University Press.

Gallant, A.R. and D.W. Nychka (1987), "Semi-nonparametric Maximum Likelihood Estimation," *Econometrica*, 55, 363–390.

Goldstein, L. and K. Messer (1990), "Optimal Plug-in Estimators for Nonparametric Functional Estimation," Technical Report No. 277, Stanford University.

Han, A.K. (1987), "Maximum Rank Correlation Estimator and Generalized Median Estimator in Censored Regression and Survival Models," *Journal of Econometrics*, 35, 303–316.

Han, A.K. and J.A. Hausman (1990), "Flexible Parametric Estimation of Duration and Competing Risk Models," *Journal of Applied Econometrics*, 5, 1–28.

Härdle, W. (1991), *Applied Nonparametric Regression*, Cambridge, Cambridge University Press (Econometric Society Monographs).

Härdle, W., P. Hall and H. Ichimura (1991), "Optimal Smoothing in Single Index Models," Discussion Paper No. 9107, CORE, Université Catholique de Louvain.

Härdle, W., J. Hart, J.S. Marron and A.B. Tsybakov (1990), "Bandwidth Choice for Average Derivative Estimation," draft, Universität Bonn.

Härdle W., W. Hildenbrand and M. Jerison (1991), "Empirical Evidence for the Law of Demand," *Econometrica*, 59, 1525–1550.

Härdle, W. and M Jerison (1989), "Evolution of Engel Curves over Time," draft, Universität Bonn.

Härdle, W. and Stoker, T.M.(1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986–995.

Harrison, D. and D.L. Rubinfeld, (1978a), "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management*, 5, 81–102.

Harrison, D. and D.L. Rubinfeld, (1978b), "The Distribution of Benefits from Improvements in Urban Air Quality", *Journal of Environmental Economics and Management*, 5, 313–332.

Hastie, T.J. and R.J. Tibshirani (1990), *Generalized Additive Models*, London, Chapman and Hall.

Hausman, J.A. and W.K. Newey (1990), "Nonparametric Estimation of Exact Consumer Surplus and Deadweight Loss," Working Paper, Department of Economics, MIT, October.

Heckman, J.J. and V. J. Hotz (1989), "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862–874.

Heckman, J.J. and B. Singer (1984), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52, 271–320.

Hildenbrand, K and W. Hildenbrand (1986), "On the Mean Income Effect: A Data Analysis of the U.K. Family Expenditure Data," in W. Hildenbrand and A. Mas-Colell, eds., *Contributions to Mathematical Economics*, Amsterdam, North-Holland.

Hoeffding, W. (1948), "A Class of Statistics with Asymptotically Normal Distribution," *Annals of Mathematical Statistics*, 19, 293–325.

Honore, B.E. (1991), "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," draft, Department of Economics, Northwestern University.

Horowitz, J.L. (1990), "A Smoothed Maximum Score for the Binary Response Model," draft, Department of Economics, Department of Iowa.

Horowitz, J.L. and G.R. Neumann (1987), "Semiparametric Estimation of Employment Duration Models," *Econometric Reviews*, 6, 1–40.

Huber, P.J. (1981), *Robust Statistics*, New York, Wiley.

Ichimura, H. (1986), "Estimation of Index Model Coefficients," doctoral dissertation, Department of Economics, MIT.

Kalbfleisch, J.D. and R.L. Prentice (1980), *Statistical Analysis of Failure Time Data*, New York, Wiley.

Kallieris, D., R. Mattern, and W. Härdle (1989), "Verhalten des EUROSID beim 90° Seitenaufprall im Vergleich zu PMTO sowie US-SID, HYBRID II und APROD," Forschungsvereinigung Automobiltechnik e.V., FAT Schriftenreihe, 79, Frankfurt am Main.

Kim, J. and D. Pollard (1990), "Cube Root Asymptotics," *Annals of Statistics*, 18, 191-219.

Klein, R.W. and R.S. Spady (1990), 'An Efficient Semiparametric Estimator of the Binary Response Model," draft, Bell Communications Research.

Lau, L.J. (1986), "Functional Forms in Econometric Model Building," in Z. Griliches and M.D. Intrilligator, eds., *Handbook of Econometrics, Volume 3*, Amsterdam, North Holland.

Lehman, E.L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco, Holden-Day.

