

Chapter 2

Single-Index Models

This chapter describes *single-index* models for conditional mean and quantile functions. Single-index models relax some of the restrictive assumptions of familiar parametric models, such as linear models and binary probit or logit models. In addition, single-index models achieve dimension reduction and, thereby, greater estimation precision than is possible with fully nonparametric estimation of $E(Y|X = x)$ when X is multidimensional. Finally, single-index models are often easy to compute, and their results are easy to interpret. Sections 2.1–2.9 present a detailed discussion of single-index models for conditional mean functions. Conditional quantile functions are discussed in Section 2.9.

2.1 Definition of a Single-Index Model of a Conditional Mean Function

Let Y be a scalar random variable and X be a $d \times 1$ random vector. In a single-index model, the conditional mean function $E(Y|X = x)$ has the form

$$E(Y|X = x) = G(x'\beta), \quad (2.1)$$

where β is an unknown $d \times 1$ constant vector and G is an unknown function. The quantity $x'\beta$ is called an *index*. The inferential problem in (2.1) is to estimate β and G from observations of (Y, X) .

Model (2.1) contains many widely used parametric models as special cases. 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. A tobit model is obtained if one assumes that $G(x'\beta) = E(Y|X = x)$ in the model

$$Y = \max(0, X'\beta + U),$$

where U is an unobserved, normally distributed random variable that is independent of X and has a mean of zero. 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.

A single-index model achieves dimension reduction and avoids the curse of dimensionality because, as will be seen later in this chapter, the index $x'\beta$ aggregates the dimension of x . Consequently, G in a single-index model can be estimated with the same rate of convergence in probability that it would have if the one-dimensional quantity $X'\beta$ were observable. Moreover, β can be estimated with the same rate of convergence, $n^{-1/2}$, that is achieved in a parametric model. Thus, in terms of rate of convergence in probability, the single-index model is as accurate as a parametric model for estimating β and as accurate as a one-dimensional nonparametric mean regression 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.

The assumptions of a single-index model are weaker than those of a parametric model and stronger than those of a fully nonparametric model. Thus, a single-index model reduces the risk of misspecification relative to a parametric model while avoiding some drawbacks of fully nonparametric methods such as the curse of dimensionality, difficulty of interpretation, and lack of extrapolation capability.

There is an important exception to the characterization of a single-index model as intermediate or as making weaker assumptions than a nonparametric model. This exception occurs in the estimation of structural economic models. A structural model is one whose components have a clearly defined relation to economic theory. It turns out that the restrictions needed to make possible a structural interpretation of a nonparametric model can cause the nonparametric model to be no more general than a single-index model. To see why, consider a simple structural model of whether an individual is employed or unemployed.

Example 2.1: A Binary-Response Model of Employment Status An important model in economic theory states that an individual is employed if his market wage exceeds his reservation wage, which is the value of his time if unemployed. Let Y^* denote the difference between an individual's market and reservation wages. Consider the problem of inferring the probability distribution of Y^* conditional on a vector of covariates, X , that characterizes the individual and, possibly, the state of the economy. Let H denote the conditional mean function. That is, $E(Y^*|X = x) = H(x)$. Then

$$Y^* = H(X) - U, \quad (2.2)$$

where U is an unobserved random variable that captures the effects of variables other than X that influence employment status (unobserved covariates). Suppose that U is independent of X , and let F be the cumulative distribution function (CDF) of U . The estimation problem is to infer H and F . It turns out, however, that this problem has

no solution unless suitable a priori restrictions are placed on H and F . The remainder of this example explains why this is so and compares alternative sets of restrictions.

To begin, suppose that Y^* were observable. Then H could be estimated nonparametrically as the nonparametric mean regression of Y^* on X . More importantly, the population distribution of the random vector (Y^*, X) would identify (that is, uniquely determine) H if H is a continuous function of the continuous components of X . F would also be identified if Y^* were observable, because F would be the CDF of the identified random variable $U = H(X) - Y^*$. F could be estimated as the empirical distribution function of the quantity that is obtained from U by replacing H with its estimator. However, Y^* is not observable because the market wage is observable only for employed individuals, and the reservation wage is never observable. An individual's employment status is observable, though. Moreover, according to the economic theory model, $Y^* \geq 0$ for employed individuals, whereas $Y^* < 0$ for individuals who are not employed. Thus, employment status provides an observation of the sign of Y^* . Let Y be the indicator of employment status: $Y = 1$ if an individual is employed and $Y = 0$ otherwise. We now investigate whether H and F can be inferred from observations of (Y, X) .

To solve this problem, let $G(x) = P(Y = 1|x)$ be the probability that $Y = 1$ conditional on $X = x$. Because Y is binary, $G(x) = E(Y|X = x)$ and G can be estimated as the nonparametric mean regression of Y on X . More importantly, the population distribution of the observable random vector (Y, X) identifies G if G is a continuous function of the continuous components of X . It follows from (2.2) that $P(Y^* \geq 0|X = x) = F[H(x)]$. Therefore, since $Y^* \geq 0$ if and only if $Y = 1$, $P(Y^* \geq 0|X = x) = P(Y = 1|x)$ and

$$F[H(x)] = G(x). \quad (2.3)$$

The problem of inferring H and F can now be seen. The population distribution of (Y, X) identifies G . H and F are related to G by (2.3). Therefore, H and F are identified and nonparametrically estimable only if (2.3) has a unique solution for H and F in terms of G .

One way to achieve identification is by assuming that H has the single-index structure

$$H(x) = x' \beta. \quad (2.4)$$

If (2.4) holds, then identification of H is equivalent to identification of β . As will be discussed in Section 2.3, β is identified if X has at least one continuously distributed component whose β coefficient is nonzero, F is differentiable and nonconstant, and certain other conditions are satisfied. F is also identified and can be estimated as the nonparametric mean regression of Y on the estimate of $X' \beta$.

The single-index model (2.4) is more restrictive than a fully nonparametric model, so it is important to ask whether H and F are identified and estimable nonparametrically. This question has been investigated by Matzkin (1992, 1994). The

answer turns out to be *no* unless H is restricted to a suitably small class of functions. To see why, suppose that X is a scalar and

$$G(x) = \frac{1}{1 + e^{-x}}.$$

Then one solution to (2.3) is

$$H(x) = x$$

and

$$F(u) = \frac{1}{1 + e^{-u}}; \quad -\infty < u < \infty.$$

Another solution is

$$H(x) = \frac{1}{1 + e^{-x}}$$

and

$$F(u) = u; \quad 0 \leq u \leq 1.$$

Therefore, (2.3) does not have a unique solution, and F and H are not identified unless they are restricted to classes that are smaller than the class of all distribution functions (for F) and the class of all functions (for H).

Matzkin (1992, 1994) gives examples of suitable classes. Each contains some single-index models but none contains all. Thus, the single-index specification consisting of (2.3) and (2.4) contains models that are not within Matzkin's classes of identifiable, nonparametric, structural models. Similarly, there are identifiable, nonparametric, structural models that are not single-index models. Therefore, Matzkin's classes of identifiable, nonparametric, structural models are neither more nor less general than the class of single-index models. It is an open question whether there are interesting and useful classes of identifiable, nonparametric, structural models of the form (2.3) that contain all identifiable single-index submodels of (2.3).

2.2 Multiple-Index Models

A multiple-index model is a generalization of a single-index model. Its form is

$$E(Y|X = x) = x'_0\beta_0 + G(x'_1\beta_1, \dots, x'_M\beta_M), \quad (2.5)$$

where $M \geq 1$ is a known integer, x_m ($m = 0, \dots, M$) is a subvector of x , β_m ($m = 0, \dots, M$) is a vector of unknown parameters, and G is an unknown function. This model has been investigated in detail by Ichimura and Lee (1991) and

Hristache et al. (2001). A different form of the model called *sliced inverse regression* has been proposed by Li (1991). If the β parameters in (2.5) are identified and certain other conditions are satisfied, then the β s can be estimated with a $n^{-1/2}$ rate of convergence in probability, the same as the rate with a parametric model. The estimator of $E(Y|X = x)$ converges at the rate of a nonparametric estimate of a conditional mean function with an M -dimensional argument. Thus, in a multiple-index model, estimation of $E(Y|X = x)$ but not of β suffers from the curse of dimensionality as M increases.

The applications in which a multiple-index model is likely to be useful are different from those in which a single-index model is likely to be useful. The curse of dimensionality associated with increasing M and the need to specify identifiable indices a priori limit the usefulness of multiple-index models for estimating $E(Y|X = x)$. There are, however, applications in which the object of interest is β , not $E(Y|X = x)$, and the specification of indices arises naturally. The following example provides an illustration.

Example 2.2: A Wage Equation with Selectivity Bias Let W denote the logarithm of an individual's market wage. Suppose we want to estimate $E(W|Z = z) \equiv E(W|z)$, where Z is a vector of covariates such as experience and level of education. Suppose, also, that the conditional mean function is assumed to be linear. Then $E(W|z) = z'\alpha$, where α is a vector of coefficients. Moreover,

$$W = z'\alpha + V, \quad (2.6)$$

where V is an unobserved random variable that represents the effects on wages of variables not included in Z (e.g., unobserved ability). If (W, Z) were observable for a random sample of individuals, then α could be estimated, among other ways, by applying ordinary least squares to (2.6). However, W is observable only for employed individuals, and a random sample of individuals is likely to include some who are unemployed. Therefore, unless attention is restricted to groups in which nearly everyone is employed, one cannot expect to observe (W, Z) for a random sample of individuals.

To see how this problem affects estimation of α and how it can lead to a multiple-index model, suppose that employment status is given by the single-index model consisting of (2.2) and (2.4). Then the mean of W conditional on $X = x$, $Z = z$, and $Y = 1$ is

$$E(W|z, x, Y = 1) = z'\alpha + E(V|z, x, U \leq x'\beta). \quad (2.7)$$

If V is independent of Z and X conditional on U , then (2.7) becomes

$$E(W|z, x, Y = 1) = z'\alpha + G(x'\beta), \quad (2.8)$$

where $G(x'\beta) = E(V|z, x, U \leq x'\beta)$. Equation (2.8) is a multiple-index model that gives the mean of log wages of employed individuals conditional on covariates

Z and X . Observe that (2.8) is not equivalent to the linear model (2.6) unless $E(V|z, x, U \leq x'\beta) = 0$. If $E(V|z, x, U \leq x'\beta) \neq 0$, estimation of (2.6) by ordinary least squares will give rise to a *selectivity bias* arising from the fact that one does not observe W for a random sample of individuals. This is also called a *sample selection* problem because the observed values of W are selected nonrandomly from the population. Gronau (1974) and Heckman (1974) used models like (2.7) under the additional assumption that V and U are bivariate normally distributed. In this case G is known up to a scalar parameter, and the model is no longer semiparametric.

In (2.8), α is identified only if X has at least one continuously distributed component that is not a component of Z and whose β coefficient is nonzero. The credibility of such an *exclusion restriction* in an application can be highly problematic. Manski (1994, 1995) provides a detailed discussion of the problems of identification in the presence of sample selection. ■

2.3 Identification of Single-Index Models

The remainder of this chapter is concerned with the semiparametric single-index model (2.1).

2.3.1 Conditions for Identification of β and G

Before estimation of β and G can be considered, restrictions must be imposed that ensure their identification. That is, β 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, by Manski (1988). Some of the restrictions required for identification are easy to see. It is clear that β is not identified if G is a constant function. It is also clear that as in a linear model, β is not identified if there is an exact linear relation among the components of X (perfect multicollinearity). In other words, β is not identified if there are a constant vector α and a constant scalar c such that $X'\alpha = c$ with probability one.

To obtain additional conditions for identification, let γ be any constant and δ be any nonzero constant. Define the function G^* by the relation $G^*(\gamma + \delta v) = G(v)$ for all v in the support of $X'\beta$. Then

$$E(Y|X = x) = G(x'\beta) \quad (2.9)$$

and

$$E(Y|X = x) = G^*(\gamma + x'\beta\delta). \quad (2.10)$$

Models (2.9) and (2.10) are observationally equivalent. They could not be distinguished empirically even if the population distribution of (Y, X) were known. Therefore, β and G are not identified unless restrictions are imposed that uniquely specify

γ and δ . The restriction on γ is called a *location normalization*, and the restriction on δ is called a *scale normalization*. Location normalization can be achieved by requiring X to contain no constant (intercept) component. Scale normalization can be achieved by setting the β coefficient of one component of X equal to one. In this chapter it will be assumed that the components of X have been arranged so that scale normalization is carried out on the coefficient of the first component. Moreover, for reasons that will now be explained, it will also be assumed that this component of X is a continuously distributed random variable.

To see why there must be at least one continuously distributed component of X , consider the following example.

Example 2.3: A Single-Index Model with Only Discrete Covariates Suppose that $X = (X_1, X_2)$ is two-dimensional and discrete with support consisting of the corners of the unit square: (0,0), (1,0), (0,1), and (1,1). Set the coefficient X_1 equal to one to achieve scale normalization. Then (2.1) becomes

$$E(Y|X = x) = G(x_1 + \beta_2 x_2).$$

Suppose that the values of $E(Y|X = x)$ at the points of support of X are as shown in Table 2.1. Then all choices of β_2 and G that equate the entry in the second column to the corresponding entry in the third column are correct models of $E(Y|X = x)$. These models are observationally equivalent and would be indistinguishable from one another even if the population distribution of (Y, X) were known. There are infinitely many such models, so β_2 and G are not identified. Bierens and Hartog (1988) provide a detailed discussion of alternative, observationally equivalent forms of β and G when all components of X are discrete. ■

Another requirement for identification is that G must be differentiable. To understand why, observe that the distinguishing characteristic of a single-index model that makes identification possible is that $E(Y|X = x)$ is constant if x changes in such a way that $x'\beta$ stays constant. However, if $X'\beta$ is a continuously distributed random variable, as it is if X has at least one continuous component with a nonzero coefficient, the set of X values on which $X'\beta = c$ has probability zero for any c . Events of probability zero happen too infrequently to permit identification. If G is differentiable, then $G(X'\beta)$ is close to $G(c)$ whenever $X'\beta$ is close to c . The set of

Table 2.1 An unidentified single-index model

(x_1, x_2) (x_1, x_2)	$E(Y X = x)$	$G(x_1 + \beta_2 x_2)$
(0, 0)	0	$G(0)$
(1, 0)	0.1	$G(1)$
(0, 1)	0.3	$G(\beta_2)$
(1, 1)	0.4	$G(1 + \beta_2)$

X values on which $X'\beta$ is within any specified nonzero distance of c has nonzero probability for any c in the interior of the support of $X'\beta$. This permits identification of β through “approximate” constancy of $X'\beta$.

It is now possible to state a complete set of conditions for identification of β in a single-index model. This theorem assumes that the components of X are all continuous random variables. Identification when some components of X are discrete is more complicated. This case is discussed after the statement of the theorem.

