Chapter 41

ESTIMATION OF SEMIPARAMETRIC MODELS*

JAMES L. POWELL

Princeton University

Contents

*This work was supported by NSF Grants 91–96185 and 92–10101 to Princeton University. I am grateful to Hyungtaik Ahn, Moshe Buchinsky, Gary Chamberlain, Songnian Chen, Gregory Chow, Angus Deaton, Bo Honoré, Joel Horowitz, Oliver Linton, Robin Lumsdaine, Chuck Manski, Rosa Matzkin, Dan McFadden, Whitney Newey, Paul Ruud, and Tom Stoker for their helpful suggestions, which were generally adopted except when they were mutually contradictory or required a lot of extra work.

Handbook of Econometrics, Volume IV, Edited by R.F. Engle and D.L. McFadden

Abstract

A semiparametric model for observational data combines a parametric form for some component of the data generating process (usually the behavioral relation between the dependent and explanatory variables) with weak nonparametric restrictions on the remainder of the model (usually the distribution of the unobservable errors). This chapter surveys some of the recent literature on semiparametric methods, emphasizing microeconometric applications using limited dependent variable models. An introductory section defines semiparametric models more precisely and reviews the techniques used to derive the large-sample properties of the corresponding estimation methods. The next section describes a number of weak restrictions on error distributions – conditional mean, conditional quantile, conditional symmetry, independence, and index restrictions – and show how they can be used to derive identifying restrictions on the distributions of observables. This general discussion is followed by a survey of a number of specific estimators proposed for particular econometric models, and the chapter concludes with a brief account of applications of these methods in practice.

1. Introduction

1.1. Overview

Semiparametric modelling is, as its name suggests, a hybrid of the parametric and nonparametric approaches to construction, fitting, and validation of statistical models. To place semiparametric methods in context, it is useful to review the way these other approaches are used to address a generic microeconometric problem – namely, determination of the relationship of a dependent variable (or variables) y to a set of conditioning variables x given a random sample $\{z_i \equiv (y_i, x_i), i = 1, \ldots, N\}$ of observations on y and x. This would be considered a "micro"-econometric problem because the observations are mutually independent and the dimension of the conditioning variables x is finite and fixed. In a "macro"-econometric application using time series data, the analysis must also account for possible serial dependence in the observations, which is usually straightforward, and a growing or infinite number of conditioning variables, e.g. past values of the dependent variable y, which may be more difficult to accommodate. Even for microeconometric analyses of cross-sectional data, distributional heterogeneity and dependence due to clustering and stratification must often be considered; still, while the random sampling assumption may not be typical, it is a useful simplification, and adaptation of statistical methods to non-random sampling is usually straightforward.

In the classical parametric approach to this problem, it is typically assumed that the dependent variable is functionally dependent on the conditioning variables

("regressors") and unobservable "errors" according to a fixed structural relation of the form

$$y = g(x, \alpha_0, \varepsilon), \tag{1.1}$$

where the structural function $g(\cdot)$ is known but the finite-dimensional parameter vector $\alpha_0 \in \mathbb{R}^p$ and the error term ε are unobserved. The form of $g(\cdot)$ is chosen to give a class of simple and interpretable data generating mechanisms which embody the relevant restrictions imposed by the characteristics of the data (e.g. $g(\cdot)$ is dichotomous if y is binary) and/or economic theory (monotonicity, homotheticity, etc.). The error terms ε are introduced to account for the lack of perfect fit of (1.1) for any fixed value of α_0 and ε, and are variously interpreted as expectational or optimization errors, measurement errors, unobserved differences in tastes or technology, or other omitted or unobserved conditioning variables; their interpretation influences the way they are incorporated into the structural function $g(\cdot)$.

To prevent (1.1) from holding tautologically for any value of α_0, the stochastic behavior of the error terms must be restricted. The parametric approach takes the error distribution to belong to a finite-dimensional family of distributions,

$$\Pr\{\varepsilon \leqslant \lambda | x\} = \int_{-\infty}^{\lambda} f_\varepsilon(u | x, \eta_0)\, \mathrm{d}\mu_\varepsilon, \tag{1.2}$$

where $f(\cdot)$ is a known density (with respect to the dominating measure μ_ε) except for an unknown, finite-dimensional "nuisance" parameter η_0. Given the assumed structural model (1.1) and the conditional error distribution (1.2), the conditional distribution of y given x can be derived,

$$\Pr\{y \leqslant \lambda | x\} = \int_{-\infty}^{\lambda} 1\{y \leqslant \lambda\} f_{y|x}(u | x, \alpha_0, \eta_0)\, \mathrm{d}\mu_{y|x},$$

for some parametric conditional density $f_{y|x}(\cdot)$. Of course, it is usually possible to posit this conditional distribution of y given x directly, without recourse to unobservable "error" terms, but the adequacy of an assumed functional form is generally assessed with reference to an implicit structural model. In any case, with this conditional density, the unknown parameters α_0 and η_0 can be estimated by maximizing the average conditional log-likelihood

$$L_N(\alpha, \eta) \equiv \frac{1}{N} \sum_{i=1}^{N} \ln f_{y|x}(y_i | x_i, \alpha, \eta)$$

over α and η.

This fully parametric modelling strategy has a number of well-known optimality properties. If the specifications of the structural equation (1.1) and error distribution (1.2) are correct (and other mild regularity conditions hold), the maximum likelihood estimators of α_0 and η_0 will converge to the true parameters at the rate of the inverse square root of the sample size ("root-N-consistent") and will be

asymptotically normally distributed, with an asymptotic covariance matrix which is no larger than that of any other regular root-N-consistent estimator. Moreover, the parameter estimates yield a precise estimator of the conditional distribution of the dependent variable given the regressors, which might be used to predict y for values of x which fall outside the observed support of the regressors. The drawback to parametric modelling is the requirement that both the structural model and the error distribution are correctly specified. Correct specification may be particularly difficult for the error distribution, which represents the unpredictable component of the relation of y to x. Unfortunately, if $g(x, \alpha, \varepsilon)$ is fundamentally nonlinear in ε – that is, it is noninvertible in ε or has a Jacobian that depends on the unknown parameters α – then misspecification of the functional form of the error distribution $f(\varepsilon|x, \eta)$ generally yields inconsistency of the MLE and inconsistent estimates of the conditional distribution of y given x.

At the other extreme, a fully nonparametric approach to modelling the relation between y and x would define any such "relation" as a characteristic of the joint distribution of y and x, which would be the primitive object of interest. A "causal" or predictive relation from the regressors to the dependent variable would be given as a particular functional of the conditional distribution of y given x,

$$g(x) = T(F_{y|x}), \tag{1.3}$$

where $F_{y,x}$ is the joint and $F_{y|x}$ is the conditional distribution. Usually the functional $T(\cdot)$ is a location measure, in which case the relation between y and x has a representation analogous to (1.1) and (1.2), but with unknown functional forms for $f(\cdot)$ and $g(\cdot)$. For example, if $g(x)$ is the mean regression function ($T(F_{y|x}) = E[y|x]$), then y can be written as

$$y = g(x) + \varepsilon,$$

with ε defined to have conditional density $f_{\varepsilon|x}$ assumed to satisfy only the normalization $E[\varepsilon|x] = 0$. In this approach the interpretation of the error term ε is different than for the parametric approach; its stochastic properties derive from its definition in terms of the functional $g(\cdot)$ rather than a prior behavioral assumption.

Estimation of the function $g(\cdot)$ is straightforward once a suitable estimator $\hat{F}_{y|x}$ of the conditional distribution of y given x is obtained; if the functional $T(\cdot)$ in (1.3) is well-behaved (i.e. continuous over the space of possible $F_{y|x}$), a natural estimator is

$$\hat{g}(x) \equiv T(\hat{F}_{y|x}).$$

Thus the problem of estimating the "relationship" $g(\cdot)$ reduces to the problem of estimating the conditional distribution function, which generally requires some smoothing across adjacent observations of the regressors x when some components

are continuously distributed (see, e.g. Prakasa Rao (1983), Silverman (1986), Bierens (1987), Härdle (1991)). In some cases, the functional $T(\cdot)$ might be a well-defined functional of the empirical c.d.f. of the data (for example, $g(x)$ might be the best linear projection of y on x, which depends only on the covariance matrix of the data); in these cases smoothing of the empirical c.d.f. will not be required. An alternative estimation strategy would approximate $g(x)$ and the conditional distribution of ε in (1.6) by a sequence of parametric models, with the number of parameters expanding as the sample size increases; this approach, termed the "method of sieves" by Grenander (1981), is closely related to the "seminonparametric" modelling approach of Gallant (1981, 1987), Elbadawi et al. (1983) and Gallant and Nychka (1987).

The advantages and disadvantages of the nonparametric approach are the opposite of those for parametric modelling. Nonparametric modelling typically imposes few restrictions on the form of the joint distribution of the data (like smoothness or monotonicity), so there is little room for misspecification, and consistency of an estimator of $g(x)$ is established under much more general conditions than for parametric modelling. On the other hand, the precision of estimators which impose only nonparametric restrictions is often poor. When estimation of $g(x)$ requires smoothing of the empirical c.d.f. of the data, the convergence rate of the estimator is usually slower than the parametric rate (square root of the sample size), due to the bias caused by the smoothing (see the chapter by Härdle and Linton in this volume). And, although some prior economic restrictions like homotheticity and monotonicity can be incorporated into the nonparametric approach (as described in the chapter by Matzkin in this volume), the definition of the "relation" is statistical, not economic. Extrapolation of the relationship outside the observed support of the regressors is not generally possible with a nonparametric model, which is analogous to a "reduced form" in the classical terminology of simultaneous equations modelling.

The semiparametric approach, the subject of this chapter, distinguishes between the "parameters of interest", which are finite-dimensional, and infinite-dimensional "nuisance parameters", which are treated nonparametrically. (When the "parameter of interest" is infinite-dimensional, like the baseline hazard in a proportional hazards model, the nonparametric methods described in the Härdle and Linton chapter are more appropriate.) In a typical parametric model, the parameters of interest, α_0, appear only in a structural equation analogue to (1.1), while the conditional error distribution is treated as a nuisance parameter, subject to certain prior restrictions. More generally, unknown nuisance functions may also appear in the structural equation. Semiparametric analogues to equations (1.1) and (1.2) are

$$y = g(x, \alpha_0, \varepsilon, \tau_0(\cdot)), \tag{1.4}$$

$$\Pr\{\varepsilon \leqslant \lambda | x\} = \int 1\{u \leqslant \lambda\} f_0(u|x)\, d\mu_\varepsilon, \tag{1.5}$$

where, as before, α_0 is unknown but known to lie in a finite-dimensional Euclidean subspace, and where the unknown nuisance parameter is

$$\eta_0 = (\tau_0(\cdot), f_0(\cdot)).$$

As with the parametric approach, prior economic reasoning and interpretational convenience are used to determine the functional form of $g(\cdot)$ in (1.4), while general regularity and identification restrictions are imposed on the nuisance parameters η_0, as in the nonparametric approach.

As a hybrid of the parametric and nonparametric approaches, semiparametric modelling shares the advantages and disadvantages of each. Because it allows a more general specification of the nuisance parameters, estimators of the parameters of interest for semiparametric models are consistent under a broader range of conditions than for parametric models, and these estimators are usually more precise (converging to the true values at the square root of the sample size) than their nonparametric counterparts. On the other hand, estimators for semiparametric models are generally less efficient than maximum likelihood estimators for a correctly-specified parametric model, and are still sensitive to misspecification of the structural function or other parametric components of the model.

This chapter will survey the econometric literature on semiparametric estimation, with emphasis on a particular class of models, nonlinear latent variable models, which have been the focus of most of the attention in this literature. The remainder of Section 1 more precisely defines the "semiparametric" categorization, briefly lists the structural functions and error distributions to be considered and reviews the techniques for obtaining large-sample approximations to the distributions of various types of estimators for semiparametric models. The next section discusses how each of the semiparametric restrictions on the behavior of the error terms can be used to construct estimators for certain classes of structural functions. Section 3 then surveys existing results in the econometric literature for several groups of latent variable models, with a variety of error restrictions for each group of structural models. A concluding section summarizes this literature and suggests topics for further work.

The coverage of the large literature on semiparametric estimation in this chapter will necessarily be incomplete; fortunately, other general references on the subject are available. A forthcoming monograph by Bickel et al. (1993) discusses much of the work on semiparametrics in the statistical literature, with special attention to construction of efficient estimators; a monograph by Manski (1988b) discusses the analogous econometric literature. Other surveys of the econometric literature include those by Robinson (1988a) and Stoker (1992), the latter giving an extensive treatment of estimation based upon index restrictions, as described in Section 2.5 below. Newey (1990a) surveys the econometric literature on semiparametric efficiency bounds, which is not covered extensively in this chapter. Finally, given the close connection between the semiparametric approach and parametric and

nonparametric approaches, the chapters by Andrews, Härdle and Linton, Manski, Matzkin, and Newey and McFadden in this volume provide more details on much of the material in the present chapter.

1.2. *Definition of "semiparametric"*

The characterization of semiparametric models as having a finite-dimensional parameter of interest (the "parametric component") and an infinite-dimensional nuisance parameter (the "nonparametric component") was given by Begun et al. (1983), who attribute the term to Oakes (1981). Although this distinction is a defining characteristic of semiparametric modelling, alone it appears to be too inclusive: many problems which would traditionally be viewed as "nonparametric" or "parametric" might well be classified as "semiparametric" along these lines. For example, the best linear predictor of y given x lies in a finite-dimensional space (indexed by the vector of projection coefficients), but this object is more closely analogous to the conditional mean of y given x (a "nonparametric" relation) than to a traditional structural relation of the form given in (1.1). The example suggests that the "dimensionality" of unknown components of a model is not sufficient to characterize it as nonparametric or semiparametric; instead, this distinction must depend somehow on the "size" of the space of nuisance parameters for the model – that is, on the generality of the restrictions imposed on η_0. At the other extreme, for the typical parametric model with ancillary regressors, the marginal distribution of the regressors might be viewed as an infinite-dimensional nuisance parameter, blurring the line between "parametric" and "semiparametric" modelling.

A refinement of the definition of semiparametric (versus nonparametric) modelling might exploit the distinction between "just-" and "over-identification" introduced in the simultaneous equations literature. In a nonparametric model, the parameters of interest can be said to be "just-identified", in that they are defined by a unique functional of the joint distribution of the data. That is, if $\alpha_0 \equiv T(F_{y,x})$ defines the parameter of interest as a characteristic of the joint distribution of y and x, then a model might be defined to be nonparametric if the functional T is unique whenever it is well-defined. In contrast, a semiparametric model would restrict the space of permissible joint distribution functions so that more than one functional would yield the same value of the parameter of interest: $\alpha_0 \equiv T^+(F_{y,x})$, where $T(G_0) \neq T^+(G_0)$ for some possible distribution function G_0 of y and x for which either side is well-defined. For example, in a nonparametric model α_0 could be the mean of the dependent variable y, whose marginal distribution is otherwise unrestricted, while a semiparametric model might restrict the distribution of y to be symmetric about the constant α_0, which could then be recovered as the mean, median, or any number of possible location measures for F_y. In a nonparametric setting, the only scope for differences in estimators of α_0 in a nonparametric model would be through differences in estimates of the distribution function $F_{y,x}$ of the data (due,

say, to different methods and degrees of "smoothing" of the empirical c.d.f.), while estimation of a semiparametric model would require an additional choice of the particular functional T^* upon which to base the estimates.

