Horowitz, Joel L.

**Working Paper**

# Semiparametric models

Papers, No. 2004,17

**Provided in Cooperation with:**
CASE - Center for Applied Statistics and Economics, Humboldt University Berlin

*Suggested Citation:* Horowitz, Joel L. (2004) : Semiparametric models, Papers, No. 2004,17, Humboldt-Universität zu Berlin, Center for Applied Statistics and Economics (CASE), Berlin

This Version is available at:
https://hdl.handle.net/10419/22191

# SEMIPARAMETRIC MODELS

by

Joel L. Horowitz

Department of Economics

Northwestern University

Evanston, IL 60208-2600

U.S.A.

Research supported in part by NSF Grant SES-9910925.

# SEMIPARAMETRIC MODELS

## 1. Introduction

Much empirical research is concerned with estimating conditional mean, median, or hazard functions. For example, labor economists are interested in estimating the mean wages of employed individuals conditional on characteristics such as years of work experience and education. The most frequently used estimation methods assume that the function of interest is known up to a set of constant parameters that can be estimated from data. Models in which the only unknown quantities are a finite set of constant parameters are called *parametric*. The use of a parametric model greatly simplifies estimation, statistical inference, and interpretation of the estimation results but is rarely justified by theoretical or other *a priori* considerations. Estimation and inference based on convenient but incorrect assumptions about the form of the conditional mean function can be highly misleading.

As an illustration, the solid line in Figure 1 shows an estimate of the mean of the logarithm of weekly wages, $\log W$, conditional on years of work experience, *EXP*, for white males with 12 years of education who work full time and live in urban areas of the North Central U.S. The estimate was obtained by applying kernel nonparametric regression (see, e.g., Härdle 1990, Fan and Gijbels 1996) to data from the 1993 Current Population Survey (CPS). The estimated conditional mean of $\log W$ increases steadily up to approximately 30 years of experience and is flat thereafter. The dashed and dotted lines in Figure 1 show two parametric estimates of the mean of the logarithm of weekly wages conditional on years of work experience. The dashed line is the ordinary least squares (OLS) estimate that is obtained by assuming that the mean of $\log W$ conditional on $EXP$ is the linear function $\boldsymbol{E}(\log W \mid EXP) = \beta_0 + \beta_1 EXP$. The dotted line is the OLS estimate that is obtained by assuming that $\boldsymbol{E}(\log W \mid EXP)$ is the quadratic function $\boldsymbol{E}(\log W \mid EXP) = \beta_0 + \beta_1 EXP + \beta_2 EXP^2$. The nonparametric estimate (solid line) places no restrictions on the shape of $\boldsymbol{E}(\log W \mid EXP)$. The linear and quadratic models give misleading estimates of $\boldsymbol{E}(\log W \mid EXP)$. The linear model indicates that $\boldsymbol{E}(\log W \mid EXP)$ increases steadily as experience increases. The quadratic model indicates that $\boldsymbol{E}(\log W \mid EXP)$ decreases after 32 years of experience. In contrast, the nonparametric estimate of $\boldsymbol{E}(\log W \mid EXP)$ becomes nearly flat at approximately 30 years of experience. Because the nonparametric estimate does not restrict the conditional mean function to be linear or quadratic, it is more likely to represent the true conditional mean function.

The opportunities for specification error increase if *Y* is binary. For example, consider a model of the choice of travel mode for the trip to work. Suppose that the available modes are

automobile and transit. Let $Y = 1$ if an individual chooses automobile and $Y = 0$ if the individual chooses transit. Let $X$ be a vector of explanatory variables such as the travel times and costs by automobile and transit. Then $\boldsymbol{E}(Y \mid x)$ is the probability that $Y = 1$ (the probability that the individual chooses automobile) conditional on $X = x$. This probability will be denoted $\boldsymbol{P}(Y = 1 \mid x)$. In applications of binary response models, it is often assumed that $\boldsymbol{P}(Y \mid x) = G(\beta' x)$, where $\beta$ is a vector of constant coefficients and $G$ is a known probability distribution function. Often, $G$ is assumed to be the cumulative standard normal distribution function, which yields a *binary probit* model, or the cumulative logistic distribution function, which yields a *binary logit* model. The coefficients $\beta$ can then be estimated by the method of maximum likelihood (Amemiya 1985). However, there are now two potential sources of specification error. First, the dependence of $Y$ on $x$ may not be through the linear *index* $\beta' x$. Second, even if the index $\beta' x$ is correct, the *response function* $G$ may not be the normal or logistic distribution function. See Horowitz (1993a, 1998) for examples of specification errors in binary response models and their consequences.

Many investigators attempt to minimize the risk of specification error by carrying out a *specification search* in which several different models are estimated and conclusions are based on the one that appears to fit the data best. Specification searches may be unavoidable in some applications, but they have many undesirable properties and their use should be minimized. There is no guarantee that a specification search will include the correct model or a good approximation to it. If the search includes the correct model, there is no guarantee that it will be selected by the investigator's model selection criteria. Moreover, the search process invalidates the statistical theory on which inference is based.

The rest of this chapter describes methods that deal with the problem of specification error by relaxing the assumptions about functional form that are made by parametric models. The possibility of specification error can be essentially eliminated through the use of nonparametric estimation methods. They assume that the function of interest is smooth but make no other assumptions about its shape or functional form. However, nonparametric methods have important disadvantages that seriously limit their usefulness in applications. One important problem is that the precision of a nonparametric estimator decreases rapidly as the dimension of the explanatory variable $X$ increases. This phenomenon is called the *curse of dimensionality*. As a result of it, impracticably large samples are usually needed to obtain acceptable estimation precision if $X$ is multidimensional, as it often is in applications. For example, a labor economist

may want to estimate mean log wages conditional on years of work experience, years of education, and one or more indicators of skill levels, thus making the dimension of $X$ at least 3.

Another problem is that nonparametric estimates can be difficult to display, communicate, and interpret when $X$ is multidimensional. Nonparametric estimates do not have simple analytic forms. If $X$ is one- or two-dimensional, then the estimate of the function of interest can be displayed graphically as in Figure 1, but only reduced-dimension projections can be displayed when $X$ has three or more components. Many such displays and much skill in interpreting them can be needed to fully convey and comprehend the shape of an estimate.

A further problem with nonparametric estimation is that it does not permit extrapolation. For example, in the case of a conditional mean function it does not provide predictions of $\boldsymbol{E}(Y \mid x)$ at points $x$ that are outside of the support (or range) of the random variable $X$. This is a serious drawback in policy analysis and forecasting, where it is often important to predict what might happen under conditions that do not exist in the available data. Finally, in nonparametric estimation, it can be difficult to impose restrictions suggested by economic or other theory. Matzkin (1994) discusses this issue.

Semiparametric methods offer a compromise. They make assumptions about functional form that are stronger than those of a nonparametric model but less restrictive than the assumptions of a parametric model, thereby reducing (though not eliminating) the possibility of specification error. Semiparametric methods permit greater estimation precision than do nonparametric methods when $X$ is multidimensional. They are easier to display and interpret than nonparametric ones and provide limited capabilities for extrapolation and imposing restrictions derived from economic or other theory models. Section 2 of this chapter describes some semiparametric models for conditional mean functions. Section 3 describes semiparametric estimators for an important class of hazard models. Section 4 is concerned with semiparametric estimation of a certain binary response model.

## 2. Semiparametric Models for Conditional Mean Functions

The term *semiparametric* refers to models in which there is an unknown function in addition to an unknown finite dimensional parameter. For example, the binary response model $\boldsymbol{P}(Y = 1 \mid x) = G(\beta' x)$ is semiparametric if the function $G$ and the vector of coefficients $\beta$ are both treated as unknown quantities. This section describes two semiparametric models of conditional mean functions that are important in applications. The section also describes a related class of models that has no unknown finite-dimensional parameters but, like semiparametric models, mitigates the disadvantages of fully nonparametric models. Finally, this section

describes a class of transformation models that is important in estimation of hazard functions among other applications. Powell (1994) discusses additional semiparametric models.