Theorem 2.1 (Identification in Single-Index Models): *Suppose that $E(Y|X = x)$ satisfies model (2.1) and X is a d -dimensional random variable. Then β and G are identified if the following conditions hold:*

- (a) G is differentiable and not constant on the support of $X'\beta$.
- (b) The components of X are continuously distributed random variables that have a joint probability density function.
- (c) The support of X is not contained in any proper linear subspace of \mathbb{R}^d .
- (d) $\beta_1 = 1$. ■

Ichimura (1993) and Manski (1988) provide proofs of several versions of this theorem. It is also possible to prove a version that permits some components of X to be discrete. Two additional conditions are needed. These are as follows: (1) varying the values of the discrete components must not divide the support of $X'\beta$ into disjoint subsets and (2) G must satisfy a nonperiodicity condition.

The following example illustrates the need for condition (1).

Example 2.4: Identification of a Single-Index Model with Continuous and Discrete Covariates Suppose that X has one continuous component, X_1 , whose support is $[0,1]$, and one discrete component, X_2 , whose support is the two-point set $\{0,1\}$. Assume that X_1 and X_2 are independent and that G is strictly increasing on $[0,1]$. Set $\beta_1 = 1$ to achieve scale normalization. Then $X'\beta = X_1 + \beta_2 X_2$. Observe that $E[Y|X = (x_1,0)] = G(x_1)$ and $E[Y|X = (x_1,1)] = G(x_1 + \beta_2)$. Observations of X for which $X_2 = 0$ identify G on $[0,1]$. However, if $\beta_2 > 1$, the support of $X_1 + \beta_2$ is disjoint from $[0,1]$, and β_2 is, in effect, an intercept term in the model for $E[Y|X = (x_1,1)]$. As was explained in the discussion of location and scale normalization, an intercept term is not identified, so β_2 is not identified in this model.

The situation is different if $\beta_2 < 1$, because the supports of X_1 and $X_1 + \beta_2$ then overlap. The interval of overlap is $[\beta_2,1]$. Because of this overlap, there is a subset of the support of X on which $X_2 = 1$ and $G(X_1 + \beta_2) = G(v)$ for some $v \in [0,1]$. The subset is $\{X: X_1 \in [\beta_2,1], X_2 = 1\}$. Since $G(v)$ is identified for $v \in [\beta_2,1]$ by observations of X_1 for which $X_2 = 0$, β_2 can be identified by solving

$$E[Y|X = (x_1,1)] = G(x_1 + \beta_2) \quad (2.11)$$

on the set of x_1 values where the ranges of $E(Y|X = (x_1,1))$ and $G(x_1 + \beta_2)$ overlap. ■

To see why G must satisfy a nonperiodicity condition, suppose that in Example 2.3 G were periodic on $[\beta_2, 1]$ instead of strictly increasing. Then (2.11) would have at least two solutions, so β_2 would not be identified. The assumption that G is strictly increasing on $[0, 1]$ prevents this kind of periodicity, but many other shapes of G also satisfy the nonperiodicity requirement. See Ichimura (1993) for details.

2.3.2 Identification Analysis When X Is Discrete

One of the conclusions reached in Section 2.3.1 is that β and G are not identified in a semiparametric single-index model if all components of X are discrete. It does not necessarily follow, however, that data are completely uninformative about β . In this section it is shown that if G is assumed to be an increasing function, then one can obtain identified *bounds* on the components of β .

To begin, it can be seen from Table 2.1 that there is a G that solves (2.11) for every possible value of β_2 in Example 2.3. Therefore, nothing can be learned about β_2 if nothing is known about G . This is not surprising. Even when the components of X are all continuous, some information about G is necessary to identify β (e.g., differentiability in the case of Theorem 2.1). Continuity and differentiability of G are not useful for identification when all components of X are discrete. A property that is useful, however, is monotonicity. The usefulness of this property is illustrated by the following example, which is a continuation of Example 2.3.

Example 2.5: Identification When X Is Discrete and G Is Monotonic Consider the model of Example 2.3 and Table 2.1 but with the additional assumption that G is a strictly increasing function. That is,

$$G(v_1) < G(v_2) \Leftrightarrow v_1 < v_2. \quad (2.12)$$

Inequality (2.12) together with the information in columns 2 and 3 of Table 2.1 implies that $\beta_2 > 1$. This result is informative, even though it does not point-identify β_2 , because any value of β_2 in $(-\infty, \infty)$ is possible in principle. Knowledge of the population distribution of (Y, X) combined with monotonicity of G excludes all values in $(-\infty, 1]$.

If the support of X is large enough, then it is possible to identify an upper bound on β_2 as well as a lower bound. For example, suppose that the point $(X_1, X_2) = (0.6, 0.5)$ is in the support of X along with the four points in Example 2.3 and that $E(Y|X_1 = 0.6, X_2 = 0.5) = G(0.6 + 0.5\beta_2) = 0.35$. This information combined with (2.12) and row 3 of Table 2.1 implies that $\beta_2 < 0.6 + 0.5\beta_2$, so $\beta_2 < 1.2$. Therefore, the available information gives the identified bounds $1 < \beta_2 < 1.2$. Any value of β_2 in the interval $(1, 1.2)$ is logically possible given the available information, so the bounds $1 < \beta_2 < 1.2$ are the tightest possible. ■

Now consider the general case in which X is d -dimensional for any finite $d \geq 2$ and has M points of support for any finite $M \geq 2$. Let x_m denote the m th point of support ($m = 1, \dots, M$). The population distribution of (Y, X) identifies $G(x'_m\beta)$ for each m . Assume without loss of generality that the support points x_m are sorted so that

$$G(x'_1\beta) \leq G(x'_2\beta) \leq \dots \leq G(x'_M\beta).$$

Achieve location and scale normalization by assuming that X has no constant component and that $\beta_1 = 1$. Also, assume that G is strictly increasing. Then tight, identified bounds on β_m ($2 \leq m \leq M$) can be obtained by solving the linear programming problems

$$\begin{aligned} &\text{maximize (minimize): } b_m \\ &\text{subject to: } x'_jb \leq x'_{j+1}b; \quad j = 1, \dots, M-1 \end{aligned} \tag{2.13}$$

with strict equality holding in the constraint if $G(x'_jb) = G(x'_{j+1}b)$. The solutions to these problems are informative whenever they are not infinite.

Bounds on other functionals of β can be obtained by suitably modifying the objective function of (2.13). For example, suppose that z is a point that is not in the support of X and that we are interested in learning whether $E(Y|X = z) = G(z'\beta)$ is larger or smaller than $E(Y|X = x_m) = G(x'_m\beta)$ for some x_m in the support of X . $G(z'\beta) - G(x'_m\beta)$ is not identified if X is discrete, but $(z - x_m)'\beta$ can be bounded by replacing b_m with $(z - x_m)'\beta$ in the objective function of (2.13). If the resulting lower bound exceeds zero, then we know that $G(z'\beta) > G(x'_m\beta)$, even though $G(z'\beta)$ is unknown. Similarly, $G(z'\beta) < G(x'_m\beta)$ if the upper bound obtained from the modified version of (2.13) is negative.

Now consider solving (2.13) with the objective function $(x_m - z)'\beta$ for each $m = 1, \dots, M$. Suppose this procedure yields the result $(x_m - z)'\beta < 0$ if $m \leq j$ for some j ($1 \leq j \leq M$). Then it follows from monotonicity of G that $G(z'\beta) > G(x'_j\beta)$. Similarly, if the solutions to the modified version of (2.13) yield the result $(x_m - z)'\beta > 0$ if $m \geq k$ for some k ($1 \leq k \leq M$), then $G(z'\beta) < G(x'_k\beta)$. Since $G(x'_j\beta)$ and $G(x'_k\beta)$ are identified, this procedure yields identified bounds on the unidentified quantity $G(z'\beta)$, thereby providing a form of extrapolation in a single-index model with a discrete X . The following example illustrates this form of extrapolation.

Example 2.6: Extrapolation When X Is Discrete and G Is Monotonic Let G , $E(Y|X = x)$, and the points of support of X be as in Example 2.5. Order the points of support as in Table 2.2. As in Example 2.5, the available information implies that

$$1 < \beta_2 < 1.2 \tag{2.14}$$

but does not further identify β_2 . Suppose that $z = (0.3, 0.25)'$. What can be said about the value of $E(Y|X = z) = G(z'\beta) = G(0.3 + 0.25\beta_2)$? This quantity is not

Table 2.2 A second unidentified single-index model

m	x_m	$E(Y X = x_m)$	$G(x_m)$
1	(0, 0)	0	$G(0)$
2	(1, 0)	0.1	$G(1)$
3	(0, 1)	0.3	$G(\beta_2)$
4	(0.6, 0.5)	0.35	$G(0.6 + 0.5\beta_2)$
5	(1, 1)	0.4	$G(1 + \beta_2)$

identified, but the following bounds may be obtained by combining the information in Table 2.2 with inequality (2.14):

$$-0.6 < (x_1 - z)' \beta < -0.55,$$

$$0.4 < (x_2 - z)' \beta < 0.45,$$

$$0.45 < (x_3 - z)' \beta < 0.60,$$

$$0.55 < (x_4 - z)' \beta < 0.60,$$

and

$$1.45 < (x_5 - z)' \beta < 1.60.$$

Therefore, monotonicity of G implies that $G(x_1' \beta) < G(z' \beta) < G(x_2' \beta)$, so identified bounds on the unidentified quantity $G(z' \beta)$ are $0 < G(z' \beta) < 0.1$. ■

2.4 Estimating G in a Single-Index Model

We now turn to the problem of estimating G and β in the single-index model (2.1). It is assumed throughout the remainder of this chapter that G and β are identified. This section is concerned with estimating G . Estimation of β is dealt with in Sections 2.5 and 2.6.

Suppose, for the moment, that β is known. Then G can be estimated as the nonparametric mean regression of Y on $X' \beta$. There are many nonparametric mean-regression estimators that can be used. See, for example, Härdle (1990), Härdle and Linton (1994), and the Appendix. This chapter uses kernel estimators. The properties of these estimators are summarized in the Appendix.

To obtain a kernel estimator of $G(z)$ at any z in the support of $X' \beta$, let the data consist of a random sample of n observations of (Y, X) . Let $\{Y_i, X_i : i = 1, \dots, n\}$ denote the sample. Here, the subscript i indexes observations, not components of X . Define $Z_i = X_i' \beta$. Let K be a kernel function, and let $\{h_n\}$ be a sequence of bandwidth parameters. Under the assumption that β is known, the kernel nonparametric estimator of $G(z)$ is

$$G_n^*(z) = \frac{1}{nh_n p_n^*(z)} \sum_{i=1}^n Y_i K\left(\frac{z - Z_i}{h_n}\right), \quad (2.15)$$

where

$$p_n^*(z) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{z - Z_i}{h_n}\right). \quad (2.16)$$

The estimator (2.15) cannot be implemented in an application because β and, therefore, Z_i are not known. This problem can be remedied by replacing the unknown β with an estimator b_n . Define $Z_{ni} = X_i' b_n$ to be the corresponding estimator of Z_i . The resulting kernel estimator of G is

$$G_n(z) = \frac{1}{nh_n p_n(z)} \sum_{i=1}^n Y_i K\left(\frac{z - Z_{ni}}{h_n}\right), \quad (2.17)$$

where

$$p_n(z) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{z - Z_{ni}}{h_n}\right). \quad (2.18)$$

It is shown in Sections 2.5 and 2.6 that β can be estimated with a $n^{-1/2}$ rate of convergence in probability. That is, there exist estimators b_n with the property that $(b_n - \beta) = O_p(n^{-1/2})$. This is faster than the fastest possible rate of convergence in probability of a nonparametric estimator of $E(Y|X'\beta = z)$. As a result, the difference between the estimators G_n^* and G_n is asymptotically negligible. Specifically,

$$(nh_n)^{1/2}[G_n(z) - G(z)] = (nh_n)^{1/2}[G_n^*(z) - G(z)] + o_p(1)$$

for any z in the support of Z . Therefore, estimation of β has no effect on the asymptotic distributional properties of the estimator of G . The reasoning behind this conclusion is easily outlined. Let \tilde{b}_n and $\tilde{\beta}$, respectively, denote the vectors obtained from b_n and β by removing their first components (the components set by scale normalization). Let \tilde{X}_i be the vector obtained from X_i , by removing its first component. Define K' to be the derivative of the kernel function K . For any \tilde{b} and $b \equiv (1, \tilde{b}')'$, define

$$\begin{aligned} A_n(\tilde{b}) &= \frac{1}{nh_n} \sum_{i=1}^n Y_i K\left(\frac{z - X_i' b}{h_n}\right), \\ A_{nz}(\tilde{b}) &= -\frac{1}{nh_n^2} \sum_{i=1}^n Y_i K'\left(\frac{z - X_i' b}{h_n}\right) \tilde{X}_i, \\ \tilde{p}_n(\tilde{b}) &= \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{z - X_i' b}{h_n}\right), \end{aligned}$$

and

$$\tilde{p}_{nz}(\tilde{b}) = -\frac{1}{nh_n^2} \sum_{i=1}^n K' \left(\frac{z - X'_i b}{h_n} \right) \tilde{X}_i.$$

Now observe that $G_n(z) = A_n(\tilde{b}_n)/\tilde{p}_n(\tilde{b}_n)$ and $G_n^*(z) = A_n(\tilde{\beta})/\tilde{p}_n(\tilde{\beta})$. Therefore, a Taylor-series expansion of the right-hand side of (2.17) about $b_n = \beta$ yields

$$G_n(z) = G_n^*(z) + \left[\frac{A_{nz}(\tilde{b}_n^*)}{\tilde{p}_n(\tilde{b}_n^*)} - \frac{A_n(\tilde{b}_n^*)\tilde{p}_{nz}(\tilde{b}_n^*)}{\tilde{p}_{nz}^2(\tilde{b}_n^*)} \right] (\tilde{b}_n - \beta), \quad (2.19)$$

where \tilde{b}_n^* is between \tilde{b}_n and $\tilde{\beta}$. By using a suitable uniform law of large numbers (see, e.g., Pakes and Pollard 1989, Lemma 2.8), it can be shown that the quantity in brackets on the right-hand side of (2.19) converges in probability to a nonstochastic limit. Therefore, there is a nonstochastic function Γ such that

$$\frac{A_{nz}(\tilde{b}_n^*)}{\tilde{p}_n(\tilde{b}_n^*)} - \frac{A_n(\tilde{b}_n^*)\tilde{p}_{nz}(\tilde{b}_n^*)}{\tilde{p}_{nz}^2(\tilde{b}_n^*)} = \Gamma(z) + o_p(1). \quad (2.20)$$

It follows from (2.19), (2.20), and $b_n - \beta = O_p(n^{-1/2})$ that

$$G_n(z) - G_n^*(z) = \Gamma(z)(\tilde{b}_n - \tilde{\beta}) + o_p(\tilde{b}_n - \tilde{\beta}) = O_p(n^{-1/2}).$$

This implies that

$$(nh_n)^{1/2}[G_n(z) - G_n^*(z)] = O_p(h_n^{1/2}), \quad (2.21)$$

which gives the desired result

The foregoing results concerning estimation of G apply with any b_n that is a $n^{-1/2}$ -consistent estimator of β . We now turn to developing such estimators.