On a related point, while it is common to refer to "semiparametric estimation" and "semiparametric estimators", this is somewhat misleading terminology. Some authors use the term "semiparametric estimator" to denote a statistic which involves a preliminary "plug-in" estimator of a nonparametric component (see, for example, Andrews' chapter in this volume); this leads to some semantic ambiguities, since the parameters of many semiparametric models can be estimated by "parametric" estimators and vice versa. Thus, though certain estimators would be hard to interpret in a parametric or nonparametric context, in general the term "semiparametric", like "parametric" or "nonparametric", will be used in this chapter to refer to classes of structural models and stochastic restrictions, and not to a particular statistic. In many cases, the same estimator can be viewed as parametric, nonparametric or semiparametric, depending on the assumptions of the model. For example, for the classical linear model

$$y = x'\beta_0 + \varepsilon,$$

the least squares estimator of the unknown coefficients β_0,

$$\hat{\beta} = \left[\sum_{i=1}^{N} x_i x_i'\right]^{-1} \sum_{i=1}^{N} x_i y_i,$$

would be considered a "parametric" estimator when the error terms are assumed to be Gaussian with zero mean and distributed independently of the regressors x. With these assumptions $\hat{\beta}$ is the maximum likelihood estimator of β_0, and thus is asymptotically efficient relative to all regular estimators of β_0. Alternatively, the least squares estimator arises in the context of a linear prediction problem, where the error term ε has a density which is assumed to satisfy the unconditional moment restriction

$$E[\varepsilon \cdot x] = 0.$$

This restriction yields a unique representation for β_0 in terms of the joint distribution of the data,

$$\beta_0 = \{E[x \cdot x']\}^{-1} E[x \cdot y],$$

so estimation of β_0 in this context would be considered a "nonparametric" problem by the criteria given above. Though other, less precise estimators of the moments $E[x \cdot x']$ and $E[x \cdot y]$ (say, based only on a subset of the observations) might be used to define alternative estimators, the classical least squares estimator $\hat{\beta}$ is, al-

most by default, an "efficient" estimator of β_0 in this model (as Levit (1975) makes precise). Finally, the least squares estimator $\hat{\beta}$ can be viewed as a special case of the broader class of weighted least squares estimators of β_0 when the error terms ε are assumed to have conditional mean zero,

$$E[\varepsilon_i | x_i] = 0 \quad \text{a.s.}$$

The model defined by this restriction would be considered "semiparametric", since β_0 is overidentified; while the least squares estimator $\hat{\beta}$ is $\sqrt{N}$-consistent and asymptotically normal for this model (assuming the relevant second moments are finite), it is inefficient in general, with an efficient estimator being based on the representation

$$\beta_0 = T^*(F_{y,x}) = \{E[\sigma^{-2}(x_i)x_i x_i']^{-1}\}E[\sigma^{-2}(x_i)x_i y_i]$$

of the parameters of interest, where $\sigma^2(x) \equiv \text{Var}(\varepsilon_i | x_i)$ (as discussed in Section 2.1 below). The least squares statistic $\hat{\beta}$ is a "semiparametric" estimator in this context, due to the restrictions imposed on the model, not on the form of the estimator.

Two categories of estimators which are related to "semiparametric estimators", but logically distinct, are "robust" and "adaptive" estimators. The term "robustness" is used informally to denote statistical procedures which are well-behaved for slight misspecifications of the model. More formally, a robust estimator $\hat{\alpha} \equiv T(\hat{F}_{y,x})$ can be defined as one for which $T(F)$ is a continuous functional at the true model (e.g. Manski (1988b)), or whose asymptotic distribution is continuous at the truth ("quantitative robustness", as defined by Huber (1981)). Other notions of robustness involve sensitivity of particular estimators to changes in a small fraction of the observations. While "semiparametric estimators" are designed to be well-behaved under weak conditions on the error distribution and other nuisance parameters (which are assumed to be correct), robust estimators are designed to be relatively efficient for correctly-specified models but also relatively insensitive to "slight" model misspecification. As noted in Section 1.4 below, robustness of an estimator is related to the boundedness (and continuity) of its influence function, defined in Section 1.4 below; whether a particular semiparametric model admits a robust estimator depends upon the particular restrictions imposed. For example, for conditional mean restrictions described in Section 2.1 below, the influence functions for semiparametric estimators will be linear (and thus unbounded) functions of the error terms, so robust estimation is infeasible under this restriction. On the other hand, the influence function for estimators under conditional quantile restrictions depends upon the sign of the error terms, so quantile estimators are generally "robust" (at least with respect to outlying errors) as well as "semiparametric".

"Adaptive" estimators are efficient estimators of certain semiparametric models for which the best attainable efficiency for estimation of the parameters of interest

does not depend upon prior knowledge of a parametric form for the nuisance parameters. That is, adaptive estimators are consistent under the semiparametric restrictions but as efficient (asymptotically) as a maximum likelihood estimator when the (infinite-dimensional) nuisance parameter is known to lie in a finite-dimensional parametric family. Adaptive estimation is possible only if the semiparametric information bound for attainable efficiency for the parameters of interest is equal to the analogous Cramér–Rao bound for any feasible parametric specification of the nuisance parameter. Adaptive estimators, which are described in more detail by Bickel et al. (1993) and Manski (1988b), involve explicit estimation of (nonparametric) nuisance parameters, as do efficient estimators for semiparametric models more generally.

1.3. *Stochastic restrictions and structural models*

As discussed above, a semiparametric model for the relationship between y and x will be determined by the parametric form of the structural function $g(\cdot)$ of (1.4) and the restrictions imposed on the error distribution and any other infinite-dimensional component of the model. The following sections of this chapter group semiparametric models by the restrictions imposed on the error distribution, describing estimation under these restrictions for a number of different structural models. A brief description of the restrictions to be considered, followed by a discussion of the structural models, is given in this section.

A semiparametric restriction on ε which is quite familiar in econometric theory and practice is a (constant) *conditional mean* restriction, where it is assumed that

$$E(\varepsilon|x) = \mu_0 \tag{1.6}$$

for some unknown constant μ_0, which is usually normalized to zero to ensure identification of an intercept term. (Here and throughout, all conditional expectations are assumed to hold for a set of regressors x with probability one.) This restriction is the basis for much of the large-sample theory for least squares and method-of-moments estimation, and estimators derived for assumed Gaussian distributions of ε (or, more generally, for error distributions in an exponential family) are often well-behaved under this weaker restriction.

A restriction which is less familiar but gaining increasing attention in econometric practice is a (constant) *conditional quantile* restriction, under which a scalar error term ε is assumed to satisfy

$$\Pr\{\varepsilon \leqslant \eta_0|x\} = \pi \tag{1.7}$$

for some fixed proportion $\pi \in (0, 1)$ and constant $\eta_0 = \eta_0(\pi)$; a *conditional median* restriction is the (leading) special case with $\pi = 1/2$. Rewriting the conditional

probability in (1.7) as the conditional expectation of an indicator function, the quantile restriction can be expressed as $E[\pi - 1\{\varepsilon \leqslant \eta_0\}|x] = 0$, which specializes to $E[\text{sgn}\{\varepsilon - \eta_0\}|x] = 0$ for a conditional median restriction. As discussed in Section 2.2 below, conditional quantile restrictions are useful for identifying the parameters of interest for structural models which are monotonic in the error term.

For scalar error terms, both conditional mean and conditional quantile restrictions are themselves special cases of a *constant conditional location* restriction, in which, for some constant v_0, the error terms satisfy

$$E[q(\varepsilon - v_0)|x] = 0, \tag{1.8}$$

where the function $q(u)$ is nonpositive for $u < 0$ and nonnegative otherwise. Often the constant term v_0 can be expressed as the solution to a conditional minimization problem, $v_0 = \text{argmin}_b E[r(\varepsilon - b)|x]$, where $r(u)$ is an antiderivative of $-q(u)$; this representation is often used as the basis for construction of estimators under these restrictions. In a limiting case, if $r(u)$ is taken to be minus the Dirac delta function, this corresponds to a *conditional mode* restriction, which asserts constancy of $v_0 = \max_u f_{\varepsilon|x}(u|x)$, where $f_{\varepsilon|x}$ is the conditional density of the errors. This restriction is useful for identification of the parameters of certain semiparametric models involving truncation.

A stronger condition which implies both the conditional mean (when it exists) and conditional median restrictions is a *conditional symmetry* restriction, under which

$$\Pr\{(\varepsilon - v_0) \leqslant u|x\} = \Pr\{(v_0 - \varepsilon) \leqslant u|x\} \tag{1.9}$$

for some constant v_0 and any conformable u. Again for scalar errors, this restriction implies (1.6) and (1.7) hold (when the expectations are well-defined) whenever $q(u)$ is an odd function of u, which may also depend upon the regressors x in general; here the value v_0 is constant across different choices of $q(\cdot)$. A different restriction which is equivalent to imposition of all possible conditional location restrictions is an *independence* restriction:

$$\Pr\{\varepsilon \leqslant u|x\} = \Pr\{\varepsilon \leqslant u\} \tag{1.10}$$

for all conformable u. Estimators based upon conditional mean or median restrictions will also be well-behaved under conditional symmetry or independence restrictions, but efficient estimation will generally require other choices for $q(u)$ in (1.8) than $q(u) = u$ or $q(u) = \text{sgn}(u)$.

Finally, a class of stochastic restrictions which can be viewed as generalizations of constant conditional mean or independence of the errors and regressors are *index* restrictions. A *strong* or *distributional index* restriction on the error terms is

an assumption that

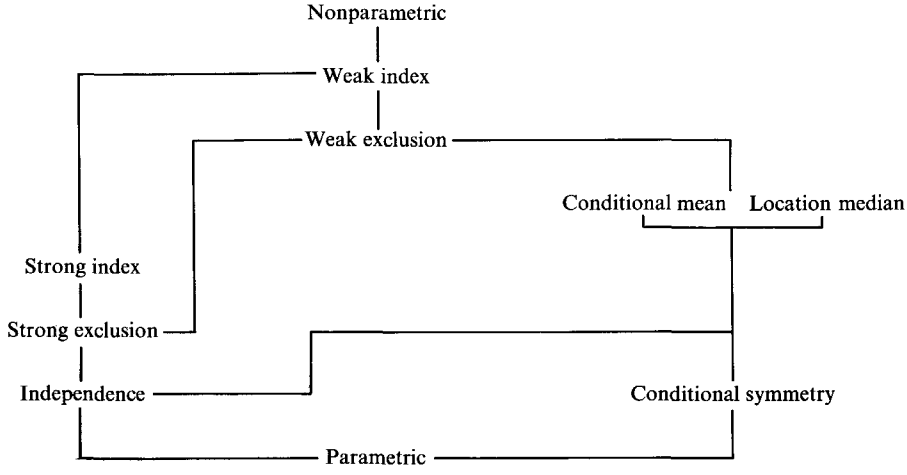
$$\Pr\{\varepsilon \leqslant u|x\} = \Pr\{\varepsilon \leqslant u|v(x)\} \tag{1.11}$$

for some "index" function $v(x)$ with $\dim\{v(x)\} < \dim\{x\}$; a *weak* or *mean index* restriction asserts a similar property only for the conditional expectation –

$$E[\varepsilon|x] = E[\varepsilon|v(x)]. \tag{1.12}$$

For different structural models, the index function $v(x)$ might be assumed to be a known function of x, or known up to a finite number of unknown parameters (e.g. $v(x) = x'\beta_0$), or an unknown function of known dimensionality (in which case some extra restriction(s) will be needed to identify the index). As a special case, the function $v(x)$ may be trivial, which yields the independence or conditional mean restrictions as special cases; more generally, $v(x)$ might be a known subvector x_1 of the regressors x, in which case (1.11) and (1.12) are strong and weak forms of an *exclusion* restriction, otherwise known as *conditional independence* and *conditional mean independence* of ε and x given x_1, respectively. When the index function is unknown, it is often assumed to be linear in the regressors, with coefficients that are related to unknown parameters of interest in the structural model.

The following diagram summarizes the hierarchy of the stochastic restrictions to be discussed in the following sections of this chapter, with declining level of generality from top to bottom:



Turning now to a description of some structural models treated in the semiparametric literature, an important class of parametric forms for the structural

functions is the class of *linear latent variable models*, in which the dependent variable y is assumed to be generated as some transformation

$$y = t(y^*; \lambda_0, \tau_0(\cdot)) \tag{1.13}$$

of some unobservable variable y^*, which itself has a linear regression representation

$$y^* = x'\beta_0 + \varepsilon. \tag{1.14}$$

Here the regression coefficients β_0 and the finite-dimensional parameters λ_0 of the transformation function are the parameters of interest, while the error distribution and any nonparametric component $\tau_0(\cdot)$ of the transformation make up the nonparametric component of the model. In general y and y^* may be vector-valued, and restrictions on the coefficient matrix β_0 may be imposed to ensure identification of the remaining parameters. This class of models, which includes the classical linear model as a special case, might be broadened to permit a nonlinear (but parametric) regression function for the latent variable y^*, as long as the additivity of the error terms in (1.14) is maintained.

One category of latent variable models, *parametric transformation models*, takes the transformation function $t(y^*; \lambda_0)$ to have no nonparametric nuisance component $\tau_0(\cdot)$ and to be invertible in y^* for all possible values of λ_0. A well-known example of a parametric transformation model is the Box–Cox regression model (Box and Cox (1964)), which has $y = t(x'\beta_0 + \varepsilon; \lambda_0)$ for

$$t^{-1}(y; \lambda) = \frac{y^\lambda - 1}{\lambda} 1\{\lambda \neq 0\} + \ln(y) 1\{\lambda = 0\}.$$

This transformation, which includes linear and log-linear (in y) regression models as special cases, requires the support of the latent variable y^* to be bounded from below (by $-1/\lambda_0$) for noninteger values of λ_0, but has been extended by Bickel and Doksum (1981) to unbounded y^*. Since the error term ε can be expressed as a known function of the observable variables and unknown parameters for these models, a stochastic restriction on ε (like a conditional mean restriction, defined below) translates directly into a restriction on y, x, β_0, and λ_0 which can be used to construct estimators.

Another category, *limited dependent variable models*, includes latent variable models in which the transformation function $t(y^*)$ which does not depend upon unknown parameters, but which is noninvertible, mapping intervals of possible y^* values into single values of y. Scalar versions of these models have received much of the attention in the econometric literature on semiparametric estimation, owing to their relative simplicity and the fact that parametric methods generally yield inconsistent estimators for β_0 when the functional form of the error distribution is misspecified. The simplest nontrivial transformation in this category is

an indicator for positivity of the latent variable y^*, which yields the *binary response model*

$$y = 1\{x'\beta_0 + \varepsilon > 0\}, \tag{1.15}$$

which is commonly used in econometric applications to model dichotomous choice problems. For this model, in which the parameters can be identified at most up to a scale normalization on β_0 or ε, the only point of variation of the function $t(y^*)$ occurs at $y^* = 0$, which makes identification of β_0 particularly difficult. A model which shares much of the structure of the binary response model is the *ordered response model*, with the latent variable y^* is only known to fall in one of $J+1$ ordered intervals $\{(-\infty, c_0], (c_0, c_1], \ldots, (c_J, \infty)\}$; that is,

$$y = \sum_{j=1}^{J} 1\{x'\beta_0 + \varepsilon > c_j\}. \tag{1.16}$$

Here the thresholds $\{c_j\}$ are assumed unknown (apart from a normalization like $c_0 \equiv 0$), and must be estimated along with β_0. The *grouped dependent variable model* is a variation with known values of $\{c_j\}$, where the values of y might correspond to prespecified income intervals.