*2.1 Single Index Models*

In a semiparametric single index model, the conditional mean function has the form

(2.1) $\boldsymbol{E}(Y \mid x) = G(\beta' x)$,

where $\beta$ is an unknown constant vector and $G$ is an *unknown* function. The quantity $\beta' x$ is called an *index*. The inferential problem is to estimate $G$ and $\beta$ from observations of $(Y, X)$. $G$ in (2.1) is analogous to a link function in a generalized linear model, except in (2.1) $G$ is unknown and must be estimated.

Model (2.1) contains many widely used parametric models as special cases. For example, if $G$ is the identity function, then (2.1) is a linear model. If $G$ is the cumulative normal or logistic distribution function, then (2.1) is a binary probit or logit model. When $G$ is unknown, (2.1) provides a specification that is more flexible than a parametric model but retains many of the desirable features of parametric models, as will now be explained.

One important property of single index models is that they avoid the curse of dimensionality. This is because the index $\beta' x$ aggregates the dimensions of $x$, thereby achieving *dimension reduction*. Consequently, the difference between the estimator of $G$ and the true function can be made to converge to zero at the same rate that would be achieved if $\beta' x$ were observable. Moreover, $\beta$ can be estimated with the same rate of convergence that is achieved in a parametric model. Thus, in terms of the rates of convergence of estimators, a single index model is as accurate as a parametric model for estimating $\beta$ and as accurate as a one-dimensional nonparametric model for estimating $G$. This dimension reduction feature of single index models gives them a considerable advantage over nonparametric methods in applications where $X$ is multidimensional and the single index structure is plausible.

A single-index model permits limited extrapolation. Specifically, it yields predictions of $\boldsymbol{E}(Y \mid x)$ at values of $x$ that are not in the support of $X$ but are in the support of $\beta' X$. Of course, there is a price that must be paid for the ability to extrapolate. A single index model makes assumptions that are stronger than those of a nonparametric model. These assumptions are testable on the support of $X$ but not outside of it. Thus, extrapolation (unavoidably) relies on untestable assumptions about the behavior of $\boldsymbol{E}(Y \mid x)$ beyond the support of $X$.

Before $\beta$ and $G$ can be estimated, restrictions must be imposed that insure their identification. That is, $\beta$ and $G$ must be uniquely determined by the population distribution of ($Y$, $X$). Identification of single index models has been investigated by Ichimura (1993) and, for the special case of binary response models, Manski (1988). It is clear that $\beta$ is not identified if $G$ is a constant function or there is an exact linear relation among the components of $X$ (perfect multicollinearity). In addition, (2.1) is observationally equivalent to the model $\boldsymbol{E}(Y \mid X) = G^*(\gamma + \delta\beta' x)$, where $\gamma$ and $\delta \neq 0$ are arbitrary and $G^*$ is defined by the relation $G^*(\gamma + \delta v) = G(v)$ for all $v$ in the support of $\beta' X$. Therefore, $\beta$ and $G$ are not identified unless restrictions are imposed that uniquely specify $\gamma$ and $\delta$. The restriction on $\gamma$ is called *location normalization* and can be imposed by requiring $X$ to contain no constant (intercept) component. The restriction on $\delta$ is called *scale normalization*. Scale normalization can be achieved by setting the $\beta$ coefficient of one component of $X$ equal to one. A further identification requirement is that $X$ must include at least one continuously distributed component whose $\beta$ coefficient is non-zero. Horowitz (1998) gives an example that illustrates the need for this requirement. Other more technical identification requirements are discussed by Ichimura (1993) and Manski (1988).

The main estimation challenge in single index models is estimating $\beta$. Given an estimator $b_n$ of $\beta$, $G$ can be estimated by carrying out the nonparametric regression of $Y$ on $b_n' X$ (e.g, by using kernel estimation). Several estimators of $\beta$ are available. Ichimura (1993) describes a nonlinear least squares estimator. Klein and Spady (1993) describe a semiparametric maximum likelihood estimator for the case in which $Y$ is binary. These estimators are difficult to compute because they require solving complicated nonlinear optimization problems. Powell, *et al*. (1989) describe a *density-weighted average derivative estimator* (DWADE) that is non-iterative and easily computed. The DWADE applies when all components of $X$ are continuous random variables. It is based on the relation

(3.2) $$\beta \propto \boldsymbol{E}[p(X)\partial G(\beta' X)/\partial X] = -2\boldsymbol{E}[Y \partial p(X)/\partial X],$$

where $p$ is the probability density function of $X$ and the second equality follows from integrating the first by parts. Thus, $\beta$ can be estimated up to scale by estimating the expression on the right-hand side of the second equality. Powell, *et al*. (1989) show that this can be done by replacing $p$ with a nonparametric estimator and replacing the population expectation $\boldsymbol{E}$ with a sample average. Horowitz and Härdle (1996) extend this method to models in which some components of $X$ are discrete. Hristache, Juditsky, and Spokoiny (2001) developed an iterated average derivative estimator that performs well when $X$ is high-dimensional. Ichimura and Lee (1991)

and Hristache, Juditsky, Polzehl and Spokoiny (2001) investigate multiple-index generalizations of (2.1).

The usefulness of single-index models can be illustrated with an example that is taken from Horowitz and Härdle (1996). The example consists of estimating a model of product innovation by German manufacturers of investment goods. The data, assembled in 1989 by the IFO Institute of Munich, consist of observations on 1100 manufacturers. The dependent variable is $Y = 1$ if a manufacturer realized an innovation during 1989 in a specific product category and 0 otherwise. The independent variables are the number of employees in the product category ($EMPLP$), the number of employees in the entire firm ($EMPLF$), an indicator of the firm's production capacity utilization ($CAP$), and a variable $DEM$, which is 1 if a firm expected increasing demand in the product category and 0 otherwise. The first three independent variables are standardized so that they have units of standard deviations from their means. Scale normalization was achieved by setting $\beta_{EMPLP} = 1$.

Table 1 shows the parameter estimates obtained using a binary probit model and the semiparametric method of Horowitz and Härdle (1996). Figure 2 shows a kernel estimate of $G'(v)$. There are two important differences between the semiparametric and probit estimates. First, the semiparametric estimate of $\beta_{EMPLF}$ is small and statistically nonsignificant, whereas the probit estimate is significant at the 0.05 level and similar in size to $\beta_{CAP}$. Second, in the binary probit model, $G$ is a cumulative normal distribution function, so $G'$ is a normal density function. Figure 2 reveals, however, that $G'$ is bimodal. This bimodality suggests that the data may be a mixture of two populations. An obvious next step in the analysis of the data would be to search for variables that characterize these populations. Standard diagnostic techniques for binary probit models would provide no indication that $G'$ is bimodal. Thus, the semiparametric estimate has revealed an important feature of the data that could not easily be found using standard parametric methods.

### *2.2 Partially Linear Models*

In a partially linear model, $X$ is partitioned into two non-overlapping subvectors, $X_1$ and $X_2$. The model has the form

$$(2.3) \qquad \boldsymbol{E}(Y \mid x_1, x_2) = \beta' x_1 + G(x_2),$$

where $\beta$ is an unknown constant vector and $G$ is an unknown function. This model is distinct from the class of single index models. A single index model is not partially linear unless $G$ is a linear function. Conversely, a partially linear model is a single index model only in this case.

Stock (1989, 1991) and Engle *et al.* (1986) illustrate the use of (2.3) in applications. Identification of $\beta$ requires the *exclusion restriction* that none of the components of $X_1$ are perfectly predictable by components of $X_2$. When $\beta$ is identified, it can be estimated with an $n^{-1/2}$ rate of convergence regardless of the dimensions of $X_1$ and $X_2$. Thus, the curse of dimensionality is avoided in estimating $\beta$.

An estimator of $\beta$ can be obtained by observing that (2.3) implies

(2.4) $$Y - \boldsymbol{E}(Y \mid x_2) = \beta'[X_1 - \boldsymbol{E}(X_1 \mid x_2)] + U,$$

where $U$ is an unobserved random variable satisfying $\boldsymbol{E}(U \mid x_1, x_2) = 0$. Robinson (1988) shows that under regularity conditions, $\beta$ can be estimated by applying OLS to (3.4) after replacing $\boldsymbol{E}(Y \mid x_2)$ and $\boldsymbol{E}(X_1 \mid x_2)$ with nonparametric estimators. The estimator of $\beta$, $b_n$, converges at rate $n^{-1/2}$ and is asymptotically normally distributed. $G$ can be estimated by carrying out the nonparametric regression of $Y - b_n' X_1$ on $X_2$. Unlike $b_n$, the estimator of $G$ suffers from the curse of dimensionality; its rate of convergence decreases as the dimension of $X_2$ increases.