2.5 Optimization Estimators of β

Estimators of β can be classified according to whether they require solving nonlinear optimization problems. This section discusses estimators that are obtained as the solutions to nonlinear optimization problems. Section 2.6 discusses estimators that do not require solving optimization problems.

2.5.1 Nonlinear Least Squares

If G were known, then β could be estimated by nonlinear least squares or weighted nonlinear least squares (WNLS). Let the data consist of the random sample $\{Y_i, X_i : i = 1, \dots, n\}$. Then the WNLS estimator of β , b_{NLS} , is the solution to

$$\text{minimize: } S_n^*(b) = \frac{1}{n} \sum_{i=1}^n W(X_i) [Y_i - G(X_i' b)]^2, \quad (2.22)$$

where W is the weight function. Under mild regularity conditions, b_{NLS} is a consistent estimator of β , and $n^{1/2}(b_{NLS} - \beta)$ is asymptotically normally distributed with a mean of zero and a covariance matrix that can be estimated consistently. See, for example, Amemiya (1985), Davidson and MacKinnon (1993), and Gallant (1987).

The estimator b_{NLS} is not available in the semiparametric case, where G is unknown. Ichimura (1993) showed that this problem can be overcome by replacing G in (2.22) with a suitable estimator. This estimator is a modified version of the kernel estimator (2.17). Carroll et al. (1997) proposed using a local-linear estimator for a more elaborate model that includes a single-index model as a special case. Ichimura (1993) makes three modifications of the usual kernel estimator. First, observe that if G_n is defined as in (2.17), then the denominator of $G_n(X_i' b)$ contains the term $p_n(X_i' b)$. To keep this term from getting arbitrarily close to zero as n increases, it is necessary to restrict the sums in (2.17) and (2.22) to observations i for which the probability density of $X' \beta$ at the point $X_i' \beta$ exceeds a small, positive number. Second, observation i is excluded from the calculation of $G_n(X_i' b)$. Third, the terms of the sums in the calculation of G_n are weighted the same way that the terms in the sum (2.22) are weighted.

To carry out these modifications, let $p(\cdot, b)$ denote the probability density function of $X' b$. Let B be a compact set that contains β . Define A_x and A_{nx} to be the following sets:

$$A_x = \{x: p(x' b, b) \geq \eta \text{ for all } b \in B\}$$

and

$$A_{nx} = \{x: \|x - x^*\| \leq 2h_n \text{ for some } x^* \in A_x\},$$

where $\eta > 0$ is a constant, h_n is the bandwidth used for kernel estimation, and $\|\cdot\|$ is the Euclidean norm. A_{nx} contains A_x and shrinks toward A_x as $h_n \rightarrow 0$. Let I denote the indicator function. $I(\cdot) = 1$ if the event in parentheses occurs and 0 otherwise. Define $J_i = I(X_i \in A_x)$ and $J_{ni} = I(X_i \in A_{nx})$. Finally, define

$$G_{ni}(z, b) = \frac{1}{nh_n p_{ni}(z, b)} \sum_{j \neq i} J_{nj} W(X_j) Y_j K\left(\frac{z - X_j' b}{h_n}\right), \quad (2.23)$$

where for any z

$$p_{ni}(z, b) = \frac{1}{nh_n} \sum_{j \neq i} J_{nj} W(X_j) K \left(\frac{z - X'_j b}{h_n} \right). \quad (2.24)$$

The estimator of $G(X'_i b)$ that is used in (2.22) is $G_{ni}(X'_i b, b)$. Thus, the semiparametric WNLS estimator of β is the solution to

$$\text{minimize: } S_n(\tilde{b}) = \frac{1}{n} \sum_{i=1}^n J_i W(X_i) [Y_i - G_{ni}(X'_i b, b)]^2. \quad (2.25)$$

The minimization is over \tilde{b} , not b , to impose scale normalization. Let \tilde{b}_n denote the resulting estimator, and call it the semiparametric WNLS estimator of $\tilde{\beta}$.

Ichimura (1993) gives conditions under which \tilde{b}_n is a consistent estimator of $\tilde{\beta}$ and

$$n^{1/2}(\tilde{b}_n - \tilde{\beta}) \xrightarrow{d} N(0, \Omega). \quad (2.26)$$

The covariance matrix, Ω , is given in (2.28) below. The conditions under which (2.26) holds are stated in Theorem 2.2.

Theorem 2.2: *Equation (2.26) holds if the following conditions are satisfied:*

- (a) $\{Y_i, X_i : i = 1, \dots, n\}$ is a random sample from a distribution that satisfies (2.1).
- (b) β is identified and is an interior point of the known compact set B .
- (c) A_x is compact, and W is bounded and positive on A_x .
- (d) $E(Y|X'b = z)$ and $p(z, b)$ are three times continuously differentiable with respect to z . The third derivatives are Lipschitz continuous uniformly over B for all $z \in \{z: z = x'b, b \in B, x \in A_x\}$.
- (e) $E|Y|^m < \infty$ for some $m \geq 3$. The variance of Y conditional on $X = x$ is bounded and bounded away from 0 for $x \in A_x$.
- (f) The kernel function K is twice continuously differentiable, and its second derivative is Lipschitz continuous. Moreover $K(v) = 0$ if $|v| > 1$, and

$$\int_{-1}^1 v^j K(v) dv = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{if } j = 1 \end{cases}.$$

- (g) The bandwidth sequence $\{h_n\}$ satisfies $(\log h_n) / [nh_n^{3+3/(m-1)}] \rightarrow 0$ and $nh_n^8 \rightarrow 0$ as $n \rightarrow \infty$. ■

There are several noteworthy features of Theorem 2.2. First, \tilde{b}_n converges in probability to $\tilde{\beta}$ at the rate $n^{-1/2}$, which is the same rate that would be obtained if G were known and faster than the rate of convergence of a nonparametric density

or mean-regression estimator. This result was used in deriving (2.21). Second, the asymptotic distribution of $n^{1/2}(\tilde{b}_n - \tilde{\beta})$ is centered at zero. This contrasts with the case of nonparametric density and mean-regression estimators, whose asymptotic distributions are not centered at zero in general when the estimators have their fastest possible rates of convergence. Third, the range of permissible rates of convergence of h_n includes the rate $n^{-1/5}$, which is the standard rate in nonparametric density and mean-regression estimation. Finally, Theorem 2.2 requires β to be contained in the known, compact set B . Therefore, in principle $S_n(\tilde{b})$ should be minimized subject to the constraint $\tilde{b} \in B$. In practice, however, the probability that the constraint is binding for any reasonable B is so small that it can be ignored. This is a useful result because solving a constrained nonlinear optimization problem is usually much more difficult than solving an unconstrained one.

Stating the covariance matrix, Ω , requires additional notation. Let $p(\cdot | \tilde{x}, b)$ denote the probability density function of $X'b$ conditional on $\tilde{X} = \tilde{x}$. Define $p(\cdot | \tilde{x}) = p(\cdot | \tilde{x}, \beta)$, $\sigma^2(x) = \text{Var}(Y|X = x)$, and

$$G(z, b) = \text{plim}_{n \rightarrow \infty} G_{ni}(z, b).$$

Calculations that are lengthy but standard in kernel estimation show that

$$\begin{aligned} G(z, b) &= \frac{E[E(Y|X'b = z, \tilde{X})I(X \in A_x)W(X)p(z|\tilde{X}, b)]}{E[I(X \in A_x)W(X)p(z|\tilde{X}, b)]} \\ &= \frac{R_1(z, b)}{R_2(z, b)}, \end{aligned}$$

where

$$R_1(z, b) = E\{G[z - \tilde{X}'(\tilde{b} - \tilde{\beta})]p[z - \tilde{X}'(\tilde{b} - \tilde{\beta})|\tilde{X}]W(X)I(X \in A_x)\}$$

and

$$R_2(z, b) = E\{p[z - \tilde{X}'(\tilde{b} - \tilde{\beta})|\tilde{X}]W(X)I(X \in A_x)\}.$$

Moreover,

$$G(z, \beta) = G(z)$$

and for $z = x'\beta$

$$\frac{\partial G(z, \beta)}{\partial \tilde{b}} = G'(z) \left\{ \tilde{x} - \frac{E[\tilde{X}W(X)|X'\beta = z, X \in A_x]}{E[W(X)|X'\beta = z, X \in A_x]} \right\}. \quad (2.27)$$

Now define

$$C = 2E \left[I(X \in A_x) W(X) \frac{\partial G(X' \beta, \beta)}{\partial \tilde{b}} \frac{\partial G(X' \beta, \beta)}{\partial \tilde{b}'} \right]$$

and

$$D = 4E \left[I(X \in A_x) W^2(X) \sigma^2(x) \frac{\partial G(X' \beta, \beta)}{\partial \tilde{b}} \frac{\partial G(X' \beta, \beta)}{\partial \tilde{b}'} \right].$$

Then

$$\Omega = C^{-1} D C^{-1}. \quad (2.28)$$

Theorem 2.2 is proved in Ichimura (1993). The technical details of the proof are complex, but the main ideas are straightforward and based on the familiar Taylor-series methods of asymptotic distribution theory. With probability approaching one as $n \rightarrow \infty$, the solution to (2.25) satisfies the first-order condition

$$\frac{\partial S_n(\tilde{b}_n)}{\partial \tilde{b}} = 0.$$

Therefore, a Taylor-series expansion gives

$$n^{1/2} \frac{\partial S_n(\tilde{\beta})}{\partial \tilde{\beta}} = - \frac{\partial^2 S_n(\tilde{b}_n)}{\partial \tilde{b} \partial \tilde{b}'} n^{1/2} (\tilde{b}_n - \tilde{\beta}), \quad (2.29)$$

where \tilde{b}_n is between \tilde{b}_n and $\tilde{\beta}$. Now consider the left-hand side of (2.29). Differentiation of S_n gives

$$n^{1/2} \frac{\partial S_n(\tilde{\beta})}{\partial \tilde{b}} = - \frac{2}{n^{1/2}} \sum_{i=1}^n J_i W(X_i) [Y_i - G_{ni}(X_i' \beta, \beta)] \frac{\partial G_{ni}(X_i' \beta, \beta)}{\partial \tilde{b}}.$$

Moreover,

$$G_{ni}(X_i' \beta, \beta) \xrightarrow{p} G(X_i' \beta)$$

and

$$\frac{\partial G_{ni}(X_i' \beta, \beta)}{\partial \tilde{b}} \xrightarrow{p} \frac{\partial G(X_i' \beta, \beta)}{\partial \tilde{b}}$$

sufficiently rapidly that we may write

$$n^{1/2} \frac{\partial S_n(\tilde{\beta})}{\partial \tilde{b}} = - \frac{2}{n^{1/2}} \sum_{i=1}^n J_i W(X_i) [Y_i - G(X_i' \beta)] \frac{\partial G(X_i' \beta, \beta)}{\partial \tilde{b}} + o_p(1). \quad (2.30)$$

The first term on the right-hand side of (2.30) is asymptotically distributed as $N(0, D)$ by the multivariate generalization of the Lindeberg–Levy central limit theorem. Therefore, the left-hand side of (2.29) is also asymptotically distributed as $N(0, D)$.

Now consider the right-hand side of (2.29). Differentiation of S_n gives

$$\begin{aligned} \frac{\partial^2 S_n(\bar{b}_n)}{\partial \tilde{b} \partial \tilde{b}'} &= \frac{2}{n} \sum_{i=1}^n J_i W(X_i) \frac{\partial G_{ni}(X_i' \bar{b}_n, \bar{b}_n)}{\partial \tilde{b}} \frac{\partial G_{ni}(X_i' \bar{b}_n, \bar{b}_n)}{\partial \tilde{b}'} \\ &\quad - \frac{2}{n} \sum_{i=1}^n J_i W(X_i) [Y_i - G_{ni}(X_i' \bar{b}_n, \bar{b}_n)] \frac{\partial^2 G_{ni}(X_i' \bar{b}_n, \bar{b}_n)}{\partial \tilde{b} \partial \tilde{b}'} . \end{aligned}$$

Because $G_{ni}(x'b, b)$ and its derivatives converge to $G(x'b, b)$ and its derivatives uniformly over both arguments, we may write

$$\begin{aligned} \frac{\partial^2 S_n(\bar{b}_n)}{\partial \tilde{b} \partial \tilde{b}'} &= \frac{2}{n} \sum_{i=1}^n J_i W(X_i) \frac{\partial G(X_i' \beta, \beta)}{\partial \tilde{b}} \frac{\partial G(X_i' \beta, \beta)}{\partial \tilde{b}'} \\ &\quad - \frac{2}{n} \sum_{i=1}^n J_i W(X_i) [Y_i - G(X_i' \beta, \beta)] \frac{\partial^2 G(X_i' \beta, \beta)}{\partial \tilde{b} \partial \tilde{b}'} + o_p(1). \end{aligned}$$

The first term on the right-hand side of this equation converges almost surely to C and the second term converges almost surely to zero by the strong law of large numbers. This result together with the previously obtained asymptotic distribution of the left-hand side of (2.29) implies that (2.29) can be written in the form

$$N(0, D) = Cn^{1/2}(\tilde{b}_n - \tilde{\beta}) + o_p(1). \quad (2.31)$$

Equation (2.26) is obtained by multiplying both sides of (2.31) by C^{-1} .

In applications, Ω is unknown, and a consistent estimator is needed to make statistical inference possible. To this end, define

$$C_n = \frac{2}{n} \sum_{i=1}^n J_i W(X_i) \frac{\partial G_{ni}(X_i' b_n, b_n)}{\partial \tilde{b}} \frac{\partial G_{ni}(X_i' b_n, b_n)}{\partial \tilde{b}'}$$

and

$$D_n = \frac{4}{n} \sum_{i=1}^n J_i W(X_i) [Y_i - G_{ni}(X_i' b_n)]^2 \frac{\partial G_{ni}(X_i' b_n, b_n)}{\partial \tilde{b}} \frac{\partial G_{ni}(X_i' b_n, b_n)}{\partial \tilde{b}'}.$$

Under the assumptions of Theorem 2.2, C_n and D_n , respectively, are consistent estimators of C and D . Ω is estimated consistently by

$$\Omega_n = C_n^{-1} D_n C_n^{-1}.$$

Intuitively, these results can be understood by observing that because G_{ni} converges in probability to G and b_n converges in probability to β ,

$$C_n = \frac{2}{n} \sum_{i=1}^n J_i W(X_i) \frac{\partial G(X_i' \beta, \beta)}{\partial \tilde{b}} \frac{\partial G(X_i' \beta, \beta)}{\partial \tilde{b}'} + o_p(1)$$

and

$$D_n = \frac{4}{n} \sum_{i=1}^n J_i W(X_i) [Y_i - G(X_i' \beta)]^2 \frac{\partial G(X_i' \beta, \beta)}{\partial \tilde{b}} \frac{\partial G(X_i' \beta, \beta)}{\partial \tilde{b}'} + o_p(1).$$

Convergence of C_n to C and D_n to D now follows from the strong law of large numbers.