A structural function for which the transformation function is more "informative" about β_0 is the *censored regression model*, also known in econometrics as the *censored Tobit model* (after Tobin (1956)). Here the observable dependent variable is assumed to be subject to a nonnegativity constraint, so that

$$y = \max\{0, x'\beta_0 + \varepsilon\}; \tag{1.17}$$

this structural function is often used as a model of individual demand or supply for some good when a fraction of individuals do not participate in that market. A variation on this model, the *accelerated failure time model with fixed censoring*, can be used as a model for duration data when some durations are incomplete. Here

$$y = \min\{x_1'\beta_0 + \varepsilon, x_2\}, \tag{1.18}$$

where y is the logarithm of the observable duration time (e.g. an unemployment spell), and x_2 is the logarithm of the duration of the experiment (following which the time to completion for any ongoing spells is unobserved); the "fixed" qualifier denotes models in which both x_1 and x_2 are observable (and may be functionally related).

These univariate limited dependent variable models have multivariate analogues which have also been considered in the semiparametric literature. One multivariate generalization of the binary response model is the *multinomial response*

model, for which the dependent variable is a J-dimensional vector of indicators, $y = \text{vec}\{y_j, j = 1, \ldots, J\}$, with

$$y_j = 1\{y_j^* \geq y_k^* \quad \text{for} \quad k \neq j\} \tag{1.19}$$

and with each latent variable y_j^* generated by a linear model

$$y_j^* = x'\beta_0^j + \varepsilon_j, \qquad \beta_0 \equiv [\beta_0^1, \ldots, \beta_0^j, \ldots, \beta_0^J]. \tag{1.20}$$

That is, $y_j = 1$ if and only if its latent variable y_j^* is the largest across alternatives. Another bivariate model which combines the binary response and censored regression models is the *censored sample selection model*, which has one binary response variable y_1 and one quantitative dependent variable y_2 which is observed only when $y_1 = 1$:

$$y_1 = 1(x_1'\beta_0^1 + \varepsilon_1 > 0) \tag{1.21}$$

and

$$y_2 = y_1[x_2'\beta_0^2 + \varepsilon_2]. \tag{1.22}$$

This model includes the censored regression model as a special case, with $\beta_0^1 = \beta_0^2 \equiv \beta_0$ and $\varepsilon_1 = \varepsilon_2 \equiv \varepsilon$. A closely related model is the *disequilibrium regression model with observed regime*, for which only the smaller of two latent variables is observed, and it is known which variable is observed:

$$y_1 = 1(x_1'\beta_0^1 + \varepsilon_1 < x_2'\beta_0^2 + \varepsilon_2) \tag{1.23}$$

and

$$y_2 = \min\{x_1'\beta_0^1 + \varepsilon_1, x_2'\beta_0^2 + \varepsilon_2\} = y_1[x_1'\beta_0^1 + \varepsilon_1] + (1 - y_1)[x_2'\beta_0^2 + \varepsilon_2]. \tag{1.24}$$

A special case of this model, the *randomly censored regression model*, imposes the restriction $\beta_0^2 = 0$, and is a variant of the duration model (1.18) in which the observable censoring threshold x_2 is replaced by a random threshold ε_2 which is unobserved for completed spells.

A class of limited dependent variable models which does not neatly fit into the foregoing latent variable framework is the class of *truncated dependent variable models*, which includes the *truncated regression* and *truncated sample selection models*. In these models, an observable dependent variable y is constructed from latent variables drawn from a particular subset of their support. For the truncated regression model, the dependent variable y has the distribution of $y^* = x'\beta_0 + \varepsilon$

conditional on $y^* > 0$:

$$y = x'\beta_0 + v, \tag{1.25}$$

with

$$\Pr\{v \leqslant c|x\} = \Pr\{\varepsilon \leqslant c|x, \quad \varepsilon > -x'\beta_0\}. \tag{1.26}$$

For the truncated selection model, the dependent variable y is generated in the same way as y_2 in (1.24), conditionally on $y_1 = 1$. Truncated models are variants of censored models for which no information on the conditioning variables x is available when the latent variable y^* cannot be observed. Since truncated samples can be constructed from their censored counterparts by deleting censored observations, identification and estimation of the parameters of interest is more challenging for truncated data.

An important class of multivariate latent dependent variable models arises in the analysis of panel data, where the dimensionality of the dependent variable y is proportional to the number of time periods each individual is observed. For concreteness, consider the special case in which a scalar dependent variable is observed for two time periods, with subscripts on y and x denoting time period; then a latent variable analogue of the standard linear "fixed effects" model for panel data has

$$\begin{aligned} y_1 &= t(\gamma + x_1'\beta_0 + \varepsilon_1, \tau_0), \\ y_2 &= t(\gamma + x_2'\beta_0 + \varepsilon_2, \tau_0), \end{aligned} \tag{1.27}$$

where $t(\cdot)$ is any of the transformation functions discussed above and γ is an unobservable error term which is constant across time periods (unlike the time-specific errors ε_1 and ε_2) but may depend in an arbitrary way on the regressors x_1 and x_2. Consistent estimation of the parameters of interest β_0 for such models is a very challenging problem; while "time-differencing" or "deviation from cell means" eliminates the fixed effect for linear models, these techniques are not applicable to nonlinear models, except in certain special cases (as discussed by Chamberlain (1984)). Even when the joint distribution of the error terms ε_1 and ε_2 is known parametrically, maximum likelihood estimators for β_0, τ_0 and the distributional parameters will be inconsistent in general if the unknown values of γ are treated as individual-specific intercept terms (as noted by Heckman and MaCurdy (1980)), so semiparametric methods will be useful even when the distribution of the fixed effects is the only nuisance parameter of the model.

The structural functions considered so far have been assumed known up to a finite-dimensional parameter. This is not the case for the *generalized regression*

model, which has

$$y = \tau_0(x'\beta_0 + \varepsilon), \tag{1.28}$$

for some transformation function $\tau_0(\cdot)$ which is of unknown parametric form, but which is restricted either to be monotonic (as assumed by Han (1987a)), or smooth (or both). Formally, this model includes the univariate limited dependent variable and parametric transformation models as special cases; however, it is generally easier to identify and estimate the parameters of interest when the form of the transformation function $t(\cdot)$ is (parametrically) known.

Another model which at first glance has a nonparametric component in the structural component is the *partially linear* or *semilinear regression model* proposed by Engle et al. (1986), who labelled it the "semiparametric regression model"; estimation of this model was also considered by Robinson (1988). Here the regression function is a nonparametric function of a subset x_1 of the regressors, and a linear function of the rest:

$$y = x_2'\beta_0 + \lambda_0(x_1) + \varepsilon, \tag{1.29}$$

where $\lambda_0(\cdot)$ is unknown but smooth. By defining a new error term $\varepsilon^* = \lambda_0(x_1) + \varepsilon$, a constant conditional mean assumption on the original error term ε translates into a mean exclusion restriction on the error terms in an otherwise-standard linear model.

Yet another class of models with a nonparametric component are *generated regressor models*, in which the regressors x appear in the structural equation for y indirectly, through the conditional mean of some other observable variable w given x:

$$y = h(E[w|x], \alpha_0, \varepsilon) \equiv g(x, \alpha_0, \delta_0(\cdot), \varepsilon), \tag{1.30}$$

with $\delta_0(x) \equiv E[w|x]$. These models arise when modelling individual behavior under uncertainty, when actions depend upon predictions (here, conditional expectations) of unobserved outcomes, as in the large literature on "rational expectations". Formally, the nonparametric component in the structural function can be absorbed into an unobservable error term satisfying a conditional mean restriction; that is, defining $\eta \equiv w - E[w|x]$ (so that $E[\eta|x] \equiv 0$), the model (1.30) with nonparametrically-generated regressors can be rewritten as $y = g(w - \eta, \alpha_0, \varepsilon)$, with a conditional mean restriction on the extra error term η. In practice, this alternative representation is difficult to manipulate unless $g(\cdot)$ is linear, and estimators are more easily constructed using the original formulation (1.30).

Although the models described above have received much of the attention in the econometric literature on semiparametrics, they by no means exhaust the set of models with parametric and nonparametric components which are used in

econometric applications. One group of semiparametric models, not considered here, include the *proportional hazards model* proposed and analyzed by Cox (1972, 1975) for duration data, and duration models more generally; these are discussed by Lancaster (1990), among many others. Another class of semiparametric models which is not considered here are *choice-based* or *response-based sampling models*; these are similar to truncated sampling models, in that the observations are drawn from sub-populations with restricted ranges of the dependent variable, eliminating the ancillarity of the regressors x. These models are discussed by Manski and McFadden (1981) and, more recently, by Imbens (1992).

1.4. *Objectives and techniques of asymptotic theory*

Because of the generality of the restrictions imposed on the error terms for semiparametric models, it is very difficult to obtain finite-sample results for the distribution of estimators except for special cases. Therefore, analysis of semiparametric models is based on large-sample theory, using classical limit theorems to approximate the sampling distribution of estimators. The goals and methods to derive this asymptotic distribution theory, briefly described here, are discussed in much more detail in the chapter by Newey and McFadden in this volume.

As mentioned earlier, the first step in the statistical analysis of a semiparametric model is to demonstrate *identification* of the parameters α_0 of interest; though logically distinct, identification is often the first step in construction of an estimator of α_0. To identify α_0, at least one function $T(\cdot)$ must be found that yields $T(F_0) = \alpha_0$, where F_0 is the true joint distribution function of $z \equiv (y, x)$ (as in (1.3) above). This functional may be implicit: for example, α_0 may be shown to uniquely solve some functional equation $T(F_0, \alpha_0) = 0$ (e.g. $E[m(y, x, \alpha_0)] = 0$, for some $m(\cdot)$). Given the functional $T(\cdot)$ and a random sample $\{z_i \equiv (y_i, x_i), i = 1, \ldots, N\}$ of observations on the data vector z, a natural estimator of α_0 is

$$\hat{\alpha} = T(\hat{F}), \tag{1.31}$$

where $\hat{F}$ is a suitable estimator of the joint distribution function F_0. *Consistency* of $\hat{\alpha}$ (i.e. $\hat{\alpha} \to \alpha_0$ in probability as $N \to \infty$) is often demonstrated by invoking a law of large numbers after approximating the estimator as a sample average:

$$\hat{\alpha} = \frac{1}{N} \sum_{i=1}^{N} \varphi_N(y_i, x_i) + o_p(1), \tag{1.32}$$

where $E[\varphi_N(y, x)] \to \alpha_0$. In other settings, consistency is demonstrated by showing that the estimator maximizes a random function which converges uniformly and almost surely to a limiting function with a unique maximum at the true value α_0. As noted below, establishing (1.31) can be difficult if construction of $\hat{\alpha}$ involves

explicit nonparametric estimators (through smoothing of the empirical distribution function).

Once consistency of the estimator is established, the next step is to determine its *rate of convergence*, i.e. the steepest function $h(N)$ such that $h(N)(\hat{\alpha} - \alpha_0) = O_p(1)$. For regular parametric models, $h(N) = \sqrt{N}$, so this is a maximal rate under weaker semiparametric restrictions. If the estimator $\hat{\alpha}$ has $h(N) = \sqrt{N}$ (in which case it is said to be *root-N-consistent*), then it is usually possible to find conditions under which the estimator has an *asymptotically linear* representation:

$$\hat{\alpha} = \alpha_0 + \frac{1}{N}\sum_{i=1}^{N} \psi(y_i, x_i) + o_p(1/\sqrt{N}), \tag{1.33}$$

where the "influence function" $\psi(\cdot)$ has $E[\psi(y, x)] = 0$ and finite second moments. The Lindeberg–Levy central limit theorem then yields *asymptotic normality* of the estimator,

$$\sqrt{N}(\hat{\alpha} - \alpha_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V_0), \tag{1.34}$$

where $V_0 = E\{\psi(y, x)[\psi(y, x)]'\}$. With a consistent estimator of V_0 (formed as the sample covariance matrix of some consistent estimator $\hat{\psi}(y_i, x_i)$ of the influence function), confidence regions and test statistics can be constructed with coverage/rejection probabilities which are approximately correct in large samples.

For semiparametric models, as defined above, there will be other functionals $T^+(F)$ which can be used to construct estimators of the parameters of interest. The *asymptotic efficiency* of a particular estimator $\hat{\alpha}$ can be established by showing that its asymptotic covariance matrix V_0 in (1.34) is equal to the semiparametric analogue to the Cramér–Rao bound for estimation of α_0. This *semiparametric efficiency bound* is obtained as the smallest of all efficiency bounds for parametric models which satisfy the semiparametric restrictions. The representation $\alpha_0 = T^*(F_0)$ which yields an efficient estimator generally depends on some component $\delta_0(\cdot)$ of the unknown, infinite-dimensional nuisance parameter $\eta_0(\cdot)$, i.e. $T^*(\cdot) = T^*(\cdot, \delta_0)$, so construction of an efficient estimator requires explicit nonparametric estimation of some characteristics of the nuisance parameter.