Lewbel, A. (1991), "Consistent Tests of Nonparametric Regression and Density Restrictions," draft, Department of Economics, Brandeis University.

Li, K.C. and N. Duan(1989) "Regression Analysis Under Link Violation," *Annals of Statistics*, 17, 1009–1052.

Li, K.C. (1991) "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–327.

Manski, C. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205–228.

Manski, C.F. (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, 313–333.

Manski, C.F. (1987), "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55, 357-362.

Manski, C.F. (1988a), *Analog Estimation in Econometrics*, London, Chapman and Hall.

Manski, C.F. (1988b), "Identification of Binary Response Models," *Journal of the American Statistical Association*, 83, 729–738.

Matzkin, R.L. (1990), "Least Concavity and the Distribution Free Estimation of Nonparametric Concave Functions," Discussion Paper No. 958, Cowles Foundation, Yale University.

McCullagh, P. and J.A. Nelder (1983), *Generalized Linear Models*, London, Chapman and Hall.

Meyer, B. (1987), "Semiparametric Estimation of Duration Models," doctoral dissertation, Department of Economics, MIT.

Mosteller, F. and J.W. Tukey (1977), *Data Analysis and Regression*, Reading, MA, Addison-Wesley.

Mroz, T.A. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55, 765–800.

Newey, W.K. (1986), "Maximum Likelihood Specification Testing and Conditional Moment Tests," *Econometrica*, 53, 1047–70.

Newey, W.K. (1988), "Two Step Series Estimation of Sample Selection Models," draft, Princeton University.

Newey, W.K. (1989), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–136.

Newey, W.K. (1991), "The Asymptotic Variance of Semiparametric Estimators," Working Paper No. 583, Department of Economics, MIT, revised July.

Newey, W.K., J.L. Powell and J.R. Walker (1990), "Semiparametric Estimation of Selection Models," *American Economic Review, Papers and Proceedings*, 80, 324–328.

Newey, W.K. and P.A. Ruud (1991), "Density Weighted Least Squares Estimation," draft, Department of Economics, MIT.

Newey, W.K. and T.M. Stoker (1989), "Efficiency Properties of Average Derivative Estimators," draft, Sloan School of Management, MIT.

Pagan, A.R. and F. Vella (1989), "Diagnostic Tests for Models Based on Individual Data: A Survey," *Journal of Applied Econometrics, 4*, S29–S59.

Pagan, A.R. and M.R. Wickens (1989), "A Survey of Some Recent Econometric Methods," *Economic Journal*, 99, 962–1025.

Powell, J.L. (1984), "Least Absolute Deviations Estimation for the Censored Regression Model," *Journal of Econometrics*, 25, 303–325.

Powell, J.L., (1986a), "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, 54, 1435-1460.

Powell, J.L. (1986a), "Censored Regression Quantiles," *Journal of Econometrics*, 32, 143–155.

Powell, J.L. (1987a), "Comments on 'Semiparametric Estimation of Employment Duration Models' by J.L. Horowitz and G.R. Neumann," *Econometric Reviews*, 6, 41–54.

Powell, J.L. (1987b), "Semiparametric Estimation of Bivariate Latent Variables Models," draft, University of Wisconsin.

Powell, J.L. (1991): "Estimation of Monotonic Regression Models Under Conditional Quantile Restrictions," in W. Barnett, J. Powell, G. Tauchen eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge, Cambridge University Press.

Powell, J.L., J.H. Stock and T.M. Stoker (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430.

Rao, C.R. (1973), *Linear Statistical Inference and Its Applications*, New York, Wiley.

Ritov, Y. (1990), "Estimation in a Linear Regression Model with Censored Data," *Annals of Statistics*, 18, 303–328.

Robinson, P.M. (1987), "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica*, 55, 875–891.

Robinson, P.M. (1988a), "Semiparametric Econometrics: A Survey," *Journal of Applied Econometrics*, 3, 35–51.

Robinson, P.M. (1988b), "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

Robinson, P.M. (1989), "Hypothesis Testing in Nonparametric and Semiparametric Models for Economic Time Series," *Review of Economic Studies*, 56, 511–534.

Robinson, P.M. (1991), "Semiparametric Methods for Time Series," Discussion Paper No. R.39, London School of Economics.