*2.3 Nonparametric Additive Models*

Let $X$ have $d$ continuously distributed components that are denoted $X_1$, ..., $X_d$. In a nonparametric additive model of the conditional mean function,

(2.5) $$\boldsymbol{E}(Y \mid x) = \mu + f_1(x_1) + \ldots + f_d(x_d),$$

where $\mu$ is a constant and $f_1, \ldots, f_d$ are unknown functions that satisfy a location normalization condition such as

(2.6) $$\int f_k(v) w_k(v) dv = 0, \quad k = 1, .., d,$$

where $w_k$ is a non-negative weight function. An additive model is distinct from a single index model unless $\boldsymbol{E}(Y \mid x)$ is a linear function of $x$. Additive and partially linear models are distinct unless $\boldsymbol{E}(Y \mid x)$ is partially linear and $G$ in (2.3) is additive.

An estimator of $f_k$ $(k = 1, \ldots, d)$ can be obtained by observing that (2.5) and (2.6) imply

(2.7) $$f_k(x_k) = \int \boldsymbol{E}(Y \mid x) w_{-k}(x_{-k}) dx_{-k},$$

where $x_{-k}$ is the vector consisting of all components of $x$ except the $k$'th and $w_{-k}$ is a weight function that satisfies $\int w_{-k}(x_{-k}) dx_{-k} = 1$. The estimator of $f_k$ is obtained by replacing $\boldsymbol{E}(Y \mid x)$ on the right-hand side of (2.7) with nonparametric estimators. Linton and Nielsen

(1995) and Linton (1997) present the details of the procedure and extensions of it. Under suitable conditions, the estimator of $f_k$ converges to the true $f_k$ at rate $n^{-2/5}$ regardless of the dimension of $X$. Thus, the additive model provides dimension reduction. It also permits extrapolation of $\boldsymbol{E}(Y \mid x)$ within the rectangle formed by the supports of the individual components of $X$. Mammen, Linton, and Nielsen (1999) describe a backfitting procedure that is likely to be more precise than the estimator based on (2.7) when $d$ is large. See Hastie and Tibshirani (1990) for an early discussion of backfitting.

Linton and Härdle (1996) describe a generalized additive model whose form is

$$\boldsymbol{E}(Y \mid x) = G[\mu + f_1(x_1) + ... + f_K(x_d)], \tag{2.8}$$

where $f_1, ..., f_d$ are unknown functions and $G$ is a known, strictly increasing (or decreasing) function. Horowitz (2001) describes a version of (2.8) in which $G$ is unknown. Both forms of (2.8) achieve dimension reduction. When $G$ is unknown, (2.8) nests additive and single index models and, under certain conditions, partially linear models.

The use of the nonparametric additive specification (2.5) can be illustrated by estimating the model $\boldsymbol{E}(\log W \mid EXP, EDUC) = \mu + f_{EXP}(EXP) + f_{EDUC}(EDUC)$, where $W$ and $EXP$ are defined as in Section 1, and $EDUC$ denotes years of education. The data are taken from the 1993 CPS and are for white males with 14 or fewer years of education who work full time and live in urban areas of the North Central U.S. The results are shown in Figure 3. The unknown functions $f_{EXP}$ and $f_{EDUC}$ are estimated by the method of Linton and Nielsen (1995) and are normalized so that $f_{EXP}(2) = f_{EDCU}(5) = 0$. The estimates of $f_{EXP}$ (Figure 3a) and $f_{EDUC}$ (Figure 3b) are nonlinear and differently shaped. Functions $f_{EXP}$ and $f_{EDUC}$ with different shapes cannot be produced by a single index model, and a lengthy specification search might be needed to find a parametric model that produces the shapes shown in Figure 3. Some of the fluctuations of the estimates of $f_{EXP}$ and $f_{EDUC}$ may be artifacts of random sampling error rather than features of $\boldsymbol{E}(\log W \mid EXP, EDUC)$. However, a more elaborate analysis that takes account of the effects of random sampling error rejects the hypothesis that either function is linear.

*2.4 Transformation Models*

A transformation model has the form

$$H(Y) = \beta' X + U, \tag{2.9}$$

where $H$ is an unknown increasing function, $\beta$ is an unknown finite dimensional vector of constants, and $U$ is an unobserved random variable. It is assumed here that $U$ is statistically

independent of $X$. The aim is to estimate $H$ and $\beta$. One possibility is to assume that $H$ is known up to a finite-dimensional parameter. For example, $H$ could be the Box-Cox transformation

$$H(y) = \begin{cases} (y^{\tau} - 1)/\tau \text{ if } \tau > 0 \\ \\ \log y \text{ if } \tau = 0 \end{cases}$$

where $\tau$ is an unknown parameter. Methods for estimating transformation models in which $H$ is parametric have been developed by Amemiya and Powell (1981) and Foster, *et al*. (2001) among others.

Another possibility is to assume that $H$ is unknown but that the distribution of $U$ is known. Cheng, Wei, and Ying (1995, 1997) have developed estimators for this version of (2.9). Consider, first, the problem of estimating $\beta$. Let $F$ denote the (known) cumulative distribution function (CDF) of $U$. Let $(Y_i, X_i)$ and $(Y_j, X_j)$ $(i \neq j)$ be two distinct, independent observations of $(Y, X)$. Then it follows from (2.9) that

(2.10) $$\boldsymbol{E}[I(Y_i > Y_j) \mid X_i = x_i, X_j = x_j)] = \boldsymbol{P}[U_i - U_j > -(x_i - x_j)].$$

Let $G(z) = \boldsymbol{P}(U_i - U_j > z)$ for any real $z$. Then

$$G(z) = \int_{-\infty}^{\infty} [1 - F(u + z)] dF(u) .$$

$G$ is a known function because $F$ is assumed known. Substituting $G$ into (2.10) gives

$$\boldsymbol{E}[I(Y_i > Y_j) \mid X_i = x_i, X_j = x_j)] = G[-\beta'(x_i - x_j)].$$

Define $X_{ij} = X_i - X_j$. Then it follows that $\beta$ satisfies the moment condition

(2.11) $$\boldsymbol{E}\{w(\beta' X_{ij}) X_{ij} [[I(Y_i > Y_j) - G(-\beta' X_{ij})]\} = 0$$

where $w$ is a weight function. Cheng, Wei, and Ying (1995) propose estimating $\beta$ by replacing the population moment condition (2.11) with the sample analog

(2.12) $$\sum_{i=1}^{n} \sum_{j=1}^{n} \{w(b' X_{ij}) X_{ij} [I(Y_i > Y_j) - G(-b' X_{ij})]\} = 0 .$$

The estimator of $\beta$, $b_n$, is the solution to (2.12). Equation (2.12) has a unique solution if $w(z) = 1$ for all $z$ and the matrix $\sum_i \sum_j X_{ij}' X_{ij}$ is positive definite. It also has a unique solution asymptotically if $w$ is positive everywhere (Cheng, Wei, and Ying 1995). Moreover, $b_n$

converges almost surely to $\beta$. Cheng, Wei, and Ying (1995) also give conditions under which $n^{1/2}(b_n - \beta)$ is asymptotically normally distributed with a mean of 0.

The problem of estimating the transformation function $H$ is addressed by Cheng, Wei, and Ying (1997). Equation (5.1) implies that for any real $y$ and vector $x$ that is conformable with $X$, $\boldsymbol{E}I[I(Y \le y) \mid X = x] - F[H(y) - \beta' x] = 0$. Cheng, Wei, and Ying (1997) propose estimating $H(y)$ by the solution to the sample analog of this equation. That is, the estimator $H_n(y)$ solves

$$n^{-1}\sum_{i=1}^{n}\{I(Y_i \le y) - F[H_n(y) - b_n' X_i]\} = 0,$$

where $b_n$ is the solution to (2.12). Cheng, Wei, and Ying (1997) show that if $F$ is strictly increasing on its support, then $H_n(y)$ converges to $H(y)$ almost surely uniformly over any interval $[0,t]$ such that $\boldsymbol{P}(Y > t) > 0$. Moreover, $n^{1/2}(H_n - H)$ converges to a mean-zero Gaussian process over this interval.