Demonstration of (root-N) consistency and asymptotic normality of an estimator depends on the complexity of the asymptotic linearity representation (1.33), which in turn depends on the complexity of the estimator. In the simplest case, where the estimator can be written in a closed form as a smooth function of sample averages,

$$\hat{\alpha} = a\left(\frac{1}{N}\sum_{i=1}^{N} m(y_i, x_i)\right), \tag{1.35}$$

the so-called "delta method" yields an influence function ψ of the form

$$\psi(y,x) = [\partial a(\mu_0)/\partial \mu][m(y,x) - \mu_0], \tag{1.36}$$

where $\mu_0 \equiv E[m(y,x)]$. Unfortunately, except for the classical linear model with a conditional mean restriction, estimators for semiparametric models are not of this simple form. Some estimators for models with weak index or exclusion restrictions on the errors can be written in closed form as functions of bivariate U-statistics,

$$\hat{\alpha} = a\left[\binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_N(z_i, z_j)\right] \equiv a(\hat{U}_N), \tag{1.37}$$

with "kernel" function p_N that has $p_N(z_i, z_j) = p_N(z_j, z_i)$ for $z_i \equiv (y_i, z_i)$; under conditions given by Powell et al. (1989), the representation (1.33) for such an estimator has influence function ψ of the same form as in (1.36), where now

$$m(y,x) = \lim_{N \to \infty} E[p_N(z_i, z_j) | z_i = (y,x)], \qquad \mu_0 = E[m(y,x)]. \tag{1.38}$$

A consistent estimator of the asymptotic covariance matrix of $\hat{\alpha}$ of (1.37) is the sample second moment matrix of

$$\hat{\psi}(y_i, x_i) = [\partial a(\hat{U}_N)/\partial \mu]\left[\frac{1}{N-1} \sum_{j \neq i} p_N(z_i, z_j) - \hat{U}_N\right]. \tag{1.39}$$

In most cases, the estimator $\hat{\alpha}$ will not have a closed-form expression like in (1.35) or (1.37), but instead will be defined implicitly as a minimizer of some sample criterion function or a solution of estimating equations. Some (generally inefficient) estimators based on conditional location or symmetry restrictions are "M-estimators", defined as minimizers of an empirical process

$$\hat{\alpha} = \underset{\alpha \in \Theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \rho(y_i, x_i, \alpha) \equiv \underset{\alpha \in \Theta}{\operatorname{argmin}}\, S_N(\alpha) \tag{1.40}$$

and/or solutions of estimating equations

$$0 = \frac{1}{N} \sum_{i=1}^{N} m(y_i, x_i, \hat{\alpha}) \equiv \bar{m}_N(\hat{\alpha}). \tag{1.41}$$

for some functions $\rho(\cdot)$ and $m(\cdot)$, with $\dim\{m(\cdot)\} = \dim(\alpha)$. When $\rho(y,x,\alpha)$ (or $m(y,x,\alpha)$) is a uniformly continuous function in the parameters over the entire parameter space Θ (with probability one), a standard uniform law of large numbers can be used to ensure that normalized versions of these criteria converge to their

expectations uniformly on the parameter space. This, along with an identification condition – namely, uniqueness of α_0 as a minimizer of $E[\rho(y,x,\alpha)]$ or solution to $0 = E[m(y,x,\alpha)]$ over $\alpha \in \Theta$ – ensures consistency of the estimator $\hat{\alpha}$ defined by (1.40) or (1.41). When $\rho(\cdot)$ (or $m(\cdot)$) is discontinuous in the parameters, uniform convergence of the empirical processes $S_N(\alpha)$ or $\bar{m}_N(\alpha)$ can usually be obtained by exploiting the special structure of $\rho(\cdot)$ or $m(\cdot)$, using the results by Huber (1967), Pollard (1985) and Pakes and Pollard (1989) described in the chapters on asymptotic theory in this volume. Under the regularity conditions imposed in these papers, the M-estimator $\hat{\alpha}$ will have an asymptotic linearity representation (1.33), with influence function

$$\psi(y,x) = -[\partial E[m(y,x,\alpha)]/\partial\alpha'|_{\alpha=\alpha_0}]^{-1} m(y,x,\alpha_0), \tag{1.42}$$

where $m(\cdot) = \partial\rho(\cdot)/\partial\alpha$ for the estimator defined by (1.40). More generally, the functions $\rho(\cdot)$ and $m(\cdot)$ may vary with the sample size N, in which case $m(\cdot) \equiv \lim m_N(\cdot) \equiv \lim \partial\rho_N(\cdot)/\partial\alpha$.

One variation on the M-estimator exploits moment restrictions $E[m(y,x,\alpha_0)] = 0$ when $\dim\{m(\cdot)\} > \dim(\alpha)$. A generalized method-of-moments (GMM) estimator is defined as

$$\hat{\alpha} = \underset{\alpha\in\Theta}{\operatorname{argmin}}\, [\bar{m}_N(\alpha)]' A_N [\bar{m}_N(\alpha)], \tag{1.43}$$

where $\bar{m}_N(\alpha)$ is defined in (1.41) and A_N is a sequence of positive semi-definite matrices converging in probability to some matrix A_0. Estimators based on conditional mean restrictions are generally of this form. Under similar regularity conditions as for M-estimators, GMM estimators will be consistent and asymptotically linear, with influence function

$$\psi(y,x) \equiv -[M_0' A_0 M_0]^{-1} M_0' A_0 m(y,x,\alpha_0), \tag{1.44}$$

where

$$M_0 \equiv \partial E[m(y,x,\alpha)]/\partial\alpha'|_{\alpha=\alpha_0}. \tag{1.45}$$

As pointed out by Hansen (1982), the asymptotic variance of the GMM estimator is minimized by choosing A_N so that its probability limit A_0 is proportional to the inverse of the covariance matrix $E[m(y,x,\alpha_0)m'(y,x,\alpha_0)]$ of the moment functions.

Another variation of the M-estimator of (1.40) defines the estimator $\hat{\alpha}$ as a minimizer of a bivariate U-process,

$$\hat{\alpha} = \underset{\alpha\in\Theta}{\operatorname{argmin}} \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_N(z_i, z_j, \alpha) \equiv \underset{\alpha\in\Theta}{\operatorname{argmin}}\, U_N(\alpha), \tag{1.46}$$

where the kernel $p_N(\cdot)$ has the same symmetry property as stated for (1.37) above; such estimators arise for models with independence or index restrictions on the error terms. Results by Nolan and Pollard (1987, 1988), Sherman (1993) and Honoré and Powell (1991) can be used to establish the consistency and asymptotic normality of this estimator, which will have an influence function of the form (1.42) when

$$m(y, x, \alpha) = \lim_{N \to \infty} \partial E[p_N(z_i, z_j, \alpha) | y_i = y, x_i = x] / \partial \alpha. \tag{1.47}$$

A more difficult class of estimators to analyze are those termed "semiparametric M-estimators" by Horowitz (1988a), for which the estimating equations in (1.41) also depend upon an estimator of a nonparametric component $\delta_0(\cdot)$; that is, $\hat{\alpha}$ solves

$$0 = \frac{1}{N} \sum_{i=1}^{N} m(y_i, x_i, \hat{\alpha}, \hat{\delta}(\cdot)) \equiv \bar{m}_N(\hat{\alpha}, \hat{\delta}(\cdot)) \tag{1.48}$$

for some nonparametric estimator $\hat{\delta}$ of δ_0. This condition might arise as a first-order condition for minimization of an empirical loss function that depends on $\hat{\delta}$,

$$\hat{\alpha} = \underset{\alpha \in \Theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \rho(y_i, x_i, \alpha, \hat{\delta}(\cdot)), \tag{1.49}$$

as considered by Andrews (1990a, b). As noted above, an efficient estimator for any semiparametric model is generally of this form and estimators for models with independence or index restrictions are often in this class. To derive the influence function for an estimator satisfying (1.48), a functional mean-value expansion of $\bar{m}_N(\hat{\alpha}, \hat{\delta})$ around $\hat{\delta} = \delta_0$ can be used to determine the effect on $\hat{\alpha}$ of estimation of δ_0. Formally, condition (1.48) yields

$$0 = \bar{m}_N(\hat{\alpha}, \hat{\delta}(\cdot)) = \bar{m}_N(\hat{\alpha}, \delta_0(\cdot)) + L_0(\hat{\delta}(\cdot) - \delta_0(\cdot)) + o_p(1/\sqrt{N}) \tag{1.50}$$

for some linear functional L_0; then, with an influence function representation of this second term

$$L_0(\hat{\delta}(\cdot) - \delta_0(\cdot)) = \frac{1}{N} \sum_{i=1}^{N} \xi(y_i, x_i) + o_p(1/\sqrt{N}) \tag{1.51}$$

(with $E[\xi(y, x)] = 0$), the form of the influence function for a semiparametric M-estimator is

$$\psi(y, x) = [\partial E(m(y, x, \alpha, \delta_0)] / \partial \alpha' |_{\alpha = \alpha_0}]^{-1} [m(y, x, \alpha_0, \delta_0) + \xi(y, x)]. \tag{1.52}$$

To illustrate, suppose δ_0 is finite-dimensional, $\delta_0 \in \mathbb{R}^k$; then the linear functional in (1.50) would be a matrix product,

$$L_0(\hat{\delta}(\cdot) - \delta_0(\cdot)) \equiv L_0(\hat{\delta} - \delta_0) \equiv [\partial E(m(y, x, \alpha, \delta)/\partial \delta'|_{\alpha=\alpha_0, \delta=\delta_0}](\hat{\delta} - \delta_0), \quad (1.53)$$

and the additional component ξ of the influence function in (1.52) would be the product of the matrix L_0 with the influence function of the preliminary estimator $\hat{\delta}$. When δ_0 is infinite-dimensional, calculation of the linear functional L_0 and the associated influence function ξ depends on the nature of the nuisance parameter δ_0 and how it enters the moment function $m(y, x, \alpha, \delta)$. One important case has δ_0 equal to the conditional expectation of some function $s(y, x)$ of the data given some other function $v(x)$ of the regressors, with $m(\cdot)$ a function only of the fitted values of this expectation; that is,

$$\delta_0 = \delta_0(v(x)) = E[s(y, x)|v(x)] \quad (1.54)$$

and

$$m(y, x, \alpha, \delta(\cdot)) = m(y, x, \alpha, \delta(v(x))), \quad (1.55)$$

with $\partial m/\partial \delta$ well-defined. For instance, this is the structure of efficient estimators for conditional location restrictions. For this case, Newey (1991) has shown that the adjustment term $\xi(y, x)$ to the influence function of a semiparametric M-estimator $\hat{\alpha}$ is of the form

$$\xi(y, x) = [\partial E(m(y, x, \alpha, \delta)|v(x)]/\partial \delta'|_{\alpha=\alpha_0}][s(y, x) - \delta_0(v(x))]. \quad (1.56)$$

In some cases the leading matrix in this expression is identically zero, so the asymptotic distribution of the semiparametric M-estimator is the same as if $\delta_0(\cdot)$ were known; Andrews (1990a, b) considered this and other settings for which the adjustment term ξ is identically zero, giving regularity conditions for validity of the expansion (1.50) in such cases. General formulae for the influence functions of more complicated semiparametric M-estimators are derived by Newey (1991) and are summarized in Andrews' and Newey and McFadden's chapters in this volume.

2. Stochastic restrictions

This section discusses how various combinations of structural equations and stochastic restrictions on the unobservable errors imply restrictions on the joint distribution of the observable data, and presents general estimation methods for the parameters of interest which exploit these restrictions on observables. The classification scheme here is the same as introduced in the monograph by Manski

(1988b) (and also in Manski's chapter in this volume), although the discussion here puts more emphasis on estimation techniques and properties. Readers who are familiar with this material or who are interested in a particular structural form, may wish to skip ahead to Section 3 (which reviews the literature for particular models), referring back to this section when necessary.

2.1. *Conditional mean restriction*

As discussed in Section 1.3 above, the class of constant conditional location restrictions for the error distribution assert constancy of

$$\nu_0 = \underset{b}{\operatorname{argmin}}\, E[r(\varepsilon - b)|x], \tag{2.1}$$

for some function $r(\cdot)$ which is nonincreasing for negative arguments and nondecreasing for positive arguments; this implies a moment condition $E[q(\varepsilon - \mu_0)|x] = 0$, for $q(u) = \partial r(u)/\partial u$. When the loss function of (2.1) is taken to be quadratic, $r(u) = u'u$, the corresponding conditional location restriction imposes constancy of the conditional mean of the error terms,

$$E(\varepsilon|x) = \mu_0 \tag{2.2}$$

for some μ_0. By appropriate definition of the dependent variable(s) y and "exogenous" variables x, this restriction may be applied to models with "endogenous" regressors (that is, some components of x may be excluded from the restriction (2.2)).

This restriction is useful for identification of the parameters of interest for structural functions $g(x, \alpha, \varepsilon)$ that are invertible in the error terms ε; that is,

$$y = g(x, \alpha_0, \varepsilon) \Leftrightarrow \varepsilon = e(y, x, \alpha_0)$$

for some function $e(\cdot)$, so that the mean restriction (2.1) can be rewritten

$$E[e(y_i, x_i, \alpha_0) - \mu_0|x_i] = 0 = E[e(y_i, x_i, \alpha_0)|x_i], \tag{2.3}$$

where the latter equality imposes the normalization $\mu_0 \equiv 0$ (i.e., the mean μ_0 is appended to the vector α_0 of parameters of interest).

Conditional mean restrictions are useful for some models that are not completely specified – that is, for models in which some components of the structural function $g(\cdot)$ are unknown or unspecified. In many cases it is more natural to specify the function $e(\cdot)$ characterizing a subset of the error terms than the structural function $g(\cdot)$ for the dependent variable; for example, the parameters of interest may be coefficients of a single equation from a simultaneous equations system and it is

often possible to specify the function $e(\cdot)$ without specifying the remaining equations of the model. However, conditional mean restrictions generally are insufficient to identify the parameters of interest in noninvertible limited dependent variable models, as Manski (1988a) illustrates for the binary response model.

The conditional moment condition (2.3) immediately yields an unconditional moment equation of the form

$$0 = E[d(x)\, e(y, x, \alpha_0)], \tag{2.4}$$

where $d(x)$ is some conformable matrix with at least as many rows as the dimension of α_0. For a given function $d(\cdot)$, the sample analogue of the right-hand side of (2.8) can be used to construct a method-of-moments or generalized method-of-moments estimator, as described in Section 1.4; the columns of the matrix $d(x)$ are "instrumental variables" for the corresponding rows of the error vector ε. More generally, the function $d(\cdot)$ may depend on the parameters of interest, α_0, and a (possibly) infinite-dimensional nuisance parameter $\delta_0(\cdot)$, so a semiparametric M-estimator for $\hat{\alpha}$ may be defined to solve

$$0 = \frac{1}{N} \sum_{i=1}^{N} d(x_i, \hat{\alpha}, \hat{\delta})\, e(y_i, x_i, \hat{\alpha}), \tag{2.5}$$

where $\dim(d(\cdot)) = \dim(\alpha) \times \dim(\varepsilon)$ and $\hat{\delta} = \hat{\delta}(\cdot)$ is a consistent estimator of the nuisance function $\delta_0(\cdot)$. For example, these sample moment equations arise as the first-order conditions for the GMM minimization given in (1.43), where the moment functions take the form $m(y, x, \alpha) = c(x)\, e(y, x, \alpha)$, for a matrix $c(x)$ of fixed functions of x with number of rows greater than or equal to the number components of α. Then, assuming differentiability of $e(\cdot)$, the GMM estimator solves (2.5) with

$$d(x, \hat{\alpha}, \hat{\delta}) \equiv \left\{ \frac{1}{N} \sum_{i=1}^{N} [\partial e(y_i, x_i, \hat{\alpha})/\partial \alpha']' [c(x_i)]' \right\} A_N c(x), \tag{2.6}$$

where A_N is the weight matrix given in (1.43).

Since the function $d(\cdot)$ depends on the data only through the conditioning variable x, it is simple to derive the form of the asymptotic distribution for the estimator $\hat{\alpha}$ which solves (2.5) using the results stated in Section 1.4:

$$\sqrt{N}(\hat{\alpha} - \alpha_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0, M_0^{-1} V_0 (M_0')^{-1}), \tag{2.7}$$

where

$$M_0 = \frac{\partial}{\partial \alpha'} E[d(x, \alpha, \delta_0)\, e(y, x, \alpha)]|_{\alpha = \alpha_0} = d(x, \alpha, \delta_0) \left\{ \frac{\partial}{\partial \alpha'} E[e(y, x, \alpha)|x_i]|_{\alpha = \alpha_0} \right\}$$

and

$$V_0 = E[d(x,\alpha_0,\delta_0)\, e(y,x,\alpha_0)\, e'(y,x,\alpha_0)\, d'(x,\alpha_0,\delta_0)]$$
$$= E[d(x,\alpha_0,\delta_0)\, \Sigma(x)\, d'(x_i,\alpha_0,\delta_0)].$$

In this expression, $\Sigma(x)$ is the conditional covariance matrix of the error terms,

$$\Sigma(x) \equiv E[e(y,x,\alpha_0)\, e'(y,x,\alpha_0)|x] = E[\varepsilon\varepsilon'|x].$$

Also, the expectation and differentiation in the definition of M_0 can often be interchanged, but the order given above is often well-defined even if $d(\cdot)$ or $e(\cdot)$ is not smooth in α.