2.5.2 Choosing the Weight Function

The choice of weight function, W , affects the efficiency of the estimator of $\tilde{\beta}$. Ideally, one would like to choose W so as to maximize the asymptotic efficiency of the estimator. Some care is needed in defining the concept of asymptotic efficiency so as to avoid the pathology of superefficiency. See Bickel et al. (1993) and Ibragimov and Has'minskii (1981) for discussions of superefficiency and methods for avoiding it. Estimators that are restricted so as to avoid superefficiency are called *regular*.

Within the class of semiparametric WNLS estimators, an estimator is asymptotically efficient if the covariance matrix Ω of its asymptotic distribution differs from the covariance matrix Ω^* of any other weighted WNLS estimator by a positive-semidefinite matrix. That is, $\Omega^* - \Omega$ is positive semidefinite. More generally, one can consider the class of all regular estimators of single-index models (2.1). This class includes estimators that may not be semiparametric WNLS estimators. The definition of an asymptotically efficient estimator remains the same, however. The covariance matrix of the asymptotic distribution of any regular estimator exceeds that of the asymptotically efficient estimator by a positive-semidefinite matrix.

The problem of asymptotically efficient estimation of β in a semiparametric single-index model is related to but more difficult than the problem of asymptotically efficient estimation in a nonlinear regression model with a known G . The case of a nonlinear regression model (not necessarily a single-index model) in which G is known has been investigated by Chamberlain (1987), who derived an asymptotic efficiency bound. The covariance matrix of the asymptotic distribution of any regular estimator must exceed this bound by a positive-semidefinite matrix. The model is $E(Y|X = x) = G(x, \beta)$. The variance function, $\sigma^2(x) = E\{[Y - G(X, \beta)]^2 | X = x\}$, is unknown. Chamberlain (1986) showed that the efficiency bound is

$$\Omega_{NLR} = \left\{ E \left[\frac{1}{\sigma^2(X)} \frac{\partial G(X, \beta)}{\partial b} \frac{\partial G(X, \beta)}{\partial b'} \right] \right\}^{-1}.$$

This is the covariance matrix of a weighted (or generalized) nonlinear least-squares estimator of β with weight function $W(x) = 1/\sigma^2(x)$. For the special case of the linear model $G(x, \beta) = x\beta$, Carroll (1982) and Robinson (1987) showed that this covariance matrix is obtained asymptotically even if $\sigma^2(x)$ is unknown by replacing $\sigma^2(x)$ with a nonparametric estimator. Thus, lack of knowledge of $\sigma^2(x)$ causes no loss of asymptotic efficiency relative to infeasible generalized least-squares estimation.

The problem of efficient estimation of β in a single-index model with an unknown G has been investigated by Hall and Ichimura (1991) and Newey and Stoker (1993). These authors showed that under regularity conditions, the efficiency bound for estimating β in a single-index model with unknown G and using only data for which $X \in A_x$ is (2.28) with weight function $W(x) = 1/\sigma^2(x)$. With this weight function, $C = D$ in (2.28), so the efficiency bound is

$$\Omega_{SI} = \left\{ E \left[\frac{I(X \in A_x)}{\sigma^2(X)} \frac{\partial G(X'\beta, \beta)}{\partial \tilde{b}} \frac{\partial G(X'\beta, \beta)}{\partial \tilde{b}'} \right] \right\}^{-1}. \quad (2.32)$$

This bound is achieved by the semiparametric WNLS estimator if $\sigma^2(X)$ is known or independent of X . The assumption that the estimator uses only observations for which $X \in A_x$ can be eliminated by letting A_x grow very slowly as n increases. Chamberlain (1986) and Cosslett (1987) derived this asymptotic efficiency bound for the case in which (2.1) is a binary-response model (that is, the only possible values of Y are 0 and 1) and G is a distribution function. Chamberlain and Cosslett also derived efficiency bounds for certain kinds of censored regression models. Except in special cases, Ω_{SI} exceeds the asymptotic efficiency bound that would be achievable if G were known. Thus, there is a cost in terms of asymptotic efficiency (but not rate of convergence of the estimator) for not knowing G . Cosslett (1987) gives formulae for the efficiency losses in binary-response and censored linear regression models.

When $\sigma^2(x)$ is unknown, as is likely in applications, it can be replaced by a consistent estimator. Call this estimator $s_n^2(x)$. The asymptotic efficiency bound will be achieved by setting $W(x) = 1/s_n^2(x)$ in the semiparametric WNLS estimator (Newey and Stoker 1993). Therefore, an asymptotically efficient estimator of β can be obtained even when $\sigma^2(x)$ is unknown.

A consistent estimator of $\sigma^2(x)$ can be obtained by using the following two-step procedure. In the first step, estimate β by using semiparametric WNLS with $W(x) = 1$. The resulting estimator is $n^{-1/2}$ -consistent and asymptotically normal but inefficient. Let e_i be the i th residual from the estimated model. That is, $e_i = Y_i - G_{ni}(X_i' b_n, b_n)$. In the second step, set $s_n^2(x)$ equal to a nonparametric estimator of the mean regression of e_i^2 on X_i . Robinson (1987) discusses technical problems that arise if X has unbounded support or a density that can be arbitrarily close to zero. He avoids these problems by using a nearest-neighbor nonparametric regression estimator. In practice, a kernel estimator will suffice if A_x is chosen so as to keep the estimated density of X away from zero.

This concludes the discussion of semiparametric weighted nonlinear least-squares estimation of single-index models. To summarize, Ichimura (1993) has given conditions under which the semiparametric WNLS estimator of β in (2.1) is $n^{-1/2}$ -consistent and asymptotically normal. The estimator of β is also asymptotically efficient if the weight function is a consistent estimator of $1/\sigma^2(x)$. A consistent estimator of $\sigma^2(x)$ can be obtained by a two-step procedure in which the first step is semiparametric WNLS estimation of β with a unit weight function and the second step is nonparametric estimation of the mean of the squared first-step residuals conditional on X .

2.5.3 Semiparametric Maximum-Likelihood Estimation of Binary-Response Models

This section is concerned with estimation of (2.1) when the only possible values of Y are 0 and 1. In this case, $G(x'\beta) = P(Y = 1|X = x)$. If G were a known function, then the asymptotically efficient estimator of β would be the maximum-likelihood estimator (MLE). The MLE solves the problem

$$\text{maximize: } \log L(b) = \frac{1}{n} \sum_{i=1}^n \{Y_i \log G(X_i'b) + (1 - Y_i) \log [1 - G(X_i'b)]\}. \quad (2.33)$$

In the semiparametric case, where G is unknown, one can consider replacing G on the right-hand side of (2.33) with an estimator such as G_{ni} in (2.23). This idea has been investigated in detail by Klein and Spady (1993). It is clear from (2.33) that care must be taken to ensure that any estimate of G is kept sufficiently far from 0 and 1. Klein and Spady (1993) use elaborate trimming procedures to accomplish this without artificially restricting X to a fixed set A_x on which $G(X'\beta)$ is bounded away from 0 and 1. They find, however, that trimming has little effect on the numerical performance of the resulting estimator. Therefore, in practice little is lost in terms of estimation efficiency and much is gained in simplicity by using only observations for which $x \in A_x$. This method will be used in the remainder of this section.

A second simplification can be obtained by observing that in the special case of a binary-response model, $\text{Var}(Y|X = x) = G(x'\beta)[1 - G(x'\beta)]$. Thus, $\sigma^2(x)$ depends only on the index $z = x'\beta$. In this case, W cancels out of the numerator and denominator terms on the right-hand side of (2.27), so

$$\frac{\partial G(z, \beta)}{\partial \tilde{b}} = G'(z)\{\tilde{x} - E[\tilde{X}|X'\beta = z, X \in A_x]\}.$$

By substituting this result into (2.28) and (2.32), it can be seen that the covariance matrix of the asymptotic distribution of the semiparametric WNLS estimator of β is the same whether the estimator of G is weighted or not. Moreover, the asymptotic

efficiency bound Ω_{SI} can be achieved without weighting the estimator of G . Accordingly, define the unweighted estimator of G

$$\hat{G}_{ni}(z, b) = \frac{1}{nh_n \hat{p}_{ni}(z, b)} \sum_{j \neq i} J_{nj} Y_j K\left(\frac{z - X'_j b}{h_n}\right),$$

where

$$\hat{p}_{ni}(z, b) = \frac{1}{nh_n} \sum_{j \neq i} J_{nj} K\left(\frac{z - X'_j b}{h_n}\right).$$

Now consider the following semiparametric analog of (2.33):

$$\text{maximize: } \log L_{SP}(\tilde{b}) = \frac{1}{n} \sum_{i=1}^n J_i \{Y_i \log \hat{G}_{ni}(X'_i b, b) + (1 - Y_i) \log [1 - \hat{G}_{ni}(X'_i b, b)]\}. \quad (2.34)$$

Let \tilde{b}_n denote the resulting estimator of $\tilde{\beta}$. If β is identified (see the discussion in Section 2.3), consistency of \tilde{b}_n for $\tilde{\beta}$ can be demonstrated by showing that $\hat{G}_{ni}(z, b)$ converges to $G(z, b)$ uniformly over z and b . Therefore, the probability limit of the solution to (2.34) is the same as the probability limit of the solution to

$$\text{maximize: } \log L_{SP}^*(\tilde{b}) = \frac{1}{n} \sum_{i=1}^n J_i \{Y_i \log G(X'_i b, b) + (1 - Y_i) \log [1 - G(X'_i b, b)]\}. \quad (2.35)$$

The solution to (2.34) is consistent for $\tilde{\beta}$ if the solution to (2.35) is. The solution to (2.35) is a parametric maximum-likelihood estimator. Consistency for $\tilde{\beta}$ can be proved using standard methods for parametric maximum-likelihood estimators. See, for example, Amemiya (1985).

By differentiating the right-hand side of (2.34), it can be seen that $b_n \equiv (1, \tilde{b}_n')'$ satisfies the first-order condition

$$\frac{1}{n} \sum_{i=1}^n J_i \frac{Y_i - \hat{G}_{ni}(X'_i b_n, b_n)}{\hat{G}_{ni}(X_i b_n, b_n)[1 - \hat{G}_{ni}(X_i b_n, b_n)]} \frac{\partial \hat{G}_{ni}(X'_i b_n, b_n)}{\partial \tilde{b}} = 0$$

with probability approaching 1 as $n \rightarrow \infty$. This is the same as the first-order condition for semiparametric WNLS estimation of β with the estimated weight function

$$\begin{aligned} W(x) &= \{\hat{G}_{ni}(x' b_n, b_n)[1 - \hat{G}_{ni}(x' b_n, b_n)]\}^{-1} \\ &= \{G(x' \beta)[1 - G(x' \beta)]\}^{-1} + o_p(1) \\ &= [\text{Var}(Y|X = x)]^{-1} + o_p(1). \end{aligned}$$

It now follows from the discussion of asymptotic efficiency in semiparametric WNLS estimation (Section 2.5.2) that the semiparametric maximum-likelihood

estimator of β in a single-index binary-response model achieves the asymptotic efficiency bound Ω_{SI} .

The conclusions of this section may be summarized as follows. The semiparametric maximum-likelihood estimator of β in a single-index binary-response model solves (2.34). The estimator is asymptotically efficient and satisfies

$$n^{1/2}(\hat{b}_n - \tilde{\beta}) \xrightarrow{d} N(0, \Omega_{SI}).$$

2.5.4 Semiparametric Maximum-Likelihood Estimation of Other Single-Index Models

Ai (1997) has extended semiparametric maximum-likelihood estimation to single-index models other than binary-response models. As in the binary-response estimator of Klein and Spady (1993), Ai (1997) forms a quasi-likelihood function by replacing the unknown probability density function of the dependent variable conditional on the index with a nonparametric estimator. To illustrate, suppose that the probability distribution of the dependent variable Y depends on the explanatory variables X only through the index $X'\beta$. Let $f(\cdot | v, \beta)$ denote the probability density function of Y conditional on $X'\beta = v$. If f were known, then β could be estimated by parametric maximum likelihood. For the semiparametric case, in which f is unknown, Ai replaces f with a kernel estimator of the density of Y conditional on the index. He then maximizes a trimmed version of the resulting quasi-likelihood function. Under suitable conditions, the resulting semiparametric estimator of β is asymptotically efficient (in the sense of achieving the semiparametric efficiency bound). See Ai (1997) for the details of the trimming procedure and regularity conditions.

Ai and Chen (2003) have given conditions for asymptotically efficient estimation of β in the moment condition model

$$E[\rho(Z, \beta, g(\cdot))|X] = 0, \quad (2.36)$$

where $Z = (Y', X'_Z)'$, Y is a random vector, X_Z is a subvector of the random vector X , ρ is a vector of known functions, β is an unknown finite-dimensional parameter, and g is a finite-dimensional vector of unknown functions that may include β among their arguments. Model (2.36) is very general and includes single-index models, partially linear models, and many others as special cases. The cost of this generality, however, is that the analysis of (2.36) is both lengthy and complicated. The details are given in Ai and Chen (2003).

2.5.5 Semiparametric Rank Estimators

If G in (2.1) is a nondecreasing function and $Y - G(X'\beta)$ is independent of X , then $X'_i\beta > X'_j\beta$ implies that $P(Y_i > Y_j) > P(Y_j > Y_i)$. This suggests estimating β by

choosing the estimator b_n so as to make the rank ordering of $\{Y_i: i = 1, \dots, n\}$ as close as possible to that of $\{X'_i\beta: i = 1, \dots, n\}$. The resulting maximum rank correlation (MRC) estimator is

$$b_{n, \text{MRC}} = \arg \max_b \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n I(Y_i > Y_j) I(X'_i b > X'_j b).$$

The MRC estimator was first proposed by Han (1987), who also gave conditions for consistency of the estimator. Cavanagh and Sherman (1998) proposed a modified estimator

$$b_{n, \text{CS}} = \arg \max_b \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n M(Y_i) I(X'_i b > X'_j b),$$

where M is an increasing function. This estimator is consistent under conditions that are weaker than those required for the MRC estimator. It is also easier to compute than the MRC estimator.