A third possibility is to assume that $H$ and $F$ are both nonparametric in (2.9). In this case, certain normalizations are needed to make identification of (2.9) possible. First, observe that (2.9) continues to hold if $H$ is replaced by $cH$, $\beta$ is replaced by $c\beta$, and $U$ is replaced by $cU$ for any positive constant $c$. Therefore, a scale normalization is needed to make identification possible. This will be done here by setting $|\beta_1| = 1$, where $\beta_1$ is the first component of $\beta$. Observe, also, that when $H$ and $F$ are nonparametric, (2.9) is a semiparametric single-index model. Therefore, identification of $\beta$ requires $X$ to have at least one component whose distribution conditional on the others is continuous and whose $\beta$ coefficient is non-zero. Assume without loss of generality that the components of $X$ are ordered so that the first satisfies this condition.

It can also be seen that (2.9) is unchanged if $H$ is replaced by $H + d$ and $U$ is replaced by $U + d$ for any positive or negative constant $d$. Therefore, a location normalization is also needed to achieve identification when $H$ and $F$ are nonparametric. Location normalization will be carried out here by assuming that $H(y_0) = 0$ for some finite $y_0$ With this location normalization, there is no centering assumption on $U$ and no intercept term in $X$.

Now consider the problem of estimating $H$, $\beta$, and $F$. Because (2.9) is a single-index model in this case, $\beta$ can be estimated using the methods described in Section 2.1. Let $b_n$

denote the estimator of $\beta$. One approach to estimating $H$ and $F$ is given by Horowitz (1996). To describe this approach, define $Z = \beta' X$. Let $G(\cdot \mid z)$ denote the CDF of $Y$ conditional on $Z = z$. Set $G_y(y \mid z) = \partial G(y \mid z)/\partial z$ and $G_z(y \mid z) = \partial G(y \mid z)/\partial z$. Then it follows from (2.9) that $H'(y) = -G_y(y \mid z)/G_z(y \mid z)$ and that

$$(2.13) \quad H(y) = -\int_{y_0}^{y} [G_y(v \mid z)/G_z(v \mid z)]dv$$

for any $z$ such that the denominator of the integrand is non-zero. Now let $w(\cdot)$ be a scalar-valued, non-negative weight function with compact support $S_w$ such that the denominator of $G_z(v \mid z)$ is bounded away from 0 for all $v \in [y_0, y]$ and $z \in S_w$. Also assume that

$$\int_{S_w} w(z)dz = 1.$$

Then

$$(2.14) \quad H(y) = -\int_{y_0}^{y} \int_{S_w} w(z)[G_y(v \mid z)/G_z(v \mid z)]dzdv$$

Horowitz (1996) obtains an estimator of $H$ from (2.14) by replacing $G_y$ and $G_z$ by kernel estimators. Specifically, $G_y$ is replaced by a kernel estimator of the probability density function of $Y$ conditional on $b_n' X = z$, and $G_z$ is replaced by a kernel estimator of the derivative with respect to $z$ of the CDF of $Y$ conditional on $b_n' X = z$. Denote these estimators by $G_{ny}$ and $G_{nz}$. Then the estimator of $H$ $T$ is

$$(2.15) \quad H_n(y) = -\int_{y_0}^{y} \int_{S_w} w(z)[G_{ny}(v \mid z)/G_{nz}(v \mid z)]dzdv$$

Horowitz (1996) gives conditions under which $H_n$ is uniformly consistent for $H$ and $n^{1/2}(H_n - H)$ converges weakly to a mean-zero Gaussian process. Horowitz (1996) also shows how to estimate $F$, the CDF of $U$, and gives conditions under which $n^{1/2}(F_n - F)$ converges to a mean-zero Gaussian process, where $F_n$ is the estimator. Gørgens and Horowitz (1999) extend these results to a censored version of (2.9). Integration over $z$ in (2.14) and (2.15) accelerates the convergence of $H_n$ to $H$. Kernel estimators converge in probability at rates slower than $n^{-1/2}$. Therefore, $G_{ny}(v \mid z)/G_{nz}(v \mid z))$ is not $n^{-1/2}$-consistent for $G_y(v \mid z)/G_z(v \mid z)$. However, integration over $z$ and $v$ in (2.15) creates an averaging effect that causes the integral

and, therefore, $H_n$ to converge at the rate $n^{-1/2}$. This is the reason for basing the estimator on (2.14) instead of (2.13).

Other estimators of $H$ when $H$ and $F$ are both nonparametric have been proposed by Ye and Duan (1997) and Chen (2002). Chen uses a rank-based approach that is in some ways simpler than that of Horowitz (1996) and may have better finite-sample performance. To describe this approach, define $d_{iy} = I(Y_i > y)$ and $d_{jy_0} = I(Y_j > y_0)$. Let $i \neq j$. Then $\boldsymbol{E}(d_{iy} - d_{jy0} \mid X_i, X_j) \geq 0$ whenever $Z_i - Z_j \geq H(y)$. This suggests that if $\beta$ were known, then $H(y)$ could be estimated by

$$H_n(y) = \arg\max_{\tau} \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} (d_{iy} - d_{iy_0}) I(Z_i - Z_j \geq \tau).$$

Since $\beta$ is unknown, Chen (2002) proposes

$$H_n(y) = \arg\max_{\tau} \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} (d_{iy} - d_{iy_0}) I(b_n' X_i - b_n' X_j \geq \tau).$$

Chen (2002) gives conditions under which $H_n$ is uniformly consistent for $H$ and $n^{1/2}(H_n - H)$ converges to a mean-zero Gaussian process. Chen (2002) also shows how this method can be extended to a censored version of (2.9).

## 3. The Proportional Hazards Model with Unobserved Heterogeneity

Let $T$ denote a duration such as that of a spell of employment or unemployment. Let $f(t \mid x)$ denote the probability density of $T$ at $t$ conditional on $X = x$, where Let $F(t \mid x) = P(T \leq t \mid X = x)$ where $X$ is a vector of covariates. Let $f(t \mid x)$ denote the corresponding conditional probability density function. The conditional hazard function is defined as

$$\lambda(t \mid x) = \frac{f(t \mid x)}{1 - F(t \mid x)}.$$

This section is concerned with an approach to modeling $\lambda(t \mid x)$ that is based on the proportional hazards model of Cox (1972).

The proportional hazards model is widely used for the analysis of duration data. Its form is

(3.1) $\lambda(t \mid x) = \lambda_0(t) e^{-x'\beta}$,

where $\beta$ is a vector of constant parameters that is conformable with $X$ and $\lambda_0$ is a non-negative function that is called the baseline hazard function. The essential characteristic of (3.1) that distinguishes it from other models is that $\lambda(t\,|\,x)$ is the product of a function of $t$ alone and a function of $x$ alone. Cox (1972) developed a partial likelihood estimator of $\beta$ and a nonparametric estimator of $\lambda_0$. Tsiatis (1981) derived the asymptotic properties of these estimators.

In the proportional hazards model with unobserved heterogeneity, the hazard function is conditioned on the covariates $X$ and an unobserved random variable $U$ that is assumed to be independent of $X$. The form of the model is

(3.2) $\lambda(t\,|\,x,u) = \lambda_0(t)e^{-(\beta' x+u)}$,

where $\lambda(\cdot\,|\,x,u)$ is the hazard conditional on $X = x$ and $U = u$. In a model of the duration of employment $U$ might represent unobserved attributes of an individual (possibly ability) that affect employment duration. A variety of estimators of $\lambda_0$ and $\beta$ have been proposed under the assumption that $\lambda_0$ or the distribution of $U$ or both are known up to a finite-dimensional parameter. See, for example, Lancaster (1979), Heckman and Singer (1984a), Meyer (1990), Nielsen, *et al*. (1992), and Murphy (1994, 1995). 'However, $\lambda_0$ and the distribution of $U$ are nonparametrically identified (Elbers and Ridder 1982, Heckman and Singer 1984b), which suggests that they can be estimated nonparametrically.