A simple extension of the Gauss–Markov argument can be used to show that an efficient choice of instrumental variable matrix $d^*(x)$ is of the form

$$d^*(x) = d^*(x,\alpha_0,\delta_0) = \left\{ \frac{\partial}{\partial \alpha'} E[e(y,x,\alpha)|x_i]|_{\alpha=\alpha_0} \right\}' [\Sigma(x)]^{-1}; \tag{2.8}$$

the resulting efficient estimator $\hat{\alpha}^*$ will have

$$\sqrt{N}(\hat{\alpha}^* - \alpha_0) \xrightarrow{d} \mathcal{N}(0, V^*), \quad \text{with} \quad V^* = \{E[d^*(x)[\Sigma(x)][d^*(x)]']\}^{-1}, \tag{2.9}$$

under suitable regularity conditions. Chamberlain (1987) showed that V^* is the semiparametric efficiency bound for any "regular" estimator of α_0 when only the conditional moment restriction (2.3) is imposed. Of course, the optimal matrix $d^*(x)$ of instrumental variables depends upon the conditional distribution of y given x, an infinite-dimensional nuisance parameter, so direct substitution of $d^*(x)$ in (2.5) is not feasible. Construction of a feasible efficient estimator for α_0 generally uses nonparametric regression and a preliminary inefficient GMM estimator of α_0 to construct estimates of the components of $d^*(x)$, the conditional mean of $\partial e(y,x,\alpha_0)/\partial\alpha'$ and the conditional covariance matrix of $e(y,x,\alpha_0)$. This is the approach taken by Carroll (1982), Robinson (1987), Newey (1990b), Linton (1992) and Delgado (1992), among others. Alternatively, a "nearly" efficient sequence of estimators can be generated as a sequence of GMM estimators with moment functions of the form $m(y,x,\alpha) = c(x)\, e(y,x,\alpha)$, when the number of rows of $c(x)$ (i.e. the number of "instrumental variables") increases slowly as the sample size increases; Newey (1988a) shows that if linear combinations of $c(x)$ can be used to approximate $d^*(x)$ to an arbitrarily high degree as the size of $c(x)$ increases, then the asymptotic variance of the corresponding sequence of GMM estimators equals V^*.

For the linear model

$$y = x'\beta_0 + \varepsilon$$

with scalar dependent variable y, the form of the optimal instrumental variable matrix $d^*(x)$ simplifies to the vector

$$d^*(x) = [\sigma^2(x)]^{-1}x,$$

where $\sigma^2(x)$ is the conditional variance of the error term ε. As noted in Section 1.2 above, an efficient estimator for β_0 would be a weighted least squares estimator, with weights proportional to a nonparametric estimator of $[\sigma^2(x)]^{-1}$, as considered by Robinson (1987).

2.2. *Conditional quantile restrictions*

In its most general form, the conditional πth quantile of a scalar error term ε is defined to be any function $\eta(x;\pi)$ for which the conditional distribution of ε has at least probability π to the left and probability $1-\pi$ to the right of $\eta_\pi(x)$:

$$\Pr\{\varepsilon \leqslant \eta(x;\pi)|x\} \geqslant \pi \quad \text{and} \quad \Pr\{\varepsilon \geqslant \eta(x;\pi)|x\} \geqslant 1-\pi. \tag{2.10}$$

A conditional quantile restriction is the assumption that, for some $\pi \in (0,1)$, this conditional quantile is independent of x,

$$\eta(x;\pi) = \eta_0(\pi) \equiv \eta_0, \quad \text{a.s.} \tag{2.11}$$

Usually the conditional distribution of ε is further restricted to have no point mass at its conditional quantile ($\Pr\{\varepsilon = \eta_0\} = 0$), which with (2.10) implies the conditional moment restriction

$$E[\pi - 1\{\varepsilon < \eta_0\}|x] = 0 \equiv E[\pi - 1\{\varepsilon < 0\}|x], \tag{2.12}$$

where again the normalization $\eta_0 \equiv 0$ is imposed (absorbing η_0 as a component of α_0). To ensure uniqueness of the solution $\eta_0 = 0$ to this moment condition, the conditional error distribution is usually assumed to be absolutely continuous with nonnegative density in some neighborhood of zero. Although it is possible in principle to treat the proportion π as an unknown parameter, it is generally assumed that π is known in advance; most attention is paid to the special case $\pi = \frac{1}{2}$ (i.e. a conditional median restriction) which is implied by the stronger assumptions of either independence of the errors and regressors or conditional symmetry of the errors about a constant.

A conditional quantile restriction can be used to identify parameters of interest in models in which the dependent variable y and the error term ε are both scalar, and the structural function $g(\cdot)$ of (1.4) is nondecreasing in ε for all possible α_0 and almost all x:

$$u_1 \leqslant u_2 \Rightarrow g(x, \alpha, u_1) \leqslant g(x, \alpha, u_2), \quad \text{a.s.} \quad (x). \tag{2.13}$$

(Of course, nonincreasing structural functions can be accommodated with a sign change on the dependent variable y.) This monotonicity and the quantile restriction (2.11) imply that the conditional πth quantile of y given x is $g(x, \alpha_0, 0)$; since

$$\varepsilon \leqslant 0 \quad \text{or} \quad \varepsilon \geqslant 0 \quad \Rightarrow \quad y = g(x, \alpha_0, \varepsilon) \leqslant g(x, \alpha_0, 0) \quad \text{or} \quad y \geqslant g(x, \alpha_0, 0),$$

it follows that

$$\begin{aligned} &\Pr\{y \leqslant g(x, \alpha_0, 0) | x\} \geqslant \Pr\{\varepsilon \leqslant 0 | x\} \geqslant \pi \quad \text{and} \\ &\Pr\{y \geqslant g(x, \alpha_0, 0) | x\} \geqslant \Pr\{\varepsilon \geqslant 0 | x\} \geqslant 1 - \pi. \end{aligned} \tag{2.14}$$

Unlike a conditional mean restriction, a conditional quantile restriction is useful for identification of α_0 even when the structural function $g(x, \alpha, \varepsilon)$ is not invertible in ε. Moreover, the equivariance of quantiles to monotonic transformations means that, when it is convenient, a transformation $l(y)$ might be analyzed instead of the original dependent variable y, since the conditional quantile of $l(y)$ is $l(g(x, \alpha_0, 0))$ if $l(\cdot)$ is nondecreasing. (Note, though, that application of a noninvertible transformation may well make the parameters α_0 more difficult to identify.)

The main drawback with the use of quantile restrictions to identify α_0 is that the approach is apparently restricted to models with a scalar error term ε, because of their lack of additivity (i.e. quantiles of convolutions are not generally the sums of the corresponding quantiles) as well as the ambiguity of a monotonicity restriction on the structural function in a multivariate setting. Estimators based upon quantile restrictions have been proposed for the linear regression, parametric transformation, binary response, ordered response and censored regression models, as described in Section 3 below.

For values of x for which $g(x, \alpha_0, \varepsilon)$ is strictly increasing and differentiable at $\varepsilon = 0$, the moment restriction given in (2.12) and monotonicity restriction (2.13) can be combined to obtain a conditional moment restriction for the observable data and unknown parameter α_0. Let

$$b(x, \alpha) = 1\left\{\frac{\partial^- g(x, \alpha, 0)}{\partial \varepsilon} = \frac{\partial^+ g(x, \alpha, 0)}{\partial \varepsilon} \equiv \frac{\partial g(x, \alpha, 0)}{\partial \varepsilon} > 0\right\}; \tag{2.15}$$

then (2.12) immediately implies

$$E\{b(x, \alpha_0)[\pi - 1\{y < g(x, \alpha_0, 0)\}] | x\} \equiv E[m(y, x, \alpha_0) | x] = 0. \tag{2.16}$$

In principle, this conditional moment condition might be used directly to define a method-of-moments estimator for α_0; however, there are two drawbacks to this approach. First, the moment function $m(\cdot)$ defined above is necessarily a discontinuous function of the unknown parameters, complicating the asymptotic theory. More importantly, this moment condition is substantially weaker than the derived quantile restriction (2.14), since observations for which $g(x, \alpha_0, u)$ is not strictly increasing at $u = 0$ may still be useful in identifying the unknown parameters. As an extreme example, the binary response model has $b(x, \alpha_0) = 0$ with probability one under standard conditions, yet (2.14) can be sufficient to identify the parameters of interest even in this case (as discussed below).

An alternative approach to estimation of α_0 can be based on a characterization of the πth conditional quantile as the solution to a particular expected loss minimization problem. Define

$$R(b, x; \pi) \equiv E[\rho_\pi(y - b) - \rho_\pi(y) | x], \tag{2.17}$$

where

$$\rho_\pi(u) \equiv u[\pi - 1(u < 0)];$$

since $|\rho_\pi(u - b) - \rho_\pi(u)| \leqslant |b|$, this minimand is well-defined irrespective of the existence of moments of the data. It is straightforward to show that $Q(b, x)$ is minimized at $b^* = g(x, \alpha_0, 0)$ when (2.14) holds (more generally, $Q(b, x)$ will be minimized at any conditional πth quantile of y given x, as noted by Ferguson (1967)). Therefore, the true parameter vector α_0 will minimize

$$Q(\alpha; w(\cdot), \pi) \equiv E[w(x) R(g(x, \alpha, 0), x; \pi)] = E\{w(x)[\rho_\pi(y - g(x, \alpha, 0)) - \rho_\pi(y)]\} \tag{2.18}$$

over the parameter space, where $w(x)$ is any scalar, nonnegative function of x which has $E[w(x) \cdot |g(x, \alpha, 0)|] < \infty$. For a particular structural function $g(\cdot)$, then, the unknown parameters will be identified if conditions on the error distribution, regressors, and weight function $w(x)$ are imposed which ensure the uniqueness of the minimizer of $Q(\alpha; w(\cdot), \pi)$ in (2.18). Sufficient conditions are uniqueness of the πth conditional quantile $\eta_0 = 0$ of the error distribution and $\Pr\{w(x) > 0, g(x, \alpha, \eta) \neq g(x, \alpha_0, 0)\} > 0$ whenever $\alpha \neq \alpha_0$.

Given a sample $\{(y_i, x_i), \quad i = 1, \ldots, N\}$ of observations on y and x, the sample analogue of the minimand in (2.18) is

$$Q_N(\alpha; w(\cdot), \pi) \equiv \frac{1}{N} \sum_{i=1}^{N} w(x_i) \rho_\pi(y_i - g(x_i, \alpha, 0)), \tag{2.19}$$

where an additive constant which does not affect the minimization problem has been deleted. In general, the weight function $w(x)$ may be allowed to depend upon

nuisance parameters, $w(x) \equiv w(x, \delta_0)$, so a feasible weighted quantile estimator of α_0 might be defined to minimize $S_N(\alpha, \eta, \hat{w}(\cdot); \pi)$, with $\hat{w}(x) = w(x, \hat{\delta})$ for some preliminary estimator $\hat{\delta}$ of δ_0. In the special case of a conditional median restriction ($\pi = \frac{1}{2}$), minimization of Q_N is equivalent to minimization of a weighted sum of absolute deviations criterion

$$S_N(\alpha; w(\cdot)) \equiv 2Q_N(\alpha; w(\cdot), \tfrac{1}{2}) = \frac{1}{N} \sum_{i=1}^{N} w(x_i)|y_i - g(x_i, \alpha, 0)|, \tag{2.20}$$

which, with $w(x) \equiv 1$, is the usual starting point for estimation of the particular models considered in the literature cited below. When the structural function $g(\cdot)$ is of the latent variable form ($g(x, \alpha, \varepsilon) = t(x'\beta + \varepsilon, \tau)$), the estimator $\hat{\alpha}$ which minimizes $Q_N(\alpha; \hat{w}, \pi)$ will typically solve an approximate first-order condition,

$$\frac{1}{N} \sum_{i=1}^{N} \hat{w}(x_i)[\pi - 1(y_i < g(x_i, \hat{\alpha}, 0))]b(x_i, \hat{\alpha}) \frac{\partial g(x_i, \hat{\alpha}, 0)}{\partial \alpha} \cong 0, \tag{2.21}$$

where $b(x, \alpha)$ is defined in (2.15) and $\partial g(\cdot)/\partial \alpha$ denotes the vector of left derivatives. (The equality is only approximate due to the nondifferentiability of $\rho_\pi(u)$ at zero and possible nondifferentiability of $g(\cdot)$ at $\hat{\alpha}$; the symbol "$\cong$" in (2.21) means the left-hand side converges in probability to zero at an appropriate rate.) These equations are of the form

$$\frac{1}{N} \sum_{i=1}^{N} m(y_i, x_i, \alpha)\, d(x_i, \hat{\alpha}, \hat{\delta}) \cong 0,$$

where the moment function $m(\cdot)$ is defined in (2.16) and

$$d(x, \hat{\alpha}, \hat{\delta}) \equiv w(x_i, \hat{\delta}) b(x_i, \hat{\alpha}) \frac{\partial g(x_i, \hat{\alpha}, 0)}{\partial \alpha}.$$

Thus the quantile minimization problem yields an analogue to the unconditional moment restriction $E[m(y, x, \alpha_0)\, d(x, \alpha_0, \delta_0)] = 0$, which follows from (2.16).

As outlined in Section 1.4 above, under certain regularity conditions (given by Powell (1991)) the quantile estimator $\hat{\alpha}$ will be asymptotically normal,

$$\sqrt{N}(\hat{\alpha} - \alpha_0) \xrightarrow{d} \mathcal{N}(0, M_0^{-1} V_0 (M_0')^{-1}), \tag{2.22}$$

where now

$$M_0 \equiv E\left[f(0|x)\, w(x, \delta_0)\, b(x, \alpha_0) \frac{\partial g(x, \alpha_0, 0)}{\partial \alpha} \frac{\partial g(x, \alpha_0, 0)}{\partial \alpha'} \right]$$

and

$$V_0 \equiv E\left[\pi(1-\pi)w^2(x,\delta_0)b(x,\alpha_0)\frac{\partial g(x,\alpha_0,0)}{\partial \alpha}\frac{\partial g(x,\alpha_0,0)}{\partial \alpha'}\right],$$

for $f(0|x)$ being the conditional density of the "residual" $y - g(x,\alpha_0,0)$ at zero (which appears from the differentiation of the expectation of the indicator function in (2.21)). The "regularity" conditions include invertibility of the matrix M_0, which is identically zero for the binary and ordered response models; as shown by Kim and Pollard (1990), the rate of convergence of the estimator $\hat{\alpha}$ is slower than $\sqrt{N}$ for these models.

When (2.22) holds, an efficient choice of weight function $w(x)$ for this problem is

$$w^*(x) \propto f(0|x), \tag{2.23}$$

for which the corresponding estimator $\hat{\alpha}^*$ has

$$\sqrt{N}(\hat{\alpha}^* - \alpha_0) \xrightarrow{d} \mathcal{N}(0, V^*), \tag{2.24}$$

with

$$V^* = \pi(1-\pi)\left\{E\left[f^2(0|x)b(x,\alpha_0)\frac{\partial g(x,\alpha_0,0)}{\partial \alpha}\frac{\partial g(x,\alpha_0,0)}{\partial \alpha'}\right]\right\}^{-1}.$$

The matrix V^* was shown by Newey and Powell (1990) to be the semiparametric efficiency bound for the linear and censored regression models with a conditional quantile restriction, and this is likely to be the case for a more general class of structural models.