Rodriguez, D. and T.M. Stoker (1992), "Semiparametric Analysis of Environmental Effects," draft, Sloan School of Management, MIT.

Rosen, S. (1974) "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 82, 34–55.

Ruud, P.A. (1983), "Sufficient Conditions for the Consistency of Maximum Likelihood Despite Misspecification of Distribution," *Econometrica*, 51, 225–228.

Ruud, P.A. (1984), "Tests of Specification in Econometrics," *Econometric Reviews*, 3, 211–242.

Ruud, P.A. (1986), "Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution," *Journal of Econometrics*, 32, 157–187.

Samarov, A.M.(1990), "Exploring Regression Structure Using Nonparametric Functional Estimation," *Proceedings of the NATO Advanced Study Institute on Nonparametric Functional Estimation and Related Topics*, Spetses, Greece.

Schmitz, H-P. (1989), Die Zeitliche Invarianz von Einkommensvertielungen, Eine Analyse der Einkommensverteilungen in Grossbritanien 1968-1983, Inaugural-Dissertation, Universitat Bonn.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, New York, Wiley.

Severini, T.A. and W.H. Wong (1987), "Profile Likelihood and Semiparametric Models," draft, University of Chicago.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, London, Chapman and Hall.

Stock, J.H. (1989), "Nonparametric Policy Analysis," *Journal of the American Statistical Association*, 84, 567–578.

Stoker, T.M. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461–1481.

Stoker, T.M. (1989), "Tests of Additive Derivative Constraints," *Review of Economic Studies*, 56, 535–552.

Stoker, T.M. (1991a), "Equivalence of Direct, Indirect and Slope Estimators of Average Derivatives," in Barnett, W.A., J.L. Powell and G.Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge, Cambridge University Press.

Stoker, T.M. (1991b), "Smoothing Bias in the Estimation of Density Derivatives," Working Paper, Sloan School of Management, MIT, August.

Stoker, T.M. (1991c), "Smoothing Bias in the Measurement of Marginal Effects," Working Paper, Sloan School of Management Working Paper, MIT, August.

Stoker, T.M. and J.M. Villas-Boas (1992), "Monte Carlo Simulation of Average Derivative Estimators", draft, Sloan School of Management, MIT.

Stone, C.J. (1980), "Optimal Rates of Convergence for Nonparametric Estimators," *Annals of Statistics*, 8, 1348–1360.

Stone, C.J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *Annals of Statistics*, 10, 1040–53.

Stone, C.J. (1986), "The Dimensionality Reduction Principle for Generalized Additive Models," *Annals of Statistics*, 14, 590–606.

Tinbergen, J. (1956) "On the Theory of Income Distribution," *Weltwirtschaftliche Archiv*, 77, 155–173.

Tsiatis, A.A. (1981), "A Large Sample Study of Cox's Regression Model," *Annals of Statistics*, 9, 93–108.

Tukey, J.W. (1977) *Exploratory Data Analysis*, Reading, MA, Addison-Wesley.

Wooldridge, J. (1990), "A Test for Functional Form Against Nonparametric Alternatives," draft, Department of Economics, MIT.

Thomas M. Stoker is Professor of Applied Economics at the Sloan School of Management, Massachusetts Institute of Technology. His research spans the areas of consumer demand and welfare measurement, aggregation over economic agents and semiparametric econometrics, and his articles have appeared in various journals, including *American Economic Review*, *Econometrica*, *Journal of Political Economy* and *Journal of the American Statistical Association*. He currently serves on the editorial boards of *Econometrica*, *Journal of the American Statistical Association*, *Journal of Business and Economic Statistics* and *Journal of Applied Econometrics*. He has held visiting appointments at Universität Bonn, and at Nuffield College, Oxford University. His recent teaching includes courses in microeconomics, econometrics, statistics and industrial organization.

The series will discuss the semiparametric approach to econometric modeling, as it has evolved for survey applications (independently distributed observations). The material will focus on three main themes: i) relative merits of semiparametric approaches, ii) precision measurement and other statistical features, and iii) the use of semiparametric methods for model assessment. Standard models of limited dependent variables (discrete choice, censored or truncated samples, and duration processes) will be discussed in the context of index model restrictions. Average derivative estimators of index model coefficients provide the main example for general statistical results and practical issues. Certain specific applications will be discussed, emphasizing interpretation of estimates and graphical methods.