Horowitz (1999) describes a nonparametric estimator of $\lambda_0$ and the density of $U$ in model (3.2). His estimator is based on expressing (3.2) as a type of transformation model. To do this, define the integrated baseline hazard funtion, $\Lambda_0$ by

$$\Lambda_0(t) = \int_0^t \lambda_0(\tau)d\tau\,.$$

Then it is not difficult to show that (3.2) is equivalent to the transformation model

(3.3) $\log\Lambda_0(T) = X'\beta + U + \varepsilon$,

where $\varepsilon$ is a random variable that is independent of $X$ and $U$ and has the CDF $F_\varepsilon(y) = 1 - \exp(-e^y)$. Now define $\sigma = |\beta_1|$, where $\beta_1$ is the first component of $\beta$ and is assumed to be non-zero. Then $\beta/\sigma$ and $H = \sigma^{-1}\log\Lambda_0$ can be estimated by using the methods of Section 2.4. Denote the resulting estimators of $\beta/\sigma$ and $H$ by $\alpha_n$ and $H_n$. If $\sigma$ were known, then $\beta$ and $\Lambda_0$ could be estimated by $b_n = \sigma\alpha_n$ and $\Lambda_{n0} = \exp(\sigma H_n)$. The baseline

hazard function $\lambda_0$ could be estimated by differentiating $\Lambda_{n0}$. Thus, it is necessary only to find an estimator of the scale parameter $\sigma$.

To do this, define $Z = \beta' X$, and let $G(\cdot \mid z)$ denote the CDF of $T$ conditional on $Z = z$. It can be shown that

$$G(t \mid z) = 1 - \int \exp[-\Lambda_0(t) e^{-(\beta' x + u)}] dF(u),$$

where $F$ is the CDF of $U$. Let $p$ denote the probability density function of $Z$. Define $G_z(t \mid z) = \partial G(t \mid z) / \partial z$ and

$$\sigma(t) = \frac{\int G_z(t \mid z) p(z)^2 dz}{\int G(t \mid z) p(z)^2 dz}.$$

Then it can be shown using l'Hospital's rule that if $\Lambda_0(t) > 0$ for all $t > 0$, then

$$\sigma = \lim_{t \to 0} \sigma(t).$$

To estimate $\sigma$, let $p_n$, $G_{nz}$ and $G_n$ be kernel estimators of $p$, $G_z$ and $G$, respectively, that are based on a simple random sample of $(T, X)$. Define

$$\sigma_n(t) = \frac{\int G_{nz}(t \mid z) p_n(z)^2 dz}{\int G_n(t \mid z) p_n(z)^2 dz}.$$

Let $c$, $d$, and $\delta$ be constants satisfying $0 < c < \infty$, $1/5 < d < 1/4$, and $1/(2d) < \delta < 1$. Let $\{t_{n1}\}$ and $\{t_{n2}\}$ be sequences of positive numbers such that $\Lambda_0(t_{n1}) = cn^{-d}$ and $\Lambda_0(t_{n2}) = cn^{-\delta d}$. Then $\sigma$ is estimated consistently by

$$\sigma_n = \frac{\sigma_n(t_{n1}) - n^{-d(1-\delta)} \sigma_n(t_{n2})}{n^{-d(1-\delta)}}.$$

Horowitz (1999) gives conditions under which $n^{(1-d)/2}(\sigma_n - \sigma)$ is asymptotically normally distributed with a mean of zero. By choosing $d$ to be close to 1/5, the rate of convergence in probability of $\sigma_n$ to $\sigma$ can be made arbitrarily close to $n^{-2/5}$, which is the fastest possible rate (Ishwaran 1996). It follows from an application of the delta method that the estimators of $\beta$, $\Lambda_0$, and $\lambda_0$ that are given by $b_n = \sigma_n \alpha_n$, $\Lambda_{n0} = \exp(\sigma_n H_n)$, and $\lambda_{n0} = d\Lambda_{n0} / dt$ are also asymptotically normally distributed with means of zero and $n^{-(1-d)/2}$ rates of convergence. The probability density function of $U$ can be estimated consistently by solving the deconvolution problem $W_n = U + \varepsilon$, where $W_n = \log \Lambda_{n0}(T) - X' \beta_n$. Because the distribution of $\varepsilon$ is

"supersmooth," the resulting rate of convergence of the estimator of the density of $U$ is $(\log n)^{-m}$, where $m$ is the number of times that the density is differentiable. This is the fastest possible rate. Horowitz (1999) also shows how to obtain data-based values for $t_{n1}$ and $t_{n2}$ and extends the estimation method to models with censoring.

If panel data on $(T, X)$ are available, then $\Lambda_0$ can be estimated with a $n^{-1/2}$ rate of convergence, and the assumption of independence of $U$ from $X$ can be dropped. Suppose that each individual in a random sample of individuals is observed for exactly two spells. Let $(T_j, X_j : j = 1,2)$ denote the values of $(T, X)$ in the two spells. Define $Z_j = \beta' X_j$. Then the joint survivor function of $T_1$ and $T_2$ conditional on $Z_1 = z_1$ and $Z_2 = z_2$ is

$$S(t_1, t_2 \mid Z_1, Z_2) \equiv \boldsymbol{P}(T_1 > t_1, T_2 > t_2 \mid Z_1, Z_2)$$

$$= \int \exp[-\Lambda_0(t_1) e^{z_1 + u} - \Lambda_0(t_2) e^{z_2 + u}] dP(u \mid Z_1 = z_1, Z_2 = z_2).$$

Honoré (1993) showed that

$$R(t_1, t_2 \mid z_1, z_2) \equiv \frac{\partial S(t_1, t_2 \mid z_1, z_2)/\partial t_1}{\partial S(t_1, t_2 \mid z_1, z_2)/\partial t_2} = \frac{\lambda_0(t_1)}{\lambda_0(t_2)} \exp(z_1 - z_2).$$

Adopt the scale normalization

$$\int_{S_T} \frac{w_t(\tau)}{\lambda_0(\tau)} d\tau = 1,$$

where $w_t$ is a non-negative weight function and $S_T$ is its support. Then

$$\lambda_0(t) = \int_{S_T} w_t(\tau) \exp(z_2 - z_1) R(t, \tau \mid z_2, z_1) d\tau .$$

Now for a weight function $\omega_z$ with support $S_Z$, define

$$w(\tau, z_1, z_2) = w_t(\tau) w_z(z_1) w_z(z_2).$$

Then,

$$\lambda_0(t) = \int_{S_T} d\tau \int_{S_Z} dz_1 \int_{S_Z} dz_2 w(\tau, z_1, z_2) \exp(z_2 - z_1) R(t, \tau \mid z_1, z_2). \tag{3.4}$$

The baseline hazard function can now be estimated by replacing $R$ with an estimator, $R_n$, in (3.4). This can be done by replacing $Z$ with $X'b_n$, where $b_n$ is a consistent estimator of $\beta$ such as a marginal likelihood estimator (Chamberlain 1985, Kalbfleisch and Prentice 1980, Lancaster 2000, Ridder and Tunali 1999), and replacing $S$ with a kernel estimator of the joint survivor function conditional $X_1' b_n = z_1$ and $X_2' b_n = z_2$. The resulting estimator of $\lambda_0$ is

$$\lambda_{n0}(t) =$$

$$\int_{S_T} d\tau \int_{S_Z} dz_1 \int_{S_Z} dz_2 w(\tau, z_1, z_2) \exp(z_2 - z_1) R_n(t, \tau \mid z_1, z_2).$$

The integrated baseline hazard function is estimated by

$$\Lambda_{n0}(t) = \int_0^t \lambda_{n0}(\tau) d\tau .$$

Horowitz and Lee (2003) give conditions under which $n^{1/2}(\Lambda_{n0} - \Lambda_0)$ converges weakly to a tight, mean-zero Gaussian process. The estimated baseline hazard function $\lambda_{n0}$ converges at the rate $n^{-q/(2q+1)}$, where $q \geq 2$ is the number of times that $\lambda_0$ is continuously differentiable. Horowitz and Lee (2003) also show how to estimate a censored version of the model.