Deriving the asymptotic distributions of these estimators is complicated because their objective functions are discontinuous. Sherman (1993) gave conditions under which $n^{1/2}(b_{n, \text{MRC}} - \beta)$ is asymptotically normally distributed with mean 0. Cavanagh and Sherman (1998) gave conditions for asymptotic normality of $n^{1/2}(b_{n, \text{CS}} - \beta)$. The derivation of these results relies on empirical process methods that are beyond the scope of this book. Sherman (1993) and Cavanagh and Sherman (1998) also give methods for estimating the covariance matrices of the asymptotic distributions of $n^{1/2}(b_{n, \text{MRC}} - \beta)$ and $n^{1/2}(b_{n, \text{CS}} - \beta)$, but these are hard to implement. Subbotin (2008) proves that the bootstrap estimates these distributions consistently, which makes the bootstrap a potentially attractive method for carrying out inference with $b_{n, \text{MRC}}$ and $b_{n, \text{CS}}$ in applied research. Rank estimators are not asymptotically efficient and can be hard to compute, but they do not require bandwidths or other smoothing parameters. This may be an advantage in some applications.

2.6 Direct Semiparametric Estimators

Semiparametric weighted nonlinear least-squares and maximum-likelihood estimators have the significant practical disadvantage of being very difficult to compute. This is because they are solutions to nonlinear optimization problems whose objective functions may be nonconvex (nonconcave in the case of the maximum-likelihood estimator) or multimodal. Moreover, computing the objective functions requires estimating a nonparametric mean regression at each data point and, therefore, can be very slow.

This section describes an estimation approach that does not require solving an optimization problem and is noniterative (hence the name *direct*). Direct estimates

can be computed very quickly. Although direct estimators are not asymptotically efficient, an asymptotically efficient estimator can be obtained from a direct estimator in one additional, noniterative, computational step. The relative computational simplicity of direct estimators makes them highly attractive for practical data analysis.

Xia et al. (2002) have proposed an iterative scheme called rMAVE (Refined Minimum Average Conditional Variance Estimation) that simplifies the computations of semiparametric weighted nonlinear least-squares and maximum-likelihood estimators. Xia (2006) has given conditions under which rMAVE yields the asymptotically efficient estimator of a single-index model. However, these conditions include starting the rMAVE iterations at a point b_0 that satisfies $\|b_0 - \beta\| = o(n^{-9/20})$. Consequently, rMAVE has no apparent advantages over taking an additional step beyond one of the direct estimators that is described in this section.

Section 2.6.1 describes a well-known direct estimation method under the assumption that X is a continuously distributed random vector. Section 2.6.2 describes a direct estimation method that overcomes an important disadvantage of the method of Section 2.6.1 though at the cost of additional complexity. Section 2.6.3 shows how the direct estimation method can be extended to models in which some components of X are discrete. Section 2.6.4 describes the one-step method for obtaining an asymptotically efficient estimator from a direct estimate.

2.6.1 Average-Derivative Estimators

The idea underlying direct estimation of a single-index model when X is a continuously distributed random vector is very simple. Let (2.1) hold. Assume that G is differentiable, as is required for identification of β . Then

$$\frac{\partial E(Y|X=x)}{\partial x} = \beta G'(x'\beta). \quad (2.37)$$

Moreover, for any bounded, continuous function W ,

$$E \left[W(X) \frac{\partial E(Y|X)}{\partial x} \right] = \beta E [W(X) G'(X'\beta)]. \quad (2.38)$$

The quantity on the left-hand side of (2.38) is called a *weighted average derivative* of $E(Y|X)$ with weight function W . Equation (2.38) shows that a weighted average derivative of $E(Y|X)$ is proportional to β . Owing to the need for scale normalization, β is identified only up to scale, so any weighted average derivative of $E(Y|X)$ is observationally equivalent to β . Thus, to estimate β , it suffices to estimate the left-hand side of (2.38) for some W . The scale normalization $\beta_1 = 1$ can be imposed, if desired, by dividing each component of the left-hand side of (2.38) by the first component.

The left-hand side of (2.38) can be estimated by replacing $\partial E(Y|X)/\partial x$ with a kernel (or other nonparametric) estimator and the population expectation $E(\cdot)$ with a sample average. Hristache et al. (2001), Härdle and Stoker (1989), Powell et al. (1989), and Stoker (1986, 1991a,b) describe various ways of doing this. The discussion in this section concentrates on the method of Powell et al. (1989), which is especially easy to analyze and implement. Section 2.6.2 describes the method of Hristache et al. (2001), which overcomes an important disadvantage of the method of Powell et al. (1989).

To describe the method of Powell et al. (1989), let $p(\cdot)$ denote the probability density function of X , and set $W(x) = p(x)$. Then the left-hand side of (2.38) can be written in the form

$$\begin{aligned} E \left[W(X) \frac{\partial E(Y|X)}{\partial x} \right] &= E \left[p(X) \frac{\partial E(Y|X)}{\partial x} \right] \\ &= \int \frac{\partial E(Y|X=x)}{\partial x} p(x)^2 dx. \end{aligned}$$

Assume that $p(x) = 0$ if x is on the boundary of the support of X . Then integration by parts gives

$$\begin{aligned} E \left[W(X) \frac{\partial E(Y|X)}{\partial x} \right] &= -2 \int E(Y|X=x) \frac{\partial p(x)}{\partial x} p(x) dx \\ &= -2E \left[Y \frac{\partial p(X)}{\partial x} \right]. \end{aligned}$$

Define

$$\delta = -2E[Y \partial p(X)/\partial x]. \quad (2.39)$$

Then δ is observationally equivalent to β up to scale normalization. A consistent estimator of δ can be obtained by replacing p with a nonparametric estimator and the expectation operator with a sample average. Let $\{Y_i, X_i: i = 1, \dots, n\}$ denote the sample. The estimator of δ is

$$\delta_n = -2 \sum_{i=1}^n Y_i \frac{\partial p_{ni}(X_i)}{\partial x}, \quad (2.40)$$

where $p_{ni}(X_i)$ is the estimator of $p(X_i)$. The quantity δ_n is called a *density-weighted average-derivative estimator*.

To implement (2.40), the estimator of p must be specified. A kernel estimator is attractive because it is relatively easily analyzed and implemented. To this end, let $d = \dim(X)$, and let K be a kernel function with a d -dimensional argument. Conditions that K must satisfy are given in Theorem 2.3 below. Let $\{h_n\}$ be a sequence of bandwidth parameters. Set

$$p_{ni}(x) = \frac{1}{n-1} \frac{1}{h_n^d} \sum_{j \neq i} K\left(\frac{x - X_j}{h_n}\right).$$

It follows from the properties of kernel density estimators (see the Appendix) that $p_{ni}(x)$ is a consistent estimator of $p(x)$. Moreover, $\partial p(x)/\partial x$ is estimated consistently by $\partial p_{ni}(x)/\partial x$. The formula for $\partial p_{ni}(x)/\partial x$ is

$$\frac{\partial p_{ni}(x)}{\partial x} = \frac{1}{n-1} \frac{1}{h_n^{d+1}} \sum_{j \neq i} K'\left(\frac{x - X_j}{h_n}\right), \quad (2.41)$$

where K' denotes the gradient of K . Substituting (2.41) into (2.40) yields

$$\delta_n = -\frac{2}{n(n-1)} \frac{1}{h_n^{d+1}} \sum_{i=1}^n \sum_{j \neq i} Y_i K'\left(\frac{X_i - X_j}{h_n}\right). \quad (2.42)$$

Observe that the right-hand side of (2.42) does not have a density estimator or other random variable in its denominator. This is because setting $W(x) = p(x)$ in the weighted average derivative defined in (2.38) cancels the density function that would otherwise be in the denominator of the estimator of $E(Y|X = x)$. This lack of a random denominator is the main reason for the relative ease with which δ_n can be analyzed and implemented.

Powell et al. (1989) give conditions under which δ_n is a consistent estimator of δ and $n^{1/2}(\delta_n - \delta)$ is asymptotically normally distributed with mean 0. The formal statement of this result and the conditions under which it holds are given in Theorem 2.3. Let $\|\cdot\|$ denote the Euclidean norm. Let $P = (d+2)/2$ if d is even and $P = (d+3)/2$ if d is odd.

Theorem 2.3: *Let the following conditions hold.*

- (a) *The support of X is a convex, possibly unbounded, subset of \mathbb{R}^d with a nonempty interior. X has a probability density function p . All partial derivatives of p up to order $P+1$ exist.*
- (b) *The components of $\partial E(Y|X)/\partial x$ and of the matrix $[\partial p(X)/\partial x](Y, X')$ have finite second moments. $E[Y\partial^r p(X)]$ exists for all positive integers $r \leq P+1$, where $\partial^r p(x)$ denotes any order r mixed partial derivative of p . $E(Y^2|X = x)$ is a continuous function of x . There is a function $m(x)$ such that*

$$E[(1 + |Y| + \|X\|)m(X)]^2 < \infty,$$

$$\left\| \frac{\partial p(x + \zeta)}{\partial x} - \frac{\partial p(x)}{\partial x} \right\| < m(x) \|\zeta\|,$$

and

$$\left\| \frac{\partial p(x + \zeta)E(Y|X = x + \zeta)}{\partial x} - \frac{\partial p(x)E(Y|X = x)}{\partial x} \right\| < m(x) \|\zeta\|.$$

- (c) The kernel function K is symmetrical about the origin, bounded, and differentiable. The moments of K through order P are finite. The moments of K of order r are all 0 if $1 \leq r < P$. In addition

$$\int K(v)dv = 1.$$

- (d) The bandwidth sequence $\{h_n\}$ satisfies $nh_n^{2P} \rightarrow 0$ and $nh_n^{d+2} \rightarrow \infty$ as $n \rightarrow \infty$. Then

$$n^{1/2}(\delta_n - \delta) \xrightarrow{d} N(0, \Omega_{AD}),$$

where

$$\Omega_{AD} = 4E[R(Y, X)R(Y, X)'] - 4\delta\delta' \quad (2.43)$$

and

$$R(y, x) = p(x) \frac{\partial E(Y|X = x)}{\partial x} - [Y - E(Y|X = x)] \frac{\partial p(x)}{\partial x}. \blacksquare$$

A consistent estimator of Ω_{AD} is given in (2.44) below.

Several comments may be made about the conditions imposed in Theorem 2.3. Condition (a) implies that X is a continuously distributed random variable and that no component of X is functionally determined by other components. Condition (b) requires the existence of various moments and imposes smoothness requirements on $p(x)$, $E(Y|X = x)$, and $E(Y^2|X = x)$. Condition (c) requires K to be a *higher-order* kernel, meaning that some of its even moments vanish. In condition (c), the order is P . Higher-order kernels are used in density estimation and nonparametric mean regression to reduce bias. See the Appendix for further discussion of this use of higher-order kernels. Here, the higher-order kernel is used to make the bias of δ_n have size $o(n^{-1/2})$, which is needed to ensure that the asymptotic distribution of $n^{1/2}(\delta_n - \delta)$ is centered at 0. Finally, the rate of convergence of h_n is faster than would be optimal if the aim were to estimate $p(x)$ or $E(Y|X = x)$ nonparametrically. Under the conditions of Theorem 2.3, the rate of convergence in probability of an estimator of $p(x)$ or $E(Y|X = x)$ is maximized by setting $h_n \propto n^{-1/(2P+d)}$, which is too slow to satisfy the requirement in condition (d) that $nh_n^{2P} \rightarrow 0$ as $n \rightarrow \infty$. The relatively fast rate of convergence of h_n required by condition (d), like the higher-order kernel required by condition (c), is needed to prevent the asymptotic distribution of $n^{1/2}(\delta_n - \delta)$ from having a nonzero mean.

Kernel density and mean-regression estimators cannot achieve $O_p(n^{-1/2})$ rates of convergence, so it may seem surprising that δ_n achieves this rate. The fast convergence of δ_n is possible because the sum over i on the right-hand side of (2.42) makes δ_n an average of kernel estimators. Averages of kernel estimators can achieve faster rates of convergence than kernel estimators that are not averaged.

A consistent estimator of Ω_{AD} can be obtained from (2.43) by replacing δ with δ_n , the population expectation with a sample average, and R with a consistent estimator. Powell et al. (1989) give the details of the calculation. The result is that Ω_{AD} is estimated consistently by

$$\Omega_{AD, n} = \frac{4}{n} \sum_{i=1}^n R_n(Y_i, X_i) R_n(Y_i, X_i)' - 4\delta_n \delta_n', \quad (2.44)$$

where

$$R_n(Y_i, X_i) = -\frac{1}{n-1} \frac{1}{h_n^{d+1}} \sum_{j \neq i} (Y_i - Y_j) K' \left(\frac{X_i - X_j}{h_n} \right).$$

2.6.2 An Improved Average-Derivative Estimator

The density-weighted average-derivative estimator of (2.42) requires the density of X to be increasingly smooth as the dimension of X increases. This is necessary to make $n^{1/2}(\hat{\delta}_n - \delta)$ asymptotically normal with a mean of 0. See assumption (a) of Theorem 2.3. The need for increasing smoothness is a form of the curse of dimensionality. Its practical consequence is that the finite-sample performance of the density-weighted average-derivative estimator is likely to deteriorate as the dimension of X increases, especially if the density of X is not very smooth. Specifically, the estimator's bias and mean-square error are likely to increase as the dimension of X increases.

Hristache et al. (2001) proposed an iterated average-derivative estimator that overcomes this problem. Their estimator is based on the observation that $G(x'\beta)$ does not vary when x varies in a direction that is perpendicular to β . Therefore, only the directional derivative of $E(Y|X = x)$ along the direction of β is needed for estimation. If this direction were known, then estimating the directional derivative would be a one-dimensional nonparametric estimation problem, and there would be no curse of dimensionality.

Of course, the direction of β is not known in applications, but Hristache et al. show that it can be estimated with sufficient accuracy through an iterative procedure. At each iteration, the gradient of $E(Y|X = x)$ is estimated using two bandwidths. The bandwidth in the estimated direction of β decreases as the iterations proceed, and the bandwidth in the estimated perpendicular direction increases. The use of two bandwidths enables the iterative procedure to mimic taking a directional derivative with increasing accuracy as the iterations proceed. The contribution to variance from estimation in the estimated perpendicular direction is small because the bandwidth in this direction is large. The contribution to bias is small despite the large bandwidth because $E(Y|X = x)$ varies little in the estimated perpendicular direction.