For the linear regression model $g(x,\alpha_0,\varepsilon) \equiv x'\beta_0 + \varepsilon$, estimation of the true coefficients β_0 using a least absolute deviations criterion dates from Laplace (1793); the extension to other quantile restrictions was proposed by Koenker and Bassett (1978). In this case $b(x,\alpha) \equiv 1$ and $\partial g(x,\alpha,\varepsilon)/\partial \alpha \equiv x$, which simplifies the asymptotic variance formulae. In the special case in which the conditional density of $\varepsilon \equiv y - x'\beta_0$ at zero is constant – $f(0|x) \equiv f_0$ – the asymptotic covariance matrix of the quantile estimator $\hat{\beta}$ further simplifies to

$$V^* = \pi(1-\pi)[f_0]^{-2}\{E[xx']\}^{-1}.$$

(Of course, imposition of the additional restriction of a constant conditional density at zero may affect the semiparametric information bound for estimation of β_0.) The monograph by Bloomfield and Steiger (1983) gives a detailed discussion of the

historical development, theoretical properties, and computational implementation of quantile estimators for linear models.

As noted in Section 1.3, conditional mean and median/quantile restrictions do not exhaust the set of conditional location restrictions appearing in the semiparametric literature. For example, M. Lee (1989) considers estimation based upon a conditional mode restriction, which imposes constancy of the maximizer of the conditional density of the error terms,

$$\max_e f_{\varepsilon|x}(e|x) = v_0.$$

For the linear model $y = x'\beta_0 + \varepsilon$, this restriction immediately implies a linear form for the conditional mode of y given x:

$$\max_y f_{y|x}(y|x) = v_0 + x'\beta_0.$$

More generally, a mode restriction may impose constancy of the modal interval of width ω for the error distribution, defined as

$$v_0(x,\omega) \equiv \operatorname*{argmax}_u \Pr\{|\varepsilon - u| \leqslant \omega/2|x\} = \operatorname*{argmax}_u E\{1(|\varepsilon - u| \leqslant \omega/2)|x\},$$

which yields the same linearity result for the modal ω-interval of y given x in a linear model. M. Lee (1989) proposes an estimator for the linear model which solves a sample maximization problem derived from this restriction, and verifies its consistency under suitable conditions; he also shows how this restriction can be used to construct consistent estimators for the truncated regression model, since modes are invariant to truncation (provided they do not overlap the truncated region).

2.3. *Conditional symmetry restrictions*

The assumption that the error terms ε are conditionally symmetrically distributed around a constant term $v_0 = 0$,

$$\Pr\{\varepsilon \leqslant u|x\} = \Pr\{-\varepsilon \leqslant u|x\} \tag{2.25}$$

for all u, clearly implies a constant conditional mean or median (when either is well-defined), so estimators which impose these weaker restrictions are also applicable under (2.25). More generally, a conditional symmetry restriction is useful for identification of the parameters of interest, α_0, for models that can be "symmetrized" in the error terms ε. Specifically, suppose that, for the structural relation $y = g(x, \alpha_0, \varepsilon)$, a function $h(y, x, \alpha)$ can be constructed where the composed function $h \circ g$ is an odd function of ε. That is,

$$h(g(x,\alpha,\varepsilon),x,\alpha) = -h(g(x,\alpha,-\varepsilon),x,\alpha) \tag{2.26}$$

for some $h(\cdot)$ and all possible x, α and ε. Then the random function $h(y, x, \alpha) = h(g(x, \alpha_0, \varepsilon), x, \alpha)$ will also be symmetrically distributed about zero when $\alpha = \alpha_0$, implying the conditional moment restriction

$$E[h(y, x, \alpha_0)|x] = E[h(g(x, \alpha_0, \varepsilon), x, \alpha_0)|x] = 0. \tag{2.27}$$

As with the previous restrictions, the conditional moment restriction can be used to generate an unconditional moment equation of the form $E[d(x)h(y, x, \alpha_0)] = 0$, with $d(x)$ a conformable matrix of instruments with a number of rows equal to the number of components of α_0. In general, the function $d(x)$ can be a function of α and nuisance parameters δ (possibly infinite-dimensional), so a semiparametric M-estimator $\hat{\alpha}$ of α_0 can be constructed to solve the sample moment equations

$$0 = \frac{1}{N}\sum_{i=1}^{N} d(x_i, \hat{\alpha}, \hat{\delta})h(y_i, x_i, \hat{\alpha}), \tag{2.28}$$

for $\hat{\delta}$ an estimator of some nuisance parameters δ_0.

For structural functions $g(x, \alpha, \varepsilon)$ which are invertible in the error terms, it is straightforward to find a transformation satisfying condition (2.26). Since $\varepsilon = e(y, x, \alpha)$ is an odd function of ε, $h(\cdot)$ can be chosen as this inverse function $e(\cdot)$. Even for noninvertible structural functions, it is still sometimes possible to find a "trimming" function $h(\cdot)$ which counteracts the asymmetry induced in the conditional distribution of y by the nonlinear transformation $g(\cdot)$. Examples discussed below include the censored and truncated regression models and a particular selectivity bias model.

As with the quantile estimators described in a preceding section, the moment condition (2.27) is sometimes insufficient to identify the parameters α_0, since the "trimming" transformation $h(\cdot)$ may be identically zero when evaluated at certain values of α in the parameter space. For example, the symmetrically censored least squares estimator proposed by Powell (1986b) for the censored regression model satisfies condition (2.27) with a function $h(\cdot)$ which is nonzero only when the fitted regression function $x_i'\hat{\beta}$ exceeds the censoring point (zero), so that the sample moment equation (2.28) will be trivially satisfied if $\hat{\beta}$ is chosen so that $x_i'\hat{\beta}$ is nonpositive for all observations. In this case, the estimator $\hat{\beta}$ was defined not only as a solution to a sample moment condition of the form (2.28), but in terms of a particular minimization problem $\hat{\beta} \equiv \operatorname{argmin}_{\Theta} S_n(\beta)$ which yields (2.28) as a first-order condition. The limiting minimand was shown to have a unique minimizer at β_0, even though the limiting first-order conditions have multiple solutions; thus, this further restriction on the acceptable solutions to the first-order condition was enough to ensure consistency of the estimator $\hat{\beta}$ for β_0. Construction of an analogous minimization problem might be necessary to fully exploit the symmetry restriction for other structural models.

Once consistency of a particular estimator $\hat{\alpha}$ satisfying (2.28) is established, the asymptotic distribution theory immediately follows from the GMM formulae pre-

sented in Section 2.1 above. For a particular choice of $h(\cdot)$, the form of the sample moment condition (2.28) is the same as condition (2.6) of Section 2.2 above, replacing the inverse transformation "$e(\cdot)$" with the more general "$h(\cdot)$" here; thus, the form of the asymptotically normal distribution of $\hat{\alpha}$ satisfying (2.28) is given by (2.7) of Section 2.2, again replacing "$e(\cdot)$" with "$h(\cdot)$".

Of course, the choice of the symmetrizing transformation $h(\cdot)$ is not unique – given any $h(\cdot)$ satisfying (2.26), another transformation $h^*(y, x, \alpha) \equiv l(h(y, x, \alpha), x, \alpha)$ will also satisfy (2.26) if $l(u, x, \alpha)$ is an odd function of u for all x and α. This multiplicity of possible symmetrizing transformations complicates the derivation of the semiparametric efficiency bounds for estimation of α_0 under the symmetry restriction, which are typically derived on a case-by-case basis. For example, Newey (1991) derived the semiparametric efficiency bounds for the censored and truncated regression models under the conditional symmetry restriction (2.25), and indicated how efficient estimators for these models might be constructed.

For the linear regression model $g(x, \alpha_0, \varepsilon) \equiv x'\beta + \varepsilon$, the efficient symmetrizing transformation $h(y, x, \beta)$ is the derivative of the log-density of ε given x, evaluated at the residual $y - x'\beta$, with optimal instruments equal to the regressors x:

$$h^*(y, x, \beta) = \partial \ln f_{\varepsilon|x}(y - x'\beta|x)/\partial \varepsilon, \qquad d^*(x, \beta, \delta) = x.$$

Here an efficient estimator might be constructed using a nonparametric estimator of the conditional density of ε given x, itself based on residuals $\tilde{e} = y - x'\tilde{\beta}$ from a preliminary fit of the model. Alternatively, as proposed by Cragg (1983) and Newey (1988a), an efficient estimator might be constructed as a sequence of GMM estimators, based on a growing number of transformation functions $h(\cdot)$ and instrument sets $d(\cdot)$, which are chosen to ensure that the sequence of GMM influence functions can approximate the influence function for the optimal estimator arbitrarily well. In either case, the efficient estimator would be "adaptive" for the linear model, since it would be asymptotically equivalent to the maximum likelihood estimator with known error density.

2.4. *Independence restrictions*

Perhaps the most commonly-imposed semiparametric restriction is the assumption of independence of the error terms and the regressors,

$$\Pr\{\varepsilon_i \leqslant \lambda | x_i\} \equiv \Pr\{\varepsilon_i \leqslant \lambda\} \quad \text{for all real } \lambda, \text{ w.p.1.} \tag{2.29}$$

Like conditional symmetry restrictions, this condition implies constancy of the conditional mean and median (as well as the conditional mode), so estimators which are consistent under these weaker restrictions are equally applicable here. In fact, for models which are invertible in the errors ($\varepsilon \equiv e(y, x, \alpha_0)$ for some $e(\cdot)$), a large

class of GMM estimators is available, based upon the general moment condition

$$E\{d(x)[l(e(y,x,\alpha_0))-v_0]\}=0 \tag{2.30}$$

for any conformable functions $d(\cdot)$ and $l(\cdot)$ for which the moment in (2.30) is well-defined, with $v_0 \equiv E[l(\varepsilon)]$. (MaCurdy (1982) and Newey (1988a) discuss how to exploit these restrictions to obtain more efficient estimators of linear regression coefficients.) Independence restrictions are also stronger than the index and exclusion restrictions to be discussed in the next section, so estimation approaches based upon those restrictions will be relevant here.

In addition to estimation approaches based on these weaker implied stochastic restrictions, certain approaches specific to independence restrictions have been proposed. One strategy to estimate the unknown parameters involves maximization of a "feasible" version of the log-likelihood function, in which the unknown distribution function of the errors is replaced by a (preliminary or concomitant) nonparametric estimator. For some structural functions (in particular, discrete response models), the conditional likelihood function for the observable data depends only on the cumulative distribution function $F_\varepsilon(\cdot)$ of the error terms, and not its derivative (density). Since cumulative distribution functions are bounded and satisfy certain monotonicity restrictions, the set of possible c.d.f.'s will be compact with respect to an appropriately chosen topology, so in such cases an estimator of the parameters of interest α_0 can be defined by maximization of the log-likelihood simultaneously over the finite-dimensional parameter α and the infinite-dimensional nuisance parameter $F_\varepsilon(\cdot)$. That is, if $f(y|x,\alpha,F_\varepsilon(\cdot))$ is the conditional density of y given x and the unknown parameters α_0 and F_ε (with respect to a fixed measure μ_y), a nonparametric maximum likelihood (NPML) estimator for the parameters can be defined as

$$\begin{pmatrix}\hat{\alpha}\\ \hat{F}(\cdot)\end{pmatrix}=\underset{\alpha\in\Theta,F\in\mathcal{F}}{\operatorname{argmax}}\frac{1}{N}\sum_{i=1}^{N}\ln f(y_i|x_i,\alpha,F(\cdot)), \tag{2.31}$$

where $\mathcal{F}$ is the space of admissible c.d.f.'s. Such estimators were proposed by, e.g. Cosslett (1983) for the binary response model and Heckman and Singer (1984) for a duration model with unobserved heterogeneity. Consistency of $\hat{\alpha}$ can be established by verification of the Kiefer and Wolfowitz (1956) conditions for consistency of NPML estimation; however, an asymptotic distribution theory for such estimators has not yet been developed, so the form of the influence function for $\hat{\alpha}$ (if it exists) has not yet been rigorously established.

When the likelihood function of the dependent variable y depends, at least for some observations, on the density function $f_\varepsilon(e) \equiv \mathrm{d}F_\varepsilon(e)/\mathrm{d}e$ of the error terms, the joint maximization problem given in (2.31) can be ill-posed: spurious maxima (at infinity) can be obtained by sending the (unbounded) density estimator $\hat{f}_\varepsilon$ to infinity at particular points (depending on α and the data). In such cases, nonparametric density estimation techniques are sometimes used to obtain a preliminary estimator

$\hat{f}_\varepsilon$ (possibly depending on α) which is substituted into the likelihood function, yielding the "profile likelihood" criterion considered by Severini and Wong (1987a), Andrews (1990a, b) and Newey (1991). More generally, the profile likelihood approach might be used whenever the likelihood function depends upon an infinite-dimensional nuisance parameter $\delta_0(\cdot)$; given a preliminary estimator $\tilde{\delta}(\cdot)$ of this nuisance parameter, an estimator of α_0 can be defined as

$$\hat{\alpha} = \underset{\alpha \in \Theta}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \ln f(y_i | x_i, \alpha, \tilde{\delta}(\cdot)), \tag{2.32}$$

which yields a first-order condition of the "semiparametric M-estimator" form discussed above. The nonparametric estimator $\tilde{\delta}(\cdot)$ typically varies with α – for example, a density estimator might be based on residuals $e_i \equiv e(y_i, x_i, \alpha)$ – which complicates derivation of the asymptotic theory, requiring particular rates of convergence of $\tilde{\delta}(\cdot)$ to $\delta_0(\cdot)$ which are uniform over α, as discussed in Andrews' and Newey and McFadden's chapters in this volume. Though the profile likelihood approach is, in principle, applicable to models which do not impose independence of the errors and regressors, it is more attractive under this restriction because the unknown nuisance parameter (the cumulative and/or density of the errors) is relatively low-dimensional, depending only on the dimension of ε and not also on the number of regressors x. Such estimators have been proposed for most of the limited dependent variable models considered below.

A variant of profile likelihood estimation is based upon squared-error minimization rather than likelihood maximization. From the structural model $y = g(x, \alpha, \varepsilon)$, it is usually possible to deduce the form of the conditional mean of y,

$$E[y|x] = E[g(x, \alpha, \varepsilon)|x] \equiv \gamma(x, \alpha, f_\varepsilon(\cdot)). \tag{2.33}$$

Given this expectation function and a nonparametric estimator $\tilde{f}(\cdot)$ of $f_\varepsilon(\cdot)$, a semiparametric analogue of a nonlinear (weighted) least squares estimator of α_0 is

$$\hat{\alpha} = \underset{\alpha \in \Theta}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} w(x_i)[y_i - \gamma(x_i, \alpha, \tilde{f}(\cdot))]^2, \tag{2.34}$$

where the weights $w(x)$ might also depend upon preliminary (parametric or nonparametric) estimators. In general, such estimators will be less efficient than their profile likelihood counterparts (just as least squares is generally less efficient than maximum likelihood), but in some circumstances may achieve the semiparametric efficiency bound. Semiparametric least squares estimators have also been proposed for many limited dependent variable models.