## 4. **A Binary Response Model**

The general binary response model has the form

(4.1) $$Y = I(\beta' X + U > 0),$$

where $U$ is an unobserved random variable. If the distribution of $U$ is unknown but depends on $X$ only through the index $\beta' X$, then (4.1) is a single-index model, and $\beta$ can be estimated by the methods described in Section 2.1. An alternative model that is non-nested with single-index models can be obtained by assuming that $median(U \mid X = x) = 0$ for all $x$. This assumption places only weak restrictions on the relation between $X$ and the distribution of $U$. Among other things, it accommodates fairly general types of heteroskedasticity of unknown form, including random coefficients. Under median centering, the inferential problem is to estimate $\beta$. The response function, $\boldsymbol{P}(Y = 1 \mid X = x)$ is not identified without making assumptions about the distribution of $U$ that are stronger than those needed to identify and estimate $\beta$. Without such assumptions, the only restriction on $\boldsymbol{P}(Y = 1 \mid X = x)$ under median centering is that

$$\boldsymbol{P}(Y = 1 \mid X = x) \begin{cases} > 0.5 & \text{if } \beta' x > 0 \\ = 0.5 & \text{if } \beta' x = 0 \\ < 0.5 & \text{if } \beta' x < 0 \end{cases}$$

Manski (1975, 1985) proposed the first estimator of $\beta$ under median centering. Let the data be the simple random sample $\{Y_i, X_i : i = 1, ..., n\}$. The estimator is called the *maximum score* estimator and is

(4.2) $$b_n = \arg\max_{\|b\|=1} n^{-1} \sum_{i=1}^{n} (2Y_i - 1) I(b'X_i \geq 0),$$

where $\|b\|$ denotes the Euclidean norm of the vector $b$. The restriction $\|b\| = 1$ is a scale normalization. Scale normalization is needed for identification because (4.1) identifies $\beta$ only up to scale. Manski (1975,1985) gave conditions under which $b_n$ consistently estimates $\beta$. The rate of convergence of $b_n$ and its asymptotic distribution were derived by Cavanagh (1987) and Kim and Pollard (1990). They showed that the rate of convergence in probability of $b_n$ to $\beta$ is $n^{-1/3}$ and that $n^{1/3}(b_n - \beta)$ converges in distribution to the maximum of a complicated multidimensional stochastic process. The complexity of the limiting distribution of the maximum score estimator limits its usefulness for statistical inference. Delgado, Rodríguez-Póo and Wolf (2001) proposed using subsampling methods to form confidence intervals for $\beta$.

The maximum score estimator has a slow rate of convergence and a complicated asymptotic distribution because it is obtained by maximizing a step function. Horowitz (1992) proposed replacing the indicator function in (4.2) by a smooth function. The resulting estimator of $\beta$ is called the *smoothed maximum score* estimator. Specifically, let $K$ be a smooth function, possibly but not necessarily a distribution function, that satisfies $K(-\infty) = 0$ and $K(\infty) = 1$. Let $\{h_n : n = 1, 2, ...\}$ be a sequence of strictly positive constants (bandwidths) that satisfies $h_n \to 0$ as $n \to \infty$. The smoothed maximum score estimator, $b_{ns}$, is

$$b_{ns} = \arg\max_{b \in B} \sum_{i=1}^{n} (2Y_i - 1) K(X_i'b / h_n),$$

where $B$ is a compact parameter set that satisfies the scale normalization $|b_1| = 1$. Horowitz (1992) shows that under assumptions that are stronger than those of Manski (1975, 1985) but still quite weak, $n^r(b_{ns} - \beta)$ is asymptotically normal, where $2/5 \leq r < 1/2$ and the exact value of $r$ depends on the smoothness of the distribution of $X'\beta$ and of $\boldsymbol{P}(Y = 1 | X = x)$. Moreover, the smoothed maximum score estimator has the fastest possible rate of convergence under its assumptions (Horowitz 1993b). Monte Carlo evidence suggests that the asymptotic normal approximation can be inaccurate with samples of practical size. However, Horowitz (2002) shows that the bootstrap, which is implemented by sampling the data randomly with replacement, provides asymptotic refinements for tests of hypotheses about $\beta$ and produces low ERPs for these tests. Thus, the bootstrap provides a practical way to carry out inference with the smoothed maximum score estimator.

Horowitz (1993c) used the smoothed maximum score method to estimate the parameters of a model of the choice between automobile and transit for work trips in the Washington, D.C., area. The explanatory variables are defined in Table 2. Scale normalization is achieved by setting the coefficient of DCOST equal to 1. The data consist of 842 observations sampled randomly from the Washington, D.C., area transportation study. Each record contains information about a single trip to work, including the chosen mode (automobile or transit) and the values of the explanatory variables. Column 2 of Table 2 shows the smoothed maximum score estimates of the model's parameters. Column 3 shows the half-widths of nominal 90% symmetrical confidence intervals based on the asymptotic normal approximation (half width equals 1.67 times the standard error of the estimate). Column 4 shows half-widths obtained from the bootstrap. The bootstrap confidence intervals are 2.5-3 times wider than the intervals based on the asymptotic normal approximation. The bootstrap confidence interval for the coefficient of DOVTT contains 0, but the confidence interval based on the asymptotic normal approximation does not. Therefore, the hypothesis that the coefficient of DOVTT is zero is not rejected at the 0.1 level based on the bootstrap but is rejected based on the asymptotic normal approximation.

## REFERENCES


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

Amemiya, T. and Powell, J.L. (1981). A Comparison of the Box-Cox Maximum Likelihood Estimator and the Non-Linear Two-Stage Least Squares Estimator, *Journal of Econometrics*, 17, 351-381.

Cavanagh, C.L. (1987). Limiting Behavior of Estimators Defined by Optimization, unpublished manuscript.

Chamberlain, G. (1985). Heterogeneity, Omitted Variable Bias, and Duration Dependence. In J.J. Heckman and B. Singer, (eds), *Longitudinal Analysis of Labor Market Data*, Cambridge University Press, Cambridge, pp. 3-38.

Chen, S. (2002). Rank Estimataion of Transformation Models, *Econometrica*, 70, 1683-1697.

Cheng, S.C., Wei, L.J., and Ying, Z. (1995). Analysis of transformation models with censored data, *Biometrika*, 82, 835-845.

Cheng, S.C., Wei, L.J., and Ying, Z. (1997). Predicting survival probabilities with semiparametric transformation models, *Journal of the American Statistical Association*, 92, 227-235.

Cox, D.R. (1972). Regression Models and Life tables, *Journal of the Royal Statistical Society*, Series B, 34, 187-220.

Delgado, M.A., J.M. Rodríguez-Poo, and M. Wolf (2001). Subsampling Inference in Cube Root Asymptotics with an Application to Manski's Maximum Score Estimator, *Economics Letters*, 73, 241-250.

Elbers, C. and G. Ridder (1982). True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model, *Review of Economic Studies*, 49, 403-409.

Engle, R F, Granger C W J, Rice J, and Weiss A (1986). Semiparametric estimates of the relationship beteween weather and electricity sales. *Journal of the American Statistical Association* 81: 310-320.

Fan, J and Gijbels I (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.

Foster, A.M., Tian, L. and Wei, L.J. (2001). Estimation for the Box-Cox Transformation Model without Assuming Prametric Error Distribution, *Journal of the American Statistical Association*, 96, 1097-1101.

Gørgens, T. and Horowitz, J.L. (1999). Semiparametric Estimation of a Censored Regression Model with an Unknown Transformation of the Dependent Variable, *Journal of Econometrics*, 90, 155-191.

Härdle, W. 1990 *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.

Hastie, T J and Tibshirani R J 1990 *Generalized Additive Models*. Chapman and Hall, London.

Heckman, J. and B. Singer (1984a). A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data, *Econometrica*, 52, 271-320.

Heckman, J. and B. Singer (1984a). The Identifiability of the Proportional Hazard Model, *Review of Economics Studies*, 51, 231-243.

Honoré, B.E. (1993). Identification Results for Duration Models with Multiple Spells, *Review of Economic Studies*, 60, 241-246.

Horowitz, J.L. (1992). A Smoothed Maximum Score Estimator for the Binary Response Model, *Econometrica*, 60, 505-531.

Horowitz, J.L (1993a) Semiparametric and Nonparametric Estimation of Quantal Response Models. In Maddala, G.S., Rao, C.R., and Vinod H.D. (eds.), *Handbook of Statistics, Vol. 11,* Elsevier, Amsterdam, pp. 45-72.

Horowitz, J.L. (1993b). Optimal Rates of Convergence of Parameter Estimators in the Binary Response Model with Weak Distributional Assumptions, *Econometric Theory*, 9, 1-18.