The details of the estimation procedure are as follows:

1. Specify the values of the tuning parameters ρ_1 , ρ_{\min} , a_ρ , a_h , h_1 , and h_{\max} . Methods for doing this are discussed below. Also, set $b_0 = 0$ and $j = 1$.
2. At iteration j ($j = 1, 2, \dots$), set $S_j = (I + \rho_j^{-2} b_{j-1} b_{j-1}')^{1/2}$, where b_{j-1} is the estimate of β at iteration $j-1$, I is the $k \times k$ identity matrix, and $k = \dim(\beta)$.
3. Let K be a kernel function. Denote the data by $\{Y_i, X_i: i = 1, \dots, n\}$. For each i , $\ell = 1, \dots, n$, define the column vector $X_{\ell i} = X_\ell - X_i$. Let $\hat{E}_j(X_i)$ and $\hat{\nabla} E_j(X_i)$ denote the estimates of $E(Y|X = X_i)$ and $\partial E(Y|X = X_i)/\partial x$ at the j th iteration. For each $i = 1, \dots, n$, these are obtained from the formula

$$\begin{bmatrix} \hat{E}_j(X_i) \\ \hat{\nabla} E_j(X_i) \end{bmatrix} = \left[\sum_{\ell=1}^n \begin{pmatrix} 1 \\ X_{\ell i} \end{pmatrix} \begin{pmatrix} 1 \\ X_{\ell i} \end{pmatrix}' K \left(\frac{\|S_j X_{\ell i}\|^2}{h_j^2} \right) \right]^{-1} \sum_{\ell=1}^n Y_\ell \begin{pmatrix} 1 \\ X_{\ell i} \end{pmatrix} K \left(\frac{\|S_j X_{\ell i}\|^2}{h_j^2} \right).$$

4. Compute the vector $b_j = n^{-1} \sum_{i=1}^n \hat{\nabla} E_j(X_i)$.
5. Set $h_{j+1} = a_h h_j$ and $\rho_{j+1} = a_\rho \rho_j$. If $\rho_{j+1} > \rho_{\min}$, set $j = j + 1$ and return to Step 2. Terminate if $\rho_{j+1} \leq \rho_{\min}$.

Let $j(n)$ denote the total number of iterations. The average-derivative estimate of β is $b_{j(n)}$. This estimate does not satisfy the scale normalization that requires its first component to equal 1, but that normalization can be achieved by division. Alternatively, $b_{j(n)}$ can be normalized to have unit length. This is the normalization that Hristache et al. use. It gives the estimate $\hat{\theta} = b_{j(n)} / \|b_{j(n)}\|$. In Step 3, $\hat{E}_j(X_i)$ and $\hat{\nabla} E_j(X_i)$ are local-linear estimates of $E(Y|X = X_i)$ and its gradient. Local-linear estimation is discussed in the Appendix. In particular, $\hat{E}_j(X_i)$ and $\hat{\nabla} E_j(X_i)$ solve the problem

$$\begin{bmatrix} \hat{E}_j(X_i) \\ \hat{\nabla} E_j(X_i) \end{bmatrix} = \arg \min_{c \in \mathbb{R}, b \in \mathbb{R}^k} \sum_{\ell=1}^n (Y_\ell - c - b' X_{\ell i})^2 K \left(\frac{\|S_j X_{\ell i}\|^2}{h_j^2} \right).$$

Hristache et al. proposed the following choices of tuning parameters. These choices are based on heuristic considerations and simulation evidence:

$$\begin{aligned} \rho_1 &= 1, & \rho_{\min} &= n^{-1/3}/h_{\max}, & a_\rho &= e^{-1/6}, \\ h_1 &= n^{-1/(4 \vee d)}, & h_{\max} &= 2d^{1/2}, & a_h &= e^{1/[2(4 \vee d)]}. \end{aligned}$$

We now state the asymptotic properties of the estimator. Make the following assumptions.

HJS1: The kernel, K , is a continuously differentiable, decreasing function on \mathbb{R}_+ with $K(0) = 1$ and $K(v) = 0$ for all $v \geq 1$.

HJS2: The model is $Y_i = G(X_i' \beta) + U_i$, where the U_i are independently and identically normally distributed with mean 0 and finite variance σ^2 .

HJS3: The function G is twice differentiable with a bounded second derivative.

HJS4: The points $\{X_i; i = 1, \dots, n\}$ are independently and identically distributed with a continuous, strictly positive density on $[0,1]^d$.

We now have the following theorem.

Theorem 2.4: *Let assumptions HJS1–HJS4 hold. Define $z_n = (1 + 2 \log n + 2 \log \log n)^{1/2}$ and $\beta^* = n^{-1} \beta' \sum_{i=1}^n G'(X_i' \beta)$. Then for all sufficiently large n ,*

$$P \left[\left\| (\hat{\theta} - \theta) - \frac{\gamma}{n^{1/2}} \right\| > \frac{C z_n^2 n^{-2/3}}{\|\beta^*\|} \right] \leq \frac{3j(n)}{n},$$

where C is a constant and γ is a normally distributed random vector in \mathbb{R}^d with mean 0 and a bounded covariance matrix. ■

Hristache et al. actually assume a fixed design (the X_i s are nonrandom), but this requires a rather complicated “design regularity” condition. A random design satisfies this condition with a probability that approaches 1 exponentially as n increases. Normality of the U_i s is not essential. The results can be extended to heteroskedastic, nonnormal U_i s that satisfy $\sup_{1 \leq i \leq n} E[\exp(\lambda U_i)] \leq D$ for some positive constants λ and D . The requirement that $X_i \in [0,1]^d$ is not restrictive because it can always be satisfied by transforming the X_i s.

Theorem 2.4 states, among other things, that the iterated average-derivative estimator is $n^{-1/2}$ -consistent and asymptotically normally distributed with a mean of 0. In contrast to the density-weighted average-derivative estimator of Section 2.6.1, this happens whenever X has a continuous, positive density, regardless of the dimension of X . Increasing smoothness and higher-order kernels are not needed to accommodate high-dimensional X s. The covariance matrix of the asymptotic distribution of the iterated average-derivative estimator is not specified, but this is unimportant because the estimator can be made asymptotically efficient with covariance matrix Ω_{SI} by taking one step toward the minimum of a suitable version of the weighted nonlinear least-squares estimator of Section 2.5. See Section 2.6.4.

2.6.3 Direct Estimation with Discrete Covariates

Average-derivative methods cannot be used to estimate components of β that multiply discrete components of X . This is because derivatives of $E(Y|X = x)$ with respect to discrete components of X are not identified. This section explains how direct (noniterative) estimation can be carried out when some components of X are discrete.

To distinguish between continuous and discrete covariates, let X denote the continuously distributed covariates and Z denote the discrete ones. Rewrite (2.1) in the form

$$E(Y|X = x, Z = z) = G(x' \beta + z' \alpha), \quad (2.45)$$

where α is the vector of coefficients of the discrete covariates. As was discussed in Section 2.3, identification requires there to be at least one continuous covariate. There need not be any discrete covariates, but it is assumed in this section that there is at least one. Let $d_z \geq 1$ denote the number of discrete covariates and components of Z .

The problem of interest in this section is estimating α . The parameter β can be estimated by using the average-derivative estimators of Sections 2.6.1 and 2.6.2 as follows. Let $S_z \equiv \{z^{(i)}: i = 1, \dots, M\}$ be the points of support of Z . Define $\delta_n^{(i)}$ to be the average-derivative estimator of δ that is obtained by applying the methods of Section 2.6.1 or 2.6.2 to the observations for which $Z = z^{(i)}$. Let $\delta_{n1}^{(i)}$ be the first component of $\delta_n^{(i)}$. Let w_{ni} ($i = 1, \dots, M$) be a set of nonnegative (possibly data-dependent) weights that sum to one. The estimator of β is

$$b_n = \frac{\sum_{i=1}^M w_{ni} \delta_n^{(i)}}{\sum_{i=1}^M w_{ni} \delta_{n1}^{(i)}}. \quad (2.46)$$

One possible set of weights is $w_{ni} = n_i/n$, where n_i is the number of observations the sample for which $Z = z^{(i)}$. However, the results presented in this section hold with any set of nonnegative weights that sum to one.

To see how α can be estimated, assume for the moment that G in (2.45) is known. Let $p(\cdot|z)$ denote the probability density function of $X'\beta$ conditional on $Z = z$. Make the following assumption.

Assumption G: There are finite numbers v_0 , v_1 , c_0 , and c_1 such that $v_0 < v_1$, $c_0 < c_1$, and $G(v) = c_0$ or c_1 at only finitely many values of v . Moreover, for each $z \in S_z$,

- (a) $G(v + z'\alpha) < c_0$ if $v < v_0$,
- (b) $G(v + z'\alpha) > c_1$ if $v > v_1$,
- (c) $p(\cdot|z)$ is bounded away from 0 on an open interval containing $[v_0, v_1]$.

Parts (a) and (b) of Assumption G impose a form of weak monotonicity on G . G must be smaller than c_0 at sufficiently small values of its argument and larger than c_1 at sufficiently large values. G is unrestricted at intermediate values of its argument. Part (c) ensures that $G(v + z'\alpha)$ is identified on $v_0 \leq v \leq v_1$.

To see the implications of Assumption G for estimating α , define

$$\begin{aligned} J(z) = & \int_{v_0}^{v_1} \{c_0 I[G(v + z'\alpha) < c_0] + c_1 I[G(v + z'\alpha) > c_1] \\ & + G(v + z'\alpha) I[c_0 \leq G(v + z'\alpha) \leq c_1]\} dv. \end{aligned}$$

Define $v_a = \max\{v_0 + z'\alpha: z \in S_z\}$ and $v_b = \min\{v_1 + z'\alpha: z \in S_z\}$. Make the change of variables $v = u - z'\alpha$ in the integrals on the right-hand side of $J(z)$.

Observe that by Assumption G, $I[G(u) < c_0] = 0$ if $u > v_b$, $I[G(u) > c_1] = 0$ if $u < v_a$, and $I[c_0 \leq G(u) \leq c_1] = 0$ if $u < v_a$ or $u > v_b$. Therefore,

$$\begin{aligned}
 J(z) &= c_0 \int_{v_0+z'\alpha}^{v_a} I[G(u) < c_0] du + c_0 \int_{v_a}^{v_b} I[G(u) < c_0] du \\
 &\quad + \int_{v_a}^{v_b} G(u) I[c_0 \leq G(u) \leq c_1] du + c_1 \int_{v_a}^{v_b} I[G(u) > c_1] du \\
 &\quad + c_1 \int_{v_b}^{v_1+z'\alpha} I[G(u) > c_1] du \\
 &= c_0(v_a - v_0 - z'\alpha) + c_0 \int_{v_a}^{v_b} I[G(u) < c_0] du + \int_{v_a}^{v_b} G(u) I[c_0 \leq G(u) \leq c_1] du \\
 &\quad + c_1 \int_{v_a}^{v_b} I[G(u) > c_1] du + c_1(v_1 - v_b + z'\alpha).
 \end{aligned}$$

It follows that for $i = 2, \dots, M$

$$J[z^{(i)}] - J[z^{(1)}] = (c_1 - c_0)[z^{(i)} - z^{(1)}]'\alpha. \quad (2.47)$$

Since c_0 , c_1 , and the support of Z are known, (2.47) constitutes $M - 1$ linear equations in the d_z unknown components of α . These equations can be solved for α if a unique solution exists. To do this, define the $(M - 1) \times 1$ vector ΔJ by

$$\Delta J = \begin{bmatrix} J[z^{(2)}] - J[z^{(1)}] \\ \vdots \\ J[z^{(M)}] - J[z^{(1)}] \end{bmatrix}.$$

Also, define the $(M - 1) \times d_z$ matrix W by

$$W = \begin{bmatrix} z^{(2)} - z^{(1)} \\ \vdots \\ z^{(M)} - z^{(1)} \end{bmatrix}.$$

Then

$$W' \Delta J = (c_1 - c_0)^{-1} W' W \alpha.$$

Therefore, if $W'W$ is a nonsingular matrix,

$$\alpha = (c_1 - c_0)(W'W)^{-1} W' \Delta J. \quad (2.48)$$

Equation (2.48) forms the basis of the estimator of α . The estimator is obtained by replacing the unknown $G(v + z'\alpha)$ that enters ΔJ with a kernel estimator of the

nonparametric mean regression of Y on $X' b_n$ conditional on $Z = z$. The resulting estimator of $G(v + z'\alpha)$ is

$$G_{nz}(v) = \frac{1}{nh_{nz}p_{nz}(v)} \sum_{i=1}^n I(Z_i = z) Y_i K\left(\frac{v - V_{ni}}{h_{nz}}\right), \quad (2.49)$$

where h_{nz} is a bandwidth, K is a kernel function, $V_{ni} = X'_i b_n$, and

$$p_{nz}(v) = \frac{1}{nh_{nz}} \sum_{i=1}^n I(Z_i = z) K\left(\frac{v - V_{ni}}{h_{nz}}\right). \quad (2.50)$$

The estimator of α is then

$$a_n = (c_1 - c_0)^{-1} (W'W)^{-1} W' \Delta J_n, \quad (2.51)$$

where

$$\Delta J_n = \begin{bmatrix} J_n[z^{(2)}] - J_n[z^{(1)}] \\ \vdots \\ J_n[z^{(M)}] - J_n[z^{(1)}] \end{bmatrix}$$

and

$$\begin{aligned} J_n(z) &= \int_{v_0}^{v_1} \{c_0 I[G_{nz}(v) < c_0] + c_1 I[G_{nz}(v) > c_1] \\ &\quad + G_{nz}(v) I[c_0 \leq G_{nz}(v) \leq c_1]\} dv. \end{aligned}$$

Horowitz and Härdle (1996) give conditions under which a_n in (2.51) is a consistent estimator of α and $n^{1/2}(a_n - \alpha)$ is asymptotically normally distributed with mean 0. The formal statement of this result is given in Theorem 2.5. Define $V = X'\beta$, $V_i = X'_i \beta$, $v = x'\beta$, and $G_z(v) = G(v + z'\alpha)$. Let $p(v|z)$ be the probability density of V conditional on $Z = z$, let $p(z)$ be the probability that $Z = z$ ($z \in S_z$), let $p(v, z) = p(v|z)p(z)$, and let $p(v, \tilde{x}|z)$ be the joint density of (V, \tilde{X}) conditional on $Z = z$. Finally, define

$$\Gamma(z) = - \int_{v_0}^{v_1} G'_z(v) E(\tilde{X}|v, z) I[c_0 \leq G(v + z'\alpha) \leq c_1] dv.$$

Theorem 2.5: *Let the following conditions hold.*

- (a) S_z is a finite set. $E(\|\tilde{X}\|^2 | Z = z) < \infty$ and $E(|Y| \|\tilde{X}\|^2 | Z = z) < \infty$ for each $z \in S_z$. $E(|Y|^2 \|\tilde{X}\|^2 | V = v, Z = z)$, $E(|Y|^2 | V = v, Z = z)$, and $p(v, z)$ are bounded uniformly over $v \in [v_0 - \varepsilon, v_1 + \varepsilon]$ for some $\varepsilon > 0$ and

all $z \in S_z$. For each $z \in S_z$, $p(v, \tilde{x}|z)$ is everywhere three times continuously differentiable with respect to v and the third derivative is bounded uniformly. $\text{Var}(Y|V = v, Z = z) > 0$ for all $z \in S_z$ and almost every v .

(b) $W'W$ is nonsingular.

(c) $E(Y|X = x, Z = z) = G(x'\beta + z'\alpha)$. G is r times continuously differentiable for some $r \geq 4$. G and its first r derivatives are bounded on all bounded intervals.

(d) Assumption G holds.