Another method for construction of estimators under independence restrictions, the "pairwise comparison" approach, uses the fact that differences of independent

and identically distributed random variables are symmetrically distributed about zero. For a particular structural model $y = g(x, \alpha, \varepsilon)$, the first step in the construction of a pairwise difference estimator is to find some transformation $e(z_i, z_j, \alpha) \equiv e_{ij}(\alpha)$ of pairs of observations $(z_i, z_j) \equiv ((y_i, x_i), (y_j, x_j))$ and the parameter vector so that, conditional on the regressors x_i and x_j, the transformations $e_{ij}(\alpha_0)$ and $e_{ji}(\alpha_0)$ are identically distributed, i.e.

$$\mathscr{L}(e_{ij}(\alpha_0)|x_i, x_j) = \mathscr{L}(e_{ji}(\alpha_0)|x_i, x_j) \quad \text{a.s.}, \tag{2.35}$$

where $\mathscr{L}(\cdot|\cdot)$ denotes the conditional sampling distribution of the random variable. In order for the parameter α_0 to be identified using this transformation, it must also be true that $\mathscr{L}(e_{ij}(\alpha_1)|x_i, x_j) \neq \mathscr{L}(e_{ji}(\alpha_1)|x_i, x_j)$ with positive probability if $\alpha_1 \neq \alpha_0$, which implies that observations i and j cannot enter symmetrically in the function $e(z_i, z_j, \alpha)$. Since ε_i and ε_j are assumed to be mutually independent given x_i and x_j, $e_{ij}(\alpha)$ and $e_{ji}(\alpha)$ will be conditionally independent given x_i and x_j; thus, if (2.35) is satisfied, then the difference $e_{ij}(\alpha) - e_{ji}(\alpha)$ will be symmetrically distributed about zero, conditionally on x_i and x_j, when evaluated at $\alpha = \alpha_0$. Given an odd function $\xi(\cdot)$ (which, in general, might depend on x_i and x_j), the conditional symmetry of $e_{ij}(\alpha) - e_{ji}(\alpha)$ implies the conditional moment restriction

$$E[\xi(e_{ij}(\alpha_0) - e_{ji}(\alpha_0))|x_i, x_j] = 0 \quad \text{a.s.}, \tag{2.36}$$

provided this expectation exists, and α_0 will be identified using this restriction if it fails to hold when $\alpha \neq \alpha_0$. When $\xi(\cdot)$ is taken to be the identity mapping $\xi(d) \equiv d$, the restriction that $e_{ij}(\alpha_0)$ and $e_{ji}(\alpha_0)$ have identical conditional distributions can be weakened to the restriction that they have identical conditional means,

$$E[e_{ij}(\alpha_0)|x_i, x_j] = E[e_{ji}(\alpha_0)|x_i, x_j] \quad \text{a.s.}, \tag{2.37}$$

which may not require independence of the errors ε_i and regressors x_i, depending on the form of the transformation $e(\cdot)$.

Given an appropriate (integrable) vector $l(x_i, x_j, \alpha)$ of functions of the regressors and parameter vector, this yields the unconditional moment restrictions

$$E[\xi(e_{ij}(\alpha_0) - e_{ji}(\alpha_0))l(x_i, x_j, \alpha_0)] = 0, \tag{2.38}$$

which can be used as a basis for estimation. If $l(\cdot)$ is chosen to have the same dimension as α, a method-of-moments estimator $\hat{\alpha}$ of α_0 can be defined as the solution to the sample analogue of this population moment condition, namely,

$$\binom{n}{2}^{-1} \sum_{i<j} \xi(e_{ij}(\hat{\alpha}) - e_{ji}(\hat{\alpha}))l(x_i, x_j, \hat{\alpha}) = 0 \tag{2.39}$$

(which may only approximately hold if $\xi(e_{ij}(\alpha) - e_{ji}(\alpha))$ is discontinuous in α). For many models (e.g. those depending on a latent variable $y_i^* \equiv g(x_i, \alpha) + \varepsilon_i$), it is possible to construct some minimization problem which has this sample moment condition as a first-order condition, i.e. for some function $s(z_i, z_j, \alpha)$ with

$$\frac{\partial s(z_i, z_j, \alpha)}{\partial \alpha} = \xi(e_{ij}(\alpha) - e_{ji}(\alpha))l(x_i, x_j, \alpha),$$

the estimator $\hat{\alpha}$ might alternatively be defined as

$$\hat{\alpha} = \underset{\alpha \in \Theta}{\operatorname{argmin}} \binom{n}{2}^{-1} \sum_{i<j} s(z_i, z_j, \alpha). \tag{2.40}$$

A simple example of a model which is amenable to the pairwise differencing approach is the linear model, $y_i = x_i'\beta_0 + \varepsilon_i$, where ε_i and x_i are assumed to be independent. For this case, one transformation function which satisfies the requirements above is

$$e(y_i, x_i, x_j, \alpha) \equiv y_i - x_i'\beta,$$

which does not depend on x_j. Choosing $l(x_i, x_j, \alpha) = x_i - x_j$, a pairwise difference estimator of β_0 can be defined to solve

$$\binom{n}{2}^{-1} \sum_{i<j} \xi((y_i - y_j) - (x_i - x_j)'\hat{\beta})(x_i - x_j) \cong 0,$$

or, if Ξ is the antiderivative of ξ, to minimize

$$S_n(\beta) = \binom{n}{2}^{-1} \sum_{i<j} \Xi((y_i - y_j) - (x_i - x_j)'\beta).$$

When $\xi(d) = d$, the estimator $\hat{\beta}$ is algebraically equal to the slope coefficient estimators of a classical least squares regression of y_i on x_i and a constant (unless some normalization on the location of the distribution of ε_i is imposed, a constant term is not identified by the independence restriction). When $\xi(d) = \operatorname{sgn}(d)$, $\hat{\beta}$ is a rank regression estimator which sets the sample covariance of the regressors x_i with the ranks of the residuals $y_i - x_i'\hat{\beta}$ equal (approximately) to zero (Jurečková (1971), Jaeckel (1972)). The same general approach has been used to construct estimators for discrete response models and censored and truncated regression models.

In all of these cases, the pairwise difference estimator $\hat{\alpha}$ is defined as a minimizer of a second-order U-statistic of the form

$$U_n(\alpha) \equiv \binom{n}{2}^{-1} \sum_{i<j} p(z_i, z_j, \alpha)$$

(with $z_i \equiv (y_i, x_i)$), and will solve an approximate first-order condition

$$\binom{n}{2}^{-1} \sum_{i<j} q(z_i, z_j, \hat{\alpha}) = o_p(n^{-1/2}),$$

where $q(\cdot) = \partial p(\cdot)/\partial \alpha$ when this derivative is well-defined. As described in Section 1.4 above, the asymptotic normal distribution of the estimator $\hat{\alpha}$ can be derived from the asymptotically linear representation

$$\hat{\alpha} = \alpha_0 - \frac{m}{n} \sum_{i=1}^{n} H_0^{-1} r(z_i, \alpha_0) + o_p(n^{-1/2}), \tag{2.41}$$

where $r(z_i, \alpha) \equiv E[q(z_i, z_j, \alpha)|z_i]$ and

$$H_0 \equiv \frac{\partial E[r(z_i, \alpha_0)]}{\partial \alpha'}.$$

The pairwise comparison approach is also useful for construction of estimators for certain nonlinear panel data models. In this setting functions of pairs of observations are constructed, not across individuals, but over time for each individual. In the simplest case, where only two observations across time are available for each individual, a moment condition analogous to (2.36) is

$$E[\xi(e_{12,i}(\alpha_0) - e_{21,i}(\alpha_0))|x_{i1}, x_{i2}] = 0 \quad \text{a.s.}, \tag{2.42}$$

where now $e_{12,i}(\alpha) \equiv e(z_{i1}, z_{i2}, \alpha)$ for the same types of transformation functions $e(\cdot)$ described above, and where the second subscripts on the random variables denote the respective time periods. To obtain the restriction (2.42), it is not necessary for the error terms $\varepsilon_i \equiv (\varepsilon_{i1}, \varepsilon_{i2})$ to be independent of the regressors $x_i = (x_{i1}, x_{i2})$ across individuals i; it suffices that the components ε_{i1} and ε_{i2} are mutually independent and identically distributed across time, given the regressors x_i. The pairwise differencing approach, when it is applicable to panel data, has the added advantage that it automatically adjusts for the presence of individual-specific fixed effects, since $\varepsilon_{i1} + \gamma_i$ and $\varepsilon_{i2} + \gamma_i$ will be identically distributed if ε_{i1} and ε_{i2} are. A familiar example is the estimation of the coefficients β_0 in the linear fixed-effects model

$$y_{it} = x_{it}'\beta_0 + \gamma_i + \varepsilon_{it}, \qquad t = 1, 2,$$

where setting the transformation $e_{12,i}(\alpha) \equiv y_{i1} - x_{i1}'\beta$ and $\xi(u) \equiv u$ in (2.42) results in the moment condition

$$E[(y_{i1} - y_{i2}) - (x_{i1} - x_{i2})'\beta_0 | x_{i1}, x_{i2}] = E[\varepsilon_{i1} - \varepsilon_{i2}|x_{i1}, x_{i2}] = 0,$$

which is the basis for the traditional least squares fixed effects estimator. As described in Section 3.5 below, this idea has been exploited to construct estimators for panel data versions of the binary response and censored and truncated regression models which are semiparametric with respect to both the error distribution and the distribution of the fixed effects.

2.5. *Exclusion and index restrictions*

Construction of estimators based on index restrictions can be based on a variety of different approaches, depending upon whether the index function $v(x)$ is completely known or depends upon (finite- or infinite-dimensional) unknown parameters, and whether the index sufficiency condition is of the "weak" (affecting only the conditional mean or median) or "strong" (applying to the entire error distribution) form. Estimators of the parameters of interest under mean index restrictions exploit modified forms of the moment conditions implied by the stronger constant conditional mean restrictions, just as estimators under distributional index restrictions use modifications of estimation strategies for independence restrictions.

Perhaps the simplest version of the restrictions to analyze are mean exclusion restrictions, for which the index function is a subset of the regressors (i.e. $v(x) \equiv x_1$, where $x \equiv (x_1', x_2')'$), so that the restriction is

$$E[\varepsilon|x] = E[\varepsilon|x_1] \quad \text{a.s.} \tag{2.43}$$

As for conditional mean restrictions, this condition can be used to identify the parameters of interest, α_0, for structural functions $y = g(x, \alpha_0, \varepsilon)$ which are invertible in the error terms ($\varepsilon = e(y, x, \alpha_0)$), so that the exclusion restriction (2.43) can be rewritten as

$$E[e(y, x, \alpha_0)|x] - E[e(y, x, \alpha_0)|x_1] = 0. \tag{2.44}$$

By iterated expectations, this implies an unconditional moment restriction which is analogous to condition (2.4) of Section 2.1, namely,

$$0 = E[\tilde{d}(x)\, e(y, x, \alpha_0)], \tag{2.45}$$

where now

$$\tilde{d}(x) \equiv d(x) - E[d(x)|x_1]\{E[A(x)|x_1]\}^{-1} A(x) \tag{2.46}$$

for any conformable matrix $d(x)$ and square matrix $A(x)$ of functions of the regressors for which the relevant expectations and inverses exist. (Note that, by construction, $E[\tilde{d}(x)|x_1] = 0$ almost surely.) Alternatively, estimation might be based on the

condition

$$0 = E[\tilde{d}(x)\,\tilde{e}(y, x, \alpha_0)], \tag{2.47}$$

where, analogously to (2.46),

$$\tilde{e}(y, x, \alpha) \equiv e(y, x, \alpha) - E[e(y, x, \alpha)|x_1]\,\{E[A(x)|x_1]\}^{-1} A(x).$$

Given a particular nonparametric method for estimation of conditional means given x_1 (denoted $\hat{E}[\cdot|x_1]$), a semiparametric M-estimator $\hat{\alpha}$ of the structural coefficients α_0 can be defined as the solution to a sample analogue of (2.45),

$$0 = \frac{1}{N}\sum_{i=1}^{N} \{d(x_i, \hat{\alpha}, \hat{\delta}) - \hat{E}[d(x_i, \hat{\alpha}, \hat{\delta})|x_{i1}](\hat{E}[A(x_i)|x_{i1}])^{-1} A(x_i)\}\, e(y_i, x_i, \hat{\alpha}), \tag{2.48}$$

where the instrumental variable matrix $d(x)$ is permitted to depend upon α and a preliminary nuisance parameter estimator $\hat{\delta}$, as in Section 2.2. Formally, the asymptotic distribution of this estimator is given by the same expression (2.7) for estimation with conditional mean restrictions, replacing d with $\tilde{d}$ throughout. However, rigorous verification of the consistency and asymptotic normality of $\hat{\alpha}$ is technically difficult, and the estimating equation (2.48) must often be modified to "trim" (i.e. delete) observations where the nonparametric regression estimator $\hat{E}[\cdot]$ is imprecise. A bound on the attainable efficiency of estimators of α_0 under condition (2.44) was derived by Chamberlain (1992), who showed that an optimal instrumental variable matrix $\tilde{d}^*(x)$ of the form (2.46) is related to the corresponding optimal instrument matrix $d^*(x)$ for the constant conditional moment restrictions of Section 2.2 by the formula

$$\tilde{d}^*(x) = d^*(x) - E[d^*(x)|x_1]\,[E\{[\Sigma(x)]^{-1}|x_1\}]^{-1}\,[\Sigma(x)]^{-1}, \tag{2.49}$$

where $d^*(x)$ is defined in (2.8) above and $\Sigma(x)$ is the conditional covariance matrix of the errors ε given the regressors x. This formula directly generalizes to the case in which the subvector x_1 is replaced by a more general (but known) index function $v(x)$.

For a linear model $y = x_2'\beta_0 + \varepsilon$, the mean exclusion restriction (2.43) yields the semilinear model considered by Robinson (1988):

$$y = x_2'\beta_0 + \theta(x_1) + \eta,$$

where $\theta(x_1) \equiv E[\varepsilon|x_1]$ and $E[\eta|x] \equiv E[\varepsilon - \theta(x_1)|x] = 0$. Defining $e(y, x, \alpha) \equiv y - x_2'\beta$, $d(x) \equiv x_2$, and $A \equiv I$, the moment condition (2.47) becomes

$$E[\{x_2 - E[x_2|x_1]\}\,\{y - E[y|x_1] - (x_2 - E[x_2|x_1])'\beta_0\}] = 0,$$

which can be solved for β_0:

$$\beta_0 = \{E[(x_2 - E[x_2|x_1])(x_2 - E[x_2|x_1])]\}^{-1} E\{(x_2 - E[x_2|x_1])(y - E[y|x_1])\}.$$

Robinson (1988) proposed an estimator of β_0 constructed from a sample analogue to (2.47), using kernel regression to nonparametrically estimate the conditional expectations and "trimming" observations where a nonparametric estimator of the density of x_1 (assumed continuously distributed) is close to zero and gave conditions under which the resulting estimator was root-N-consistent and asymptotically normal. Linton (1992) constructs higher-order approximations to the distribution of this estimator.

Strengthening the mean exclusion restriction to a distributional exclusion condition widens the class of moment restrictions which can be exploited when the structural function is invertible in the errors. Imposing

$$\Pr\{\varepsilon \leqslant u|x\} = \Pr\{\varepsilon \leqslant u|x_1\} \tag{2.50}$$

for all possible values of u yields the general moment conditions

$$0 = E[\tilde{d}(x)\, l(e(y, x, \alpha_0))] \tag{2.51}$$

for any square-integrable function $l(\varepsilon)$ of the errors, which includes (2.45) as a special case. As with independence restrictions, precision of estimators of α_0 can be improved by judicious choice of the transformation $l(\cdot)$.