Horowitz, J.L. (1993c). Semiparametric Estimation of a Work-Trip Mode Choice Model, *Journal of Econometrics*, 58, 49-70.

Horowitz, J.L. (1996). Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable, *Econometrica*, 64, 103-137.

Horowitz, J.L. (1998). *Semiparametric Methods in Econometrics*. Springer-Verlag, New York.

Horowitz, J.L. (1999). Semiparametric Estimation of a Proportional Hazard Model with Unobserved Heterogeneity, *Econometrica*, 67, 1001-1028.

Horowitz, J.L. (2001). Nonparametric Estimation of a Generalized Additive Model with an Unknown Link Function. *Econometrica*, 69, 499-513.

Horowitz, J.L. (2002). Bootstrap Critical Values for Tests Based on the Smoothed Maximum Score Estimator, *Journal of Econometrics*, 111, 141-167.

Horowitz J.L. and Härdle W. (1996). Direct Semiparametric Estimation of Single-Index Models with Discrete Covariates, *Journal of the American Statistical Association* 91, 1632-1640.

Horowitz, J.L. and Lee, S. (2003). Semiparametric Estimation of a Panel Data Proportional Hazards Model with Fixed Effects, *Journal of Econometrics*, forthcoming.

Hristache, M., Juditsky, A., Polzehl, J., and Spokoiny, V. (2001). Structure Adaptive Approach for Dimension Reduction, *Annals of Statistics*, 29, 1537-1566.

Hristache, M., Juditsky, A., and Spokoiny, V. (2001). Structure Adaptive Approach for Dimension Reduction, *Annals of Statistics*, 29, 1-32.

Ichimura, H. (1993). Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models, *Journal of Econometrics* 58, 71-120.

Ichimura, H. and Lee L.-F. (1991). Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation. In Barnett W A, Powell J, and Tauchen G (eds), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge University Press, Cambridge, pp. 3-49.

Ishwaran, H. (1996). Identifiability and Rates of Estimation for Scale Parameters in Location Mixture Models, *Annals of Statistics*, 24, 1560-1571.

Kalbfleisch and Prentice (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

Kim, J. and D. Pollard (1990). Cube Root Asymptotics, *Annals of Statistics*, 15, 541-551.

Klein R W and Spady R H 1993 An efficient semiparametric estimator for binary response models. *Econometrica* 61: 387-421.

Lancaster, T. (1979). Econometric Methods for the Duration of Unemployment, *Econometrica*, 47, 939-956.

Lancaster, T. (2000). The Incidental Parameter Problem Since 1948, *Journal of Econometrics*, 95, 391-413.

Linton, O.B. (1997). Efficient Estimation of Additive Nonparametric Regression Models, *Biometrika* 84, 469-473.

Linton, O.B. and Härdle, W. (1996). Estimating Additive Regression Models with Known Links, *Biometrika*, 83, 529-540.

Linton, O.B. and Nielsen J.P. (1995). A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration, *Biometrika*, 82, 93-100.

Mammen, E., Linton, O.B., and Nielsen, J.P. (1999). The Existence and Asymptotic Properties of Backfitting Projection Algorithm under Weak Conditions, *Annals of Statistics*, 27, 1443-1490.

Manski, C.F. (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. (1988). Identification of Binary Response Models, *Journal of the American Statistical Association,* 83, 729-738.

Matzkin, R.L. (1994). Restrictions of Economic Theory in Nonparametric Methods. In Engle, R.F. and McFadden, D.L. (eds) *Handbook of Econometrics*, Vol. 4. North-Holland, Amsterdam, pp. 2523-2558.

Meyer, B. D. (1990). Unemployment Insurance and Unemployment Spells, *Econometrica*, 58, 757-782.

Murphy, S. A. (1994). Consistency in a Proportional Hazards Model Incorporating a Random Effect, *Annals of Statistics*, 22, 712-731.

Murphy, S. A. (1995). Asymptotic Theory for the Frailty Model, *Annals of Statistics*, 23, 182-198.

Nielsen, G. G., Gill, R.D., Andersen, P.K., and Sørensen, T.I.A. (1992). A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models, *Scandinavian Journal of Statistics*, 19, 25-43, 1992.

Powell, J.L. (1994). Estimation of Semiparametric Models. In Engle, R.F. and McFadden, D.L. (eds) *Handbook of Econometrics*, Vol. 4. North-Holland, Amsterdam, pp. 2444-2521.

Powell, J.L., Stock, J.H., and Stoker, T.M. (1989). Semiparametric Estimation of Index Coefficients, *Econometrica*, 51, 1403-1430.

Ridder, G. and Tunali, I. (1999). Stratified Partial Likelihood Estimation, *Journal of Econometrics*, 92, 193-232.

Robinson, P.M. (1988). Root-*N*-Consistent Semiparametric Regression. *Econometrica*, 56, 931-954.

Stock, J.H. (1989). Nonparametric Policy Analysis, *Journal of the American Statistical Association*, 84, 567-575.

Stock, J.H. (1991). Nonparametric Policy Analysis: An Application to Estimating Hazardous Waste Cleanup Benefits. In Barnett, W.A., Powell, J., and Tauchen, G. (eds) *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge University Press, Cambridge, pp. 77-98.

Tsiatis, A. A. (1981). A Large Sample Study of Cox's Regression Model, *Annals of Statistics*, 9, 93-108.

Ye, J. and Duan, N. (1997). Nonparametric $n^{-1/2}$-Consistent Estimation for the General Transformation Model, *Annals of Statistics*, 25, 2682-2717.

**Table 1: Estimated Coefficients (Standard Errors) for Model of Product Innovation**

| EMPLP | EMPLF | CAP | DEM |
|---|---|---|---|
| | Semiparametric Model | | |
| 1 | 0.032 | 0.346 | 1.732 |
| | (0.023) | (0.078) | (0.509) |
| | Probit Model | | |
| 1 | 0.516 | 0.520 | 1.895 |
| | (0.024) | (0.163) | (0.387) |

**Table 2: Smoothed Maximum Score Estimates of a Work-Trip Mode-Choice Model**

| Variable | Estimated Coefficient | Half-Width of Nominal 90% Conf. Interval Based on Asymp. Normal Approximation | Half-Width of Nominal 90% Conf. Interval Based on Bootstrap |
|---|---|---|---|
| INTRCPT | -1.5761 | 0.2812 | 0.7664 |
| AUTOS | 2.2418 | 0.2989 | 0.7488 |
| DOVTT | 0.0269 | 0.0124 | 0.0310 |
| DIVTT | 0.0143 | 0.0033 | 0.0087 |
| DCOST | 1.0[b] | | |

[a] Definitions of variables: INTRCPT: Intercept term equal to 1; AUTOS: Number of cars owned by traveler's household; DOVTT: Transit out-of-vehicle travel time minus automobile out-of-vehicle travel time (minutes); DIVTT: Transit in-vehicle travel time minus automobile in-vehicle travel time; DCOST: Transit fare minus automobile travel cost ($).