(e) If $d = \dim(X) > 1$, there is a $(d-1) \times 1$ vector-valued function $\omega(y, x, z)$ satisfying $E[\omega(Y, X, Z)] = 0$,

$$n^{1/2}(b_n - \beta) = \frac{1}{n^{1/2}} \sum_{i=1}^n \omega(Y_i, X_i, Z_i) + o_p(1),$$

and

$$\frac{1}{n^{1/2}} \sum_{i=1}^n \omega(Y_i, X_i, Z_i) \xrightarrow{d} N(0, V_\omega)$$

for some finite matrix V_ω .

(f) K in (2.49) and (2.50) is a bounded, symmetrical, differentiable function that is nonzero only on $[-1, 1]$. K' is Lipschitz continuous. For each integer j between 0 and r ($r \geq 4$),

$$\int_{-1}^1 v^j K(v) dv = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{if } 1 \leq j \leq r-1 \end{cases}.$$

(g) As $n \rightarrow \infty$, $nh_n^{r+3} \rightarrow \infty$ and $nh_n^{2r} \rightarrow 0$, where h_n is the bandwidth in (2.49) and (2.50).

Then a_n is a consistent estimator of α . Moreover, $n^{1/2}(a_n - \alpha)$ is asymptotically distributed as $N(0, \Omega_\alpha)$, where Ω_α is the covariance matrix of the $d_z \times 1$ random vector Λ_n whose m th component is

$$\begin{aligned} & \sum_{j=2}^m [(W'W)^{-1}W']_{mj} n^{-1/2} \sum_{i=1}^n \{I(Z_i = z^{(j)})p(V_i, z^{(j)})^{-1} \\ & [Y_i - G_{z^{(j)}}(V_i)]I[c_0 \leq G_{z^{(j)}}(V_i) \leq c_1] - I(Z_i = z^{(1)})p(V_i, z^{(1)})^{-1} \\ & [Y_i - G_{z^{(1)}}(V_i)]I[c_0 \leq G_{z^{(1)}}(V_i) \leq c_1] + (\Gamma_{z^{(j)}} - \Gamma_{z^{(1)}})\omega(Y_i, X_i, Z_i)\}. \blacksquare \end{aligned}$$

Condition (a) makes Z a discrete random variable with finite support and establishes the existence and properties of certain moments. The need for conditions (b) and (d) has already been discussed. Condition (c) makes $E(Y|X = x, Z = z)$ a single-index model. Condition (e) is satisfied by the estimators of β discussed in Sections 2.6.1 and 2.6.2 but does not require the use of these estimators. Conditions (f) and (g) require K to be a higher-order kernel with undersmoothing. As in

Section 2.6.1, conditions (f) and (g) are used to insure that the asymptotic distribution of $n^{1/2}(a_n - \alpha)$ is centered at 0.

The covariance matrix Ω_α can be estimated consistently by replacing unknown quantities with consistent estimators. Γ_z is estimated consistently by

$$\Gamma_{nz} = -\frac{1}{n} \sum_{i=1}^n \tilde{X}_i I(Z_i = z) I(v_0 \leq V_{ni} \leq v_1) I[c_0 \leq G_{nz}(V_{ni}) \leq c_1] G'_{nz}(V_{ni}) / p_{nz}(V_{ni}),$$

where $G'_{nz}(v) = dG_{nz}(v)/dv$. Define $\lambda(y, v, z)$ to be the $(M - 1) \times 1$ vector whose $(j - 1)$ component ($j = 2, \dots, M$) is

$$\begin{aligned} \lambda_j(y, v, z) = & I(z = z^{(j)}) \frac{y - G_{nz^{(j)}}(v)}{p_{nz^{(j)}}(v)} I[c_0 \leq G_{nz^{(j)}}(v) \leq c_1] \\ & - I(z = z^{(1)}) \frac{y - G_{nz^{(1)}}(v)}{p_{nz^{(1)}}(v)} I[c_0 \leq G_{nz^{(1)}}(v) \leq c_1]. \end{aligned}$$

Let ω_n be a consistent estimator of ω . Then Ω_α is estimated consistently by the sample covariance of the $d_z \times 1$ vector whose m th component ($m = 1, \dots, d_z$) is

$$\sum_{j=2}^m [(W'W)^{-1}W']_{mj} [\lambda_j(Y_i, V_{ni}, Z_i) + (\Gamma_{nz^{(j)}} - \Gamma_{nz^{(1)}})\omega_n(Y_i, X_i, Z_i)].$$

Horowitz and Härdle (1996) show how to estimate ω when the estimator of β is (2.46) and the $\delta_n^{(i)}$ are density-weighted average-derivative estimates (Section 2.6.1). To state their result, let $p_{ni}(x)$ be a kernel estimator of the probability density of X conditional on $Z = z^{(i)}$. That is,

$$p_{ni}(x) = \frac{1}{n_i s_n} \sum_{j=1}^n I(Z_j = z^{(i)}) K^* \left(\frac{x - X_j}{s_n} \right),$$

where K^* is a kernel function of a k -dimensional argument, n_i is the number of observations for which $Z = z^{(i)}$, and s_n is a bandwidth. Let $x^{(1)}$ be the first component of x . Then the estimator of ω is

$$\omega_n(y, x, z^{(i)}) = -2 \frac{n_i}{n \delta_n^{(i)}} [y - G(x' b_n + z^{(i)'} a_n)] \left[\frac{\partial p_{ni}(x)}{\partial \tilde{x}} - \tilde{b}_n \frac{\partial p_{ni}(x)}{\partial x^{(1)}} \right].$$

2.6.4 One-Step Asymptotically Efficient Estimators

In parametric estimation, an asymptotically efficient estimator can be obtained by taking one Newton step from any $n^{-1/2}$ -consistent estimator toward the maximum-likelihood estimator. This procedure is called *one-step asymptotically*

efficient estimation. The resulting estimator is called a *one-step asymptotically efficient estimator*. This section shows that the same idea applies to estimation of β in a semiparametric single-index model. Specifically, let S_n be the objective function of the semiparametric WNLS estimator (2.25) with $W(x) = 1/s_n^2(x)$. Then an asymptotically efficient estimator of β can be obtained by taking one Newton step from any $n^{1/2}$ -consistent estimator toward the minimum of S_n . In the case of a single-index binary-response model, the step may be taken toward the maximum of the semiparametric log-likelihood function (2.34).

One-step asymptotically efficient estimation is especially useful in semiparametric single-index models because the direct estimators described in Sections 2.6.1–2.6.3 can be computed very rapidly. Therefore, one-step estimators can be obtained with much less computation than is needed to minimize S_n or maximize the semiparametric log-likelihood function.

Consider one-step asymptotically efficient estimation based on S_n . Let X denote the entire vector of covariates, continuous and discrete. Let β denote the entire vector of coefficients of X in (2.1). Let \tilde{b}_n^* be any $n^{-1/2}$ -consistent estimator of $\tilde{\beta}$. It is convenient in applications but not essential to the arguments made here to let \tilde{b}_n^* be a direct estimator. The one-step asymptotically efficient estimator of $\tilde{\beta}$ is

$$\tilde{b}_n = \tilde{b}_n^* - \left[\frac{\partial^2 S_n(\tilde{b}_n^*)}{\partial \tilde{b} \partial \tilde{b}'} \right]^{-1} \frac{\partial S_n(\tilde{b}_n^*)}{\partial \tilde{b}}. \quad (2.52)$$

To see why \tilde{b}_n is asymptotically efficient, write (2.52) in the form

$$n^{1/2}(\tilde{b}_n - \tilde{\beta}) = n^{1/2}(\tilde{b}_n^* - \tilde{\beta}) - \left[\frac{\partial^2 S_n(\tilde{b}_n^*)}{\partial \tilde{b} \partial \tilde{b}'} \right]^{-1} n^{1/2} \frac{\partial S_n(\tilde{b}_n^*)}{\partial \tilde{b}}. \quad (2.53)$$

Observe that just as in the arguments leading to (2.31),

$$\frac{\partial^2 S_n(\tilde{b}_n^*)}{\partial \tilde{b} \partial \tilde{b}'} = C + o_p(1). \quad (2.54)$$

Moreover, a Taylor-series expansion gives

$$\frac{\partial S_n(\tilde{b}_n^*)}{\partial \tilde{b}} = \frac{\partial S_n(\tilde{\beta})}{\partial \tilde{b}} + \frac{\partial^2 S_n(\tilde{b}_n)}{\partial \tilde{b} \partial \tilde{b}'} (\tilde{b}_n^* - \tilde{\beta}),$$

where \tilde{b}_n is between \tilde{b}_n^* and $\tilde{\beta}$. The second-derivative term in this equation converges in probability to C , so

$$n^{1/2} \frac{\partial S_n(\tilde{b}_n^*)}{\partial \tilde{b}} = n^{1/2} \frac{\partial S_n(\tilde{\beta})}{\partial \tilde{b}} + C n^{1/2} (\tilde{b}_n^* - \tilde{\beta}) + o_p(1). \quad (2.55)$$

Substitution of (2.54) and (2.55) into (2.53) yields

$$n^{1/2}(\tilde{b}_n - \tilde{\beta}) = -C^{-1}n^{1/2}\frac{\partial S_n(\tilde{\beta})}{\partial \tilde{b}} + o_p(1).$$

As in (2.30)

$$n^{1/2}\frac{\partial S_n(\tilde{\beta})}{\partial \tilde{b}} \xrightarrow{d} N(0, D).$$

Therefore, $n^{1/2}(\tilde{b}_n - \tilde{\beta}) \xrightarrow{d} N(0, C^{-1}DC^{-1})$. Since $C^{-1}DC^{-1} = \Omega_{SI}$ when $W(x) = 1/s_n^2(x)$,

$$n^{1/2}(\tilde{b}_n - \tilde{\beta}) \xrightarrow{d} N(0, \Omega_{SI}).$$

This establishes the asymptotic efficiency of the one-step semiparametric WNLS estimator. The same arguments apply to the one-step semiparametric maximum-likelihood estimator after replacing S_n with the semiparametric log-likelihood function.

2.7 Bandwidth Selection

Implementation of any of the semiparametric estimators for single-index models that are discussed in this chapter requires choosing the numerical values of one or more bandwidth parameters and, possibly, of other tuning parameters. The selection of tuning parameters for the average-derivative estimator of Section 2.6.2 was discussed in that section. This section summarizes what is known about selecting tuning parameters for other estimators.

Härdle et al. (1993) investigated bandwidth selection for the semiparametric weighted nonlinear least-squares estimator of (2.25). They proposed optimizing the objective function over \tilde{b} and the bandwidth h_n . They gave conditions under which this yields an estimate of the bandwidth that minimizes the asymptotic integrated mean-square error of a kernel estimator of G . Thus, the resulting bandwidth estimate is an estimate of the asymptotically optimal bandwidth for kernel estimation of G . This bandwidth does not necessarily have any optimality properties for estimation of β .

As can be seen from the results in Sections 2.4, 2.5, and 2.6, in semiparametric single-index models, the asymptotic distribution of $n^{1/2}(b_n - \beta)$ does not depend on the bandwidth h_n . Therefore, bandwidth selection must be based on a higher-order approximation to the distribution of $n^{1/2}(b_n - \beta)$. Härdle and Tsybakov (1993) used such an approximation to obtain a formula for the bandwidth that minimizes the asymptotic approximation to $E \|\delta_n - \delta\|^2$, where δ and δ_n , respectively, are as in (2.39) and (2.42), and $\|\cdot\|$ is the Euclidean norm. This is an asymptotically optimal bandwidth for estimating β . Powell and Stoker(1996) obtained the

bandwidth that minimizes the asymptotic mean-square error of a single component of $\delta_n - \delta$.

Two aspects of the results of Härdle and Tsybakov (1993) and Powell and Stoker (1996) are especially noteworthy. First, the asymptotically optimal bandwidth has the form

$$h_{n,opt} = h_0 n^{-2/(2P+d+2)},$$

where P and d are defined as in Theorem 2.3 and h_0 is a constant. Second, Powell and Stoker (1996) provide a method for estimating h_0 in an application. To state this method, let h_{n1} be an initial bandwidth estimate that satisfies $h_{n1} \rightarrow 0$ and $nh_{n1}^c \rightarrow \infty$ as $n \rightarrow \infty$, where $c = \max(\eta + 2d + 4, P + d + 2)$ for some $\eta > 0$. Define

$$q_n(y_1, x_1, y_2, x_2) = -\frac{1}{h_{n1}^{d+1}}(y_1 - y_2)K'\left(\frac{x_1 - x_2}{h_{n1}}\right)$$

and

$$\hat{Q} = \frac{2h_{n1}^{d+2}}{n(n-1)} \sum_{i < j} q_n(Y_i, X_i, Y_j, X_j)^2.$$

Let $\delta_n(h)$ denote the density-weighted average-derivative estimator of δ based on bandwidth h . Let $\tau \neq 1$ be a positive number. Define

$$\hat{S} = \frac{\delta_n(\tau h_{n1}) - \delta_n(h_{n1})}{(\tau h_{n1})^P - h_{n1}^P}.$$

The estimator of h_0 is

$$\hat{h}_0 = \left[\frac{(d+2)\hat{Q}}{P\hat{S}^2} \right]^{1/(2P+d+2)}.$$

Another possible approach to bandwidth selection is based on resampling the data. Suppose that the asymptotically optimal bandwidth has the form

$$h_{n,opt} = h_0 n^{-\gamma}$$

for some known γ . For example, in density-weighted average-derivative estimation, $\gamma = 2P + d + 2$. Let $m < n$ be a positive integer. Let $\{Y_i^*, X_i^*: i = 1, \dots, m\}$ be a sample of size m that is obtained by sampling the estimation data randomly without replacement. Then $\{Y_i^*, X_i^*\}$ is a random sample from the population distribution of (Y, X) . Repeat this resampling process J times. Let $b_{mj}(h)$ ($j = 1, \dots, J$) be the estimate of β that is obtained from the j th sample using bandwidth $h = \tau m^{-\gamma}$, where τ is a constant. Let b_n be the estimate of β that is obtained from the full

sample by using a preliminary bandwidth estimate that satisfies the requirements needed to make b_n a $n^{-1/2}$ -consistent estimator of β . Let τ_m be the solution to the problem

$$\text{minimize: } \frac{1}{J} \sum_{j=1}^J [b_{mj}(h) - b_n]^2.$$

Then τ_m estimates h_0 , and $h_{n,opt}$ is estimated by

$$\hat{h}_{n,opt} = \tau_m n^{-\gamma}.$$

Horowitz and Härdle (1996) used Monte Carlo methods to obtain rules of thumb for selecting the tuning parameters required for the estimator of α described in Section 2.6.3. Horowitz and Härdle (1996) obtained good numerical results in Monte Carlo experiments by setting $h_{nz} = s_{vz} n_z^{-1/7.5}$, where s_{vz} is the sample standard deviation of $X'b_n$ conditional on $Z = z \in S_z$ and n_z is the number of observations with $Z = z$. In these experiments, the values of the other tuning parameters were

$$v_1 = \min_{z \in S_z} \max_{1 \leq i \leq n} \{X'_i b_n - h_{nz}: Z_i = z\},$$

$$v_0 = \max_{z \in S_z} \min_{1 \leq i \leq n} \{X'_i b_n + h_{nz}: Z_i = z\},$$

$$c_0 = \max_{z \in S_z} \max_{X_i b_n \leq v_0} G_{nz}^*(X'_i b_n),$$

and

$$c_1 = \min_{z \in S_z} \min_{X_i b_n \geq v_1} G_{nz}^*(X'_i b_n).$$

In the formulae for c_0 and c_1 , G_{nz}^* is the kernel estimator of G_z that is obtained using a second-order kernel instead of the higher-order kernel used to estimate α . Horowitz and Härdle (1996) found that using a second-order kernel produced estimates of c_0 and c_1 that were more stable than those obtained with a higher-order kernel.