Even for noninvertible structural functions, the pairwise comparison approach considered for index restrictions can be modified to be applicable for distributional exclusion (or known index) restrictions. For any pair of observations z_i and z_j which have the same value of the index function $v(x_i) = v(x_j)$, the corresponding error terms ε_i and ε_j will be independently and identically distributed, given the regressors x_i and x_j, under the distributional index restriction

$$\Pr\{\varepsilon \leqslant u|x\} = \Pr\{\varepsilon \leqslant u|v(x)\}. \tag{2.52}$$

Given the pairwise transformation function $e(z_i, z_j, \alpha) \equiv e_{ij}(\alpha)$ described in the previous section, an analogue to restriction (2.35) holds under this additional restriction of equality of index functions:

$$\mathscr{L}(e_{ij}(\alpha_0)|x_i, x_j) = \mathscr{L}(e_{ji}(\alpha_0)|x_i, x_j) \quad \text{a.s. if} \quad v(x_i) = v(x_j). \tag{2.53}$$

As for independence restrictions, (2.53) implies the weaker conditional mean restriction

$$E[e_{ij}(\alpha_0)|x_i, x_j] = E[e_{ji}(\alpha_0)|x_i, x_j] \quad \text{a.s. if} \quad v(x_i) = v(x_j), \tag{2.54}$$

which is relevant for invertible structural functions (with $e_{ij}(\alpha)$ equated with the inverse function $e(y_i, x_i, \alpha)$ in this case).

These restrictions suggest estimation of α_0 by modifying the estimating equation (2.39) or the minimization problem (2.40) of the preceding subsection to exclude pairs of observations for which $v(x_i) \neq v(x_j)$. However, in general $v(x_i) - v(x_j)$ may be continuously distributed around zero, so direct imposition of this restriction would exclude all pairs of observations. Still, if the sampling distributions $\mathcal{L}(e_{ij}(\alpha_0) | x_i, x_j, v(x_i) - v(x_j) = c)$ or conditional expectations $E[e_{ij}(\alpha_0) | x_i, x_j, v(x_i) - v(x_j) = c]$ are smooth functions of c at $c = 0$, the restrictions (2.53) or (2.54) will approximately hold if $v(x_i) - v(x_j)$ is close to zero. Then appropriate modifications of the estimating equations (2.39) and minimization problem (2.40) are

$$\binom{N}{2}^{-1} \sum_{i<j} \xi(e_{ij}(\hat{\alpha}) - e_{ji}(\hat{\alpha}))\, l(x_i, x_j, \hat{\alpha})\, w_N(v(x_i) - v(x_j)) = 0 \tag{2.55}$$

and

$$\hat{\alpha} = \underset{\theta \in \Theta}{\operatorname{argmin}} \binom{N}{2}^{-1} \sum_{i<j} s(z_i, z_j, \alpha)\, w_N(v(x_i) - v(x_j)), \tag{2.56}$$

for some weighting function $w_N(\cdot)$ which tends to zero as the magnitude of its argument increases and, at a faster rate, as the sample size N increases (so that, ultimately, only observations with $v(x_i) - v(x_j)$ very close to zero are included in the summations).

Returning to the semilinear regression model $y = x_2'\beta_0 + \theta(x_1) + \eta$, $E[\eta | x] = 0$, the same transformation as used in the previous subsection can be used to construct a pairwise difference, provided the nonparametric components $\theta(x_{i1})$ and $\theta(x_{j1})$ are equal for the two observations; that is, if $e(y_i, x_i, x_j, \alpha) \equiv e_{ij}(\alpha) = y_i - x_{i2}'\beta$ and $v(x_i) \equiv x_{i1}$, then

$$e_{ij}(\alpha_0) - e_{ji}(\alpha_0) = (\varepsilon_i - \varepsilon_j) = (\eta_i - \eta_j)$$

if $v(x_i) = v(x_j)$. Provided $\theta(x_{i1})$ is a smooth (continuous and differentiable) function, relation (2.36) will hold approximately if $x_{i1} \cong x_{j1}$. Defining the weight function $w_N(\cdot)$ to be a traditional kernel weight,

$$w_N(d) = k(h_N^{-1} d), \qquad k(0) > 0, k(\lambda) \to 0 \text{ as } \|\lambda\| \to \infty, h_N \to 0 \text{ as } N \to \infty, \tag{2.57}$$

and taking $l(x_i, x_j, \alpha) = x_{i2} - x_{j2}$ and $\xi(d) = d$, a pairwise difference estimator of β_0 using either (2.55) or (2.56) reduces to a weighted least squares regression of the distinct differences $(y_i - y_j)$ in dependent variables on the differences $(x_{i2} - x_{j2})$ in regressors, using $k(h_N^{-1}(x_{i1} - x_{j1}))$ as weights (as proposed by Powell (1987)).

Consistency of the resulting estimator $\hat{\beta}$ requires only the weak exclusion restriction (2.43); when the strong exclusion restriction (2.53) is imposed, other choices of odd function $\xi(d)$ besides the identity function are permissible in (2.55). Thus, an estimator of β_0 using $\xi(d) = \text{sgn}(d)$ might solve

$$\binom{N}{2}^{-1} \sum_{i<j} \text{sgn}((y_i - y_j) - (x_{i1} - x_{j1})'\hat{\beta})(x_{i1} - x_{j1})\, k((x_{i2} - x_{j2})/h_N) \cong 0. \tag{2.58}$$

This is the first-order condition of a "smoothed" version of the minimization problem defining the rank regression estimator,

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \binom{N}{2}^{-1} \sum_{i<j} |(y_i - y_j) - (x_i - x_j)'\beta| k((x_{i2} - x_{j2})/h_N), \tag{2.59}$$

which is a "robust" alternative to estimators proposed by Robinson (1988b) and Powell (1987) for the semilinear model. Although the asymptotic theory for such estimators has yet to be developed, it is likely that reasonable conditions can be found to ensure their root-N-consistency and asymptotic normality.

So far, the discussion has been limited to models with known index functions $v(x)$. When the index function depends upon unknown parameters δ_0 which are functionally unrelated to the parameters of interest α_0, and when preliminary consistent estimators $\hat{\delta}$ of δ_0 are available, the estimators described above are easily adapted to use an estimated index function $\hat{v}(x) = v(x, \hat{\delta})$. The asymptotic distribution theory for the resulting estimator must properly account for the variability of the preliminary estimator $\hat{\delta}$. When δ_0 is related to α_0, and that relation is exploited in the construction of an estimator of α_0, the foregoing estimation theory requires more substantial modification, both conceptually and technically.

A leading special case occurs when the index governing the conditional error distribution appears in the same form in the structural function for the dependent variable y. For example, suppose the structural function has a linear latent variable form,

$$y = g(x, \alpha_0, \varepsilon) = t(x'\beta_0 + \varepsilon), \tag{2.60}$$

and index $v(x)$ is the latent linear regression function $x'\beta_0$,

$$\Pr\{\varepsilon \leqslant u | x\} = \Pr\{\varepsilon \leqslant u | x'\beta_0\}. \tag{2.61}$$

This particular index restriction on the unobservable error terms immediately implies the same index restriction for the observable dependent variable,

$$\Pr\{y \leqslant u | x\} = \Pr\{y \leqslant u | x'\beta_0\}, \tag{2.62}$$

which can be used to generate moment restrictions for estimation of β_0. For example, (2.62) implies the weaker restriction

$$E[y|x] = G(x'\beta_0), \tag{2.63}$$

on the conditional mean of the dependent variable, where $G(\cdot)$ is some unknown nuisance function. (Clearly β_0 is at most identified up to a location and scale normalization without stronger restrictions on the form of $G(\cdot)$.) Defining $\tilde{e}(y, x, b) \equiv y - E[y|x'b]$, condition (2.63) implies that

$$E[d(x)\,\tilde{e}(y, x, \beta_0)] = 0 \tag{2.64}$$

for any conformable, square-integrable $d(x)$. Thus, with a nonparametric estimator $\hat{E}[y|x'b]$ of the conditional expectation function $E[y|x'b]$, a semiparametric M-estimator of β_0 can be constructed as a sample analogue to (2.64). Alternatively, a weighted pairwise difference approach might be used: assuming $G(\cdot)$ is continuous, the difference in the conditional means of the dependent variables for observations i and j satisfies

$$E[y_i - y_j|x_i, x_j] = G(x_i'\beta_0) - G(x_j'\beta_0) \cong 0 \quad \text{if} \quad x_i'\beta_0 \cong x_j'\beta_0. \tag{2.65}$$

So by estimating $E[y_i - y_j|x_i, x_j]$ nonparametrically and determining when it is near zero, the corresponding pair of observations will have $(x_i - x_j)'\beta_0 \cong 0$, which is useful in determining β_0. When $G(\cdot)$ is known to be monotonic (which follows, for example, if the transformation $t(\cdot)$ of the latent variable in (2.60) is monotonic and ε is assumed to be independent of x), a variation on the pairwise comparison approach could exploit the resulting inequality $E[y_i - y_j|x_i, x_j] = G(x_i'\beta_0) - G(x_j'\beta_0) > 0$ only if $x_i'\beta_0 > x_j'\beta_0$.

Various estimators based upon these conditions have been proposed for the monotone regression model, as discussed in Section 3.2 below. More complicated examples involve multiple indices, with some indices depending upon parameters of interest and others depending upon unrelated nuisance parameters, as for some of the proposed estimators for selectivity bias models. The methods of estimation of the structural parameters α_0 vary across the particular models but generally involve nonparametric estimation of regression or density functions involving the index $v(x)$.

3. Structural models

3.1. *Discrete response models*

The parameters of the binary response model

$$y = 1(x'\beta_0 + \varepsilon > 0) \tag{3.1}$$

are traditionally estimated by maximization of the average log-likelihood function

$$\mathscr{L}_N(\beta; F) = \frac{1}{N}\sum_{i=1}^{N} (y_i \ln[F(x_i'\beta)] + (1 - y_i)\ln[1 - F(x_i'\beta)]), \tag{3.2}$$

where the error term ε is assumed to be distributed independently of x with known distribution function $F(\cdot)$ (typically standard normal or logistic). Estimators for semiparametric versions of the binary response model usually involve maximization of a modified form of this log-likelihood, one which does not presuppose knowledge of the distribution of the errors. For the more general multinomial response model, in which J indicator variables $\{y_j, j = 1, \ldots, J\}$ are generated as

$$y_j = 1\{x'\beta_0^j + \varepsilon_j > x'\beta_0^k + \varepsilon_k \quad \text{for all} \quad k \neq j\}, \tag{3.3}$$

the average log-likelihood has the analogous form

$$\mathscr{L}_N(\beta^1, \ldots, \beta^J; F) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J} y_{ij} \ln[F_j(x_i'\beta^1, \ldots, x_i'\beta^J)], \tag{3.4}$$

where $F_j(\cdot)$ is the conditional probability that $y_j = 1$ given the regressors x. This form easily specializes to the ordered response or grouped dependent variable models, replacing $F_j(\cdot)$ with $F(x_i'\beta_0 - c_j) - F(x_i'\beta_0 - c_{j-1})$, where the $\{c_j\}$ are the (known or unknown) group boundaries.

The earliest example of a semiparametric approach for estimation of a limited dependent variable model in econometrics is the *maximum score* estimation method proposed by Manski (1975). For the binary response mode, Manski suggested that β_0 be estimated by maximizing the number of correct predictions of y by the sign of the latent regression function $x'\beta$; that is, $\hat{\beta}$ was defined to maximize the predictive score function

$$S_n(\beta) \equiv \sum_{i=1}^{N} (y_i 1\{x_i'\beta > 0\} + (1 - y_i) 1\{x_i'\beta \leqslant 0\}) \tag{3.5}$$

over a suitable parameter space Θ (e.g. the unit sphere). The error terms ε were restricted to have conditional median zero to ensure consistency of the estimator. A later interpretation of the estimator (Manski (1985)) characterized the maximum score estimator $\hat{\beta}$ as a least absolute deviations estimator, since the estimator solved the minimization problem

$$\hat{\beta} = \underset{\Theta}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} |y_i - 1\{x_i'\beta > 0\}|. \tag{3.6}$$

This led to the extension of the maximum score idea to more general quantile estimation of β_0, under the assumption that the corresponding conditional quantile of the error terms was constant (Manski (1985)). The maximum score approach was also applied to the multinomial response model by Manski (1975); in this case, the score criterion becomes

$$S_n(\beta_1, \ldots, \beta_J) = \sum_{i=1}^{N} \sum_{j=1}^{J} y_{ij} 1\{x_i'\beta^j > x_i'\beta^k, \quad k \neq j\}, \tag{3.7}$$

and its consistency was established under the stronger condition of mutual independence of the alternative specific errors $\{\varepsilon_j\}$. M. Lee (1992) used conditional median restrictions to define a least absolute deviations estimator of the parameters of the ordered response model along the same lines.

Although consistency of the maximum score estimator for binary response was rigorously established by Manski (1985) and Amemiya (1985), its asymptotic distribution cannot be established by the methods described in Section 2.2 above, because of lack of continuity of the median regression function $1\{x'\beta_0 > 0\}$ of the dependent variable y. More importantly, because this median regression function is flat except at its discontinuity points, the estimator is not root-N-consistent under standard regularity conditions on the errors and regressors. Kim and Pollard (1990) found that the rate of convergence of the maximum score estimator to β_0 under such conditions is $N^{1/3}$, with a nonstandard asymptotic distribution (involving the distribution of the maximum value of a particular Gaussian process with quadratic drift). This result was confirmed for finite samples by the simulation study of Manski and Thompson (1986).

Chamberlain (1986) showed that this slow rate of convergence of the maximum score estimator was not particular to the estimation method, but a general consequence of estimation of the binary response model with a conditional median restriction. Chamberlain showed that the semiparametric version of the information matrix for this model is identically zero, so that no regular root-N-consistent estimator of β_0 exists in this case. An extension by Zheng (1992) derived the same result – a zero semiparametric information matrix – even if the conditional median restriction is strengthened to an assumption of conditional symmetry of the error distribution. Still, consistency of the maximum score estimator $\hat{\beta}$ illustrates the fact that the parameters β_0 of the binary response model are identified under conditional quantile or symmetry assumptions on the error terms, which is not the case if the errors are restricted only to have constant conditional mean.

If additional smoothness restrictions on the distribution of the errors and regressors are imposed, the maximum score (quantile) approach can be modified to obtain estimators which converge to the true parameters at a faster rate than $N^{1/3}$. Nawata (1992) proposed an estimator which, in essence, estimates β_0 by maximizing the fit of an estimator of the conditional median function $1(x'\beta_0 > 0)$ of the binary variable to a nonparametric estimator of the conditional median of y given x. In a

first stage, the observations are grouped by a partition of the space of regressors, and the median value of the dependent variable y is calculated for each of these regressor bins. These group medians, along with the average value of the regression vector in each group, are treated as raw data in a second-stage fit of the binary response model using the likelihood function (3.2) with a standard normal cumulative and a correction for heteroskedasticity induced by the grouping scheme. Nawata (1992) gives conditions under which