[b] Coefficient equal to 1 by scale normalization

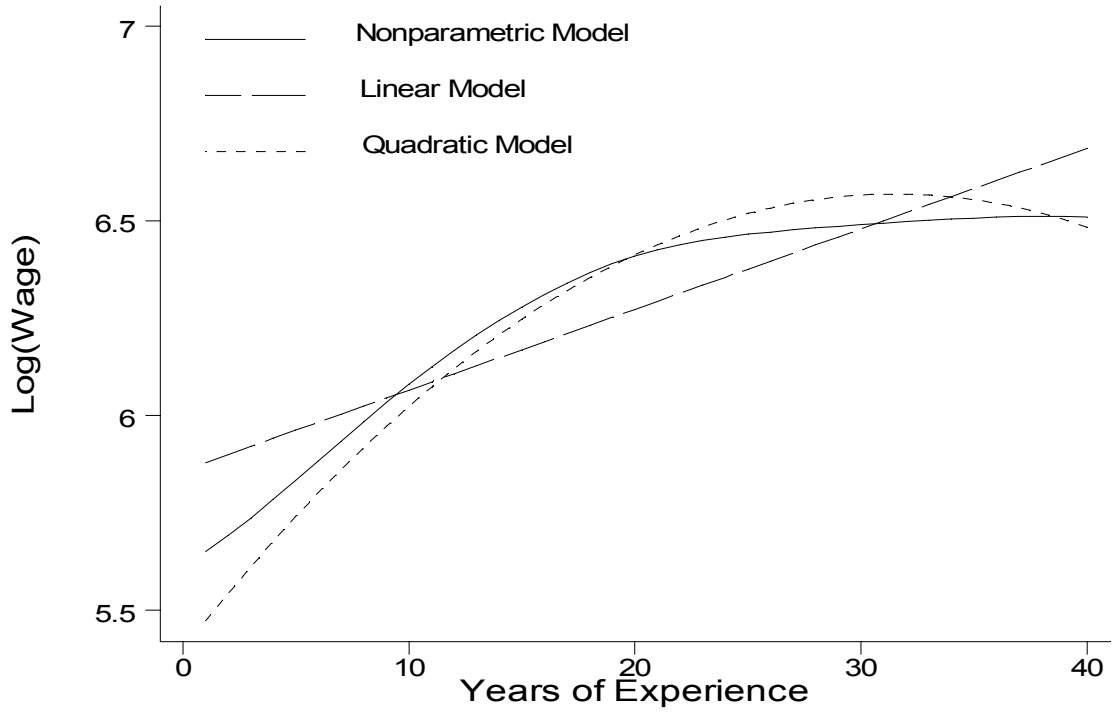


Figure 1: Nonparametric and Parametric Estimates of Mean Log Wages

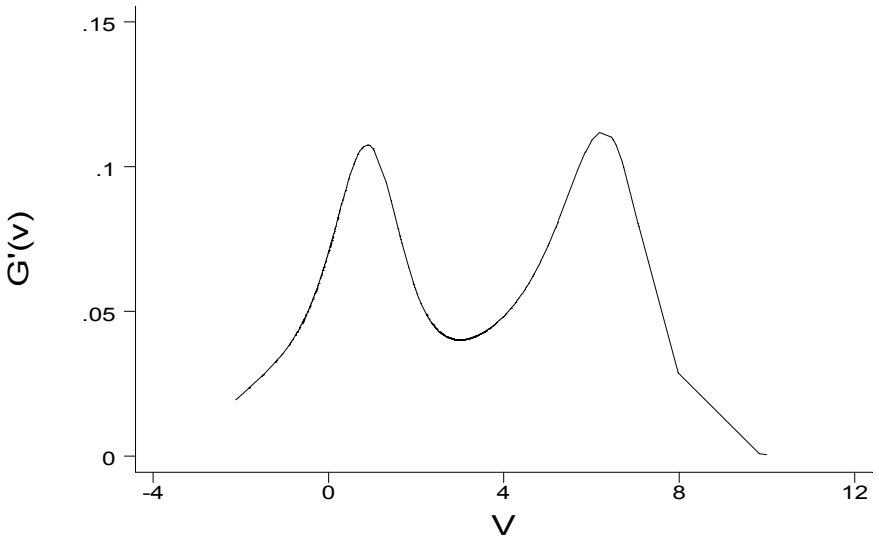


Figure 2: Estimate of $G'(v)$ for model of product innovation.

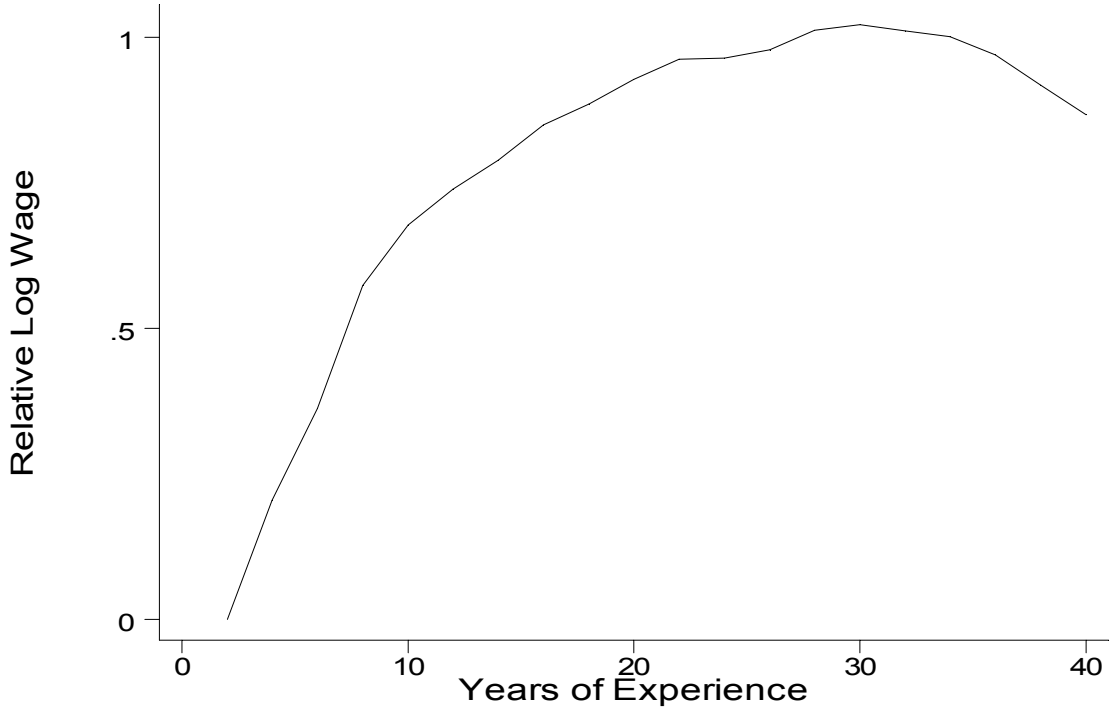


Figure 3a

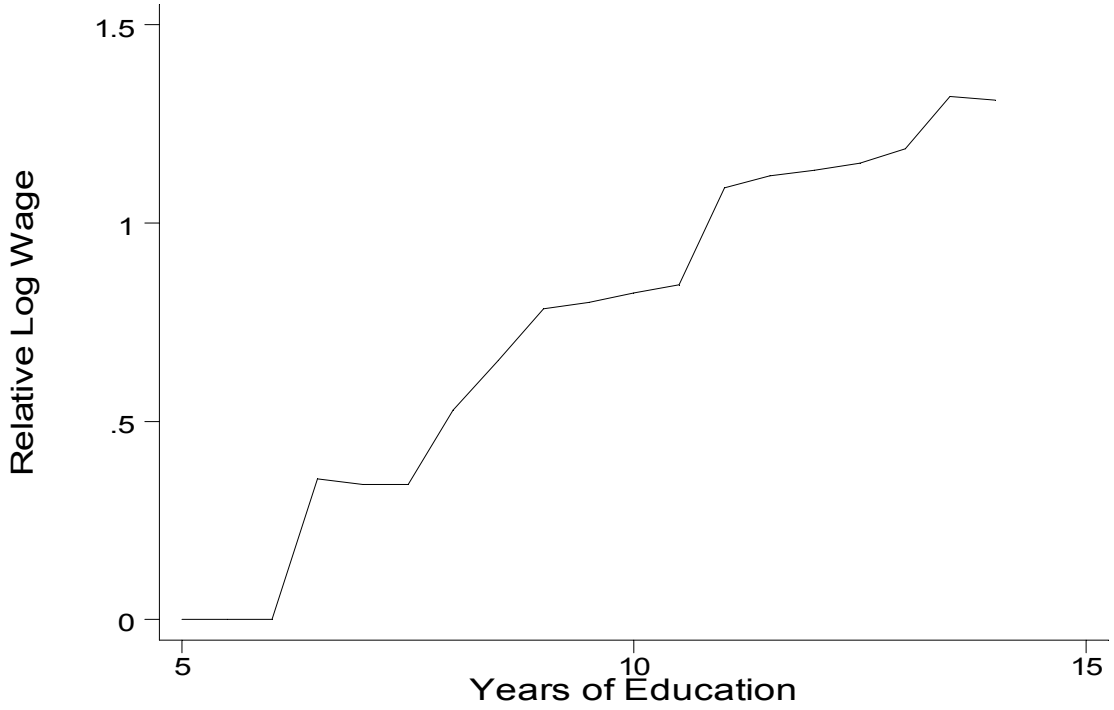


Figure 3b