2.8 An Empirical Example

This section presents an empirical example that illustrates the usefulness of semi-parametric single-index models. The example is taken from Horowitz and Härdle (1996) and 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 discrete 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 2.3 shows the parameter estimates obtained using a binary probit model and the direct semiparametric methods of Sections 2.6.1 and 2.6.3. Figure 2.1 shows a kernel estimate of $G'(v)$. There are two important differences between the semiparametric and probit estimates. First, the semiparametric estimate of β_{EMPLF} is small and statistically nonsignificant, whereas the probit estimate is significant at the 0.05 level and similar in size to β_{CAP} . Second, in the binary probit model, G is a cumulative normal distribution function, so G' is a normal density function. Figure 2.1 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.

Table 2.3 Estimated coefficients (standard errors) for model of product innovation

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

Source: Horowitz and Härdle (1996). The coefficient of EMPLP is 1 by scale normalization.

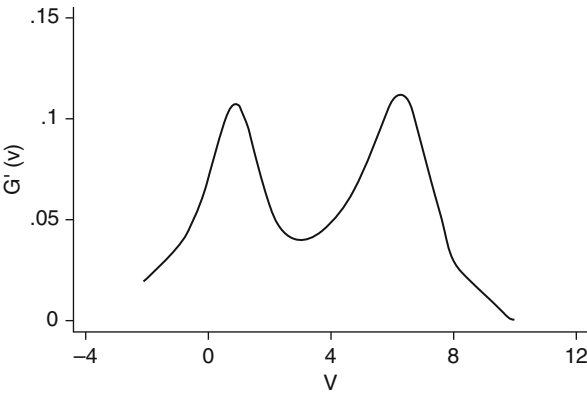


Fig. 2.1 Plot of $G'(v)$ for model of product innovation. Source: Horowitz and Härdle (1996)

2.9 Single-Index Models of Conditional Quantile Functions

Let $Q_\alpha(Y|X)$ denote the α -quantile of Y conditional on X , where $0 < \alpha < 1$. Then $P[Y \leq Q_\alpha(Y|X)|X] = \alpha$. In a single-index model of the conditional quantile function,

$$Q_\alpha(Y|X = x) = G(x'\beta), \quad (2.56)$$

where β is an unknown constant vector and G is an unknown function. It is not difficult to show that G and β are identified under the assumptions of Theorem 2.1. Moreover, if b_n is a $n^{-1/2}$ -consistent estimator of β , then G can be estimated with a one-dimensional nonparametric rate of convergence by carrying out a nonparametric quantile regression of Y on $X'b_n$. Nonparametric quantile regression is discussed briefly in the Appendix and in more detail by Chaudhuri (1991a,b), Fan et al. (1994), and Yu and Jones (1998). This section concentrates on $n^{-1/2}$ -consistent estimation of β . As is explained in the Appendix, estimating a conditional quantile function requires optimizing a nonsmooth objective function. Consequently, quantile estimation is more complex technically than estimation of conditional mean functions, and it requires regularity conditions that are more elaborate and difficult to interpret intuitively.

As with single-index models of conditional mean functions, β in (2.56) is proportional to $\partial Q_\alpha(Y|X = x)/\partial x$. Let W be a weight function. Define

$$\delta = E \left[\frac{\partial Q_\alpha(Y|X = x)}{\partial x} W(x) \right]. \quad (2.57)$$

Then δ and β are equal up to a proportionality constant. Replacing $\partial Q_\alpha(Y|X = x)/\partial x$ with a nonparametric estimator and the population expectation with a sample average in (2.57) yields an average-derivative estimator of δ and, hence, β up to a proportionality constant. Specifically, let the data $\{Y_i, X_i : i = 1, \dots, n\}$ be a simple random sample of (Y, X) , and let $\partial \hat{Q}_\alpha(Y|X_i)/\partial x$ be a nonparametric estimator of $\partial Q_\alpha(Y|X = x)/\partial x|_{x=X_i}$. Then the average-derivative estimator is

$$\hat{\delta}_{AD} = \frac{1}{n} \sum_{i=1}^n \left[\frac{\partial \hat{Q}_\alpha(Y|X_i)}{\partial x} \right] W(X_i).$$

Chaudhuri et al. (1997) have derived the asymptotic distributional properties of $\hat{\delta}_{AD}$. Their result requires the following definition of smoothness of a function. The definition is somewhat technical but basically requires derivatives of the function to be continuous. Let V be an open, convex subset of \mathbb{R}^d , where $d = \dim(X)$. Let m be a function from \mathbb{R}^d to \mathbb{R} . Define m to have order of smoothness p on V and write

$m \in S_p(V)$ if $p = \ell + \gamma$ for some integer $\ell \geq 0$ and γ satisfies $0 < \gamma \leq 1$, all partial derivatives of m through order ℓ exist, and each order ℓ partial derivative satisfies

$$\left| D^\ell m(x_1) - D^\ell m(x_2) \right| \leq C \|x_1 - x_2\|^\gamma$$

for all $x_1, x_2 \in V$, where $D^\ell m$ denotes any order ℓ partial derivative of m and $C > 0$ is a finite constant.

Let p_X denote the probability density function of X and $p_{Y|X}$ denote the density of Y conditional on X . For sequences of numbers $\{c_n\}$ and $\{d_n\}$, let $c_n \asymp d_n$ mean that c_n/d_n is bounded away from 0 and ∞ as $n \rightarrow \infty$. Let $[p]$ denote the largest integer that is less than or equal to p . For data $\{Y_i, X_i : i = 1, \dots, n\}$ define $U_i = Y_i - Q_\alpha(Y|X_i)$. Let ∇ denote the gradient operator. Define $l(x) = \nabla \log [p_X(x)]$.

Now make the following assumptions.

QAD1: The probability density of X is positive on V and $p_X \in S_{p_1}(V)$, where $p_1 = 1 + \gamma$ for some $\gamma \in (0, 1]$.

QAD2: The weight function W is supported on a compact set with nonempty interior that is contained in V , and $W \in S_{p_1}(V)$.

QAD3: Define $U = Y - Q_\alpha(Y|X = x)$. Let $p_{U|X}(u|x)$ denote the probability density function of U at u conditional on $X = x$. Then $p_{U|X}(u|x)$ considered as a function of x belongs to $S_{p_1(V)}$ for all u in a neighborhood of 0. Moreover $p_{U|X}(u|x) > 0$ and $\partial p_{U|X}(u|x)/\partial u$ exists and is continuous for all u in a neighborhood of 0 and all $x \in V$.

QAD4: $Q_\alpha(Y|X = x) \in S_{p_4(V)}$, where $p_4 > 3 + 3d/2$.

QAD5: $\hat{Q}_\alpha(Y|X = x)$ is the local polynomial estimator of Chaudhuri (1991a,b) with a polynomial of degree $[p_4]$ and bandwidth h_n satisfying $h_n \asymp n^{-\kappa}$ with $\frac{1}{2(p_4-1)} < \kappa < \frac{1}{4+3d}$.

The next theorem states the result of Chaudhuri et al. (1997).

Theorem 2.6: *Let assumptions QAD1–QAD4 hold. Then as $n \rightarrow \infty$,*

$$\begin{aligned} \hat{\delta}_{AD} - \delta &= \frac{1}{n} \sum_{i=1}^n \left\{ W(X_i) \nabla Q_\alpha(Y|X_i) - [\alpha - I(U_i \leq 0)] \frac{\nabla W(X_i) + W(X_i) l(X_i)}{p_{Y|X}[Q_\alpha(Y|X_i)|X_i]} \right\} \\ &\quad - \beta + o_p(n^{-1/2}). \blacksquare \end{aligned}$$

Theorem 2.6 implies that $n^{1/2}(\hat{\delta}_{AD} - \delta)$ is asymptotically normally distributed with mean 0 and variance equal to the variance of the summand on the right-hand side of (2.58). As in average-derivative estimation of a conditional mean function (Section 2.6), averaging of the nonparametric estimator $\partial \hat{Q}_\alpha / \partial x$ in (2.58) enables $\hat{\delta}_{AD}$ to achieve a $n^{-1/2}$ rate of convergence instead of the slower rate for nonparametric estimation of derivatives. It follows from QAD5 that Q_α must be increasingly smooth as d increases. Thus, the average-derivative estimator of Chaudhuri et al. (1997), like the density-weighted average-derivative estimator of Powell et al.

(1989) for conditional mean functions, has a curse of dimensionality. Methods for choosing W and h_n in applications and for avoiding the curse of dimensionality in average-derivative estimation of conditional quantile functions have not yet been developed.

Khan (2001) has developed a rank estimator of β in (2.56) that is $n^{-1/2}$ -consistent and asymptotically normal if G is monotonic. The average-derivative estimator does not require monotonicity, but the rank estimator requires less smoothness than does the average-derivative estimator. In addition, the rank estimator can accommodate discrete components of X , although at least one component must be continuously distributed. Khan's estimator is based on an estimator of Cavanagh and Sherman (1998) and is

$$b_n = \arg \min_{\tilde{b} \in \tilde{B}, b_1=1} \frac{1}{n} \sum_{\substack{i,j=1 \\ i \neq j}}^n W(X_i) \hat{Q}_\alpha(Y|X_i) I(X'_i b > X'_j b),$$

where \tilde{b} denotes the vector consisting of all components of b except the first, $\tilde{B} \in \mathbb{R}^{d-1}$ is a compact parameter set, W is a weight function, and \hat{Q}_α is Chaudhuri's (1991a,b) nonparametric estimator of Q_α .

To obtain the asymptotic distribution of Khan's estimator, define

$$\begin{aligned} \tau_1(x, b) &= \int W(x) Q_\alpha(Y|X=x) I(x'b > v'b) p_X(v) dv \\ &\quad + \int W(v) Q_\alpha(Y|X=v) I(v'b > x'b) p_X(v) dv \end{aligned}$$

and

$$\tau_2(x, b) = \int I(x'b > v'b) p_X(v) dv.$$

Let $\tilde{\beta}$ be the vector consisting of all components of β except the first. Let N be a neighborhood of $\tilde{\beta}$. Now make the following assumptions.

RAD1: $\tilde{\beta}$ is in the interior of the compact parameter set \tilde{B} .

RAD2: $Q_\alpha(Y|X=x) = G(x'\beta)$ and G is a nonconstant, increasing function.

RAD3: $Q_\alpha(Y|X) \in S_p(V)$, where $p > 3d/2$ and V is the support of X .

RAD4: The weight function W is continuous, bounded, and bounded away from 0 on its support, S_W . S_W has the form $S_{W1} \times \tilde{S}_W$, where S_{W1} is a compact subset of the support of the first component of X and has a nonempty interior. \tilde{S}_W is a compact subset of the remaining $d-1$ components of X and has a nonempty interior. S_W is not contained in any proper linear subspace of \mathbb{R}^d .

RAD5: The support of X is a convex subset of \mathbb{R}^d with a nonempty interior.

RAD6: X has a probability density function, p_X , that is continuous and bounded on its support. Moreover, $p_X(x) \geq c$ for some constant $c > 0$ and all $x \in S_W$.

RAD7: Let t_0 satisfy $G(t) < G(t_0)$ if $t < t_0$. Assume that

$$T \equiv \max_{\tilde{x} \in \tilde{S}_W, \tilde{b} \in \tilde{B}} |\tilde{x}'\tilde{b}| < \infty.$$

Then $[t_0 - 3T, t_0 + 3T] \in S_{W1}$.

RAD8: Define $U = Y - Q_\alpha(Y|X = x)$. Let $p_{U|X}(u|x)$ denote the probability density function of U at u conditional on $X = x$. Then $p_{U|X}(u|x)$ considered as a function of x is Lipschitz continuous for all u in a neighborhood of 0. Moreover $p_{U|X}(u|x)$ considered as a function of u is continuous, bounded, and bounded away from 0 for all u in a neighborhood of 0.

RAD9: For each x in the support of X and all $\tilde{b} \in N$, $\nabla^2 \tau_1(x, b) \equiv \partial^2 \tau_1(x, b) / \partial \tilde{b} \partial \tilde{b}'$ exists and is Lipschitz continuous. Moreover, $E[\nabla^2 \tau_1(X, \beta)]$ is negative definite.

RAD10: For each x in the support of X and all $\tilde{b} \in N$, $\nabla \tau_2(x, b) \equiv \partial \tau_2(x, b) / \partial \tilde{b}$ exists and is continuous. Moreover, $E \|\nabla \tau_2(X, \beta)\| < \infty$.

RAD11: \hat{Q}_α is a local polynomial estimator based on a polynomial of degree $[p]$ and bandwidth h_n satisfying $n^{1/2}h_n^p \rightarrow 0$ and $(\log n)/(nh_n^{3d})^{1/2} \rightarrow 0$ as $n \rightarrow \infty$.

The following theorem shows that $n^{1/2}(\tilde{b}_n - \tilde{\beta})$ is asymptotically normal under RAD1–RAD11.

Theorem 2.7: Let RAD1–RAD11 hold. Then $n^{1/2}(\tilde{b}_n - \tilde{\beta}) \xrightarrow{d} N(0, D^{-1}\Sigma D^{-1})$, where $D = 0.5E[\nabla^2 \tau_1(X, \beta)]$, $\Sigma = E[s(Y, X)s(Y, X)']$, and

$$s(y, x) = \frac{W(x)}{p_{U|X}(0|x)} \{I[y \leq Q_\alpha(Y|X = x)] - \alpha\} \nabla \tau_2(x, \beta). \blacksquare$$

Khan (2001) proves Theorem 2.7, provides an estimator of the covariance matrix $D^{-1}\Sigma D^{-1}$, and shows that under slightly modified assumptions, the conclusion of the theorem holds if some components of X are discrete. Like the average-derivative estimator, the rank estimator requires fully nonparametric estimation of Q_α and has a curse of dimensionality, but the rank estimator's smoothness assumptions are weaker than the smoothness assumptions of the average-derivative estimator.



<http://www.springer.com/978-0-387-92869-2>

Semiparametric and Nonparametric Methods in
Econometrics

Horowitz, J.L.

2009, X, 276 p., Hardcover

ISBN: 978-0-387-92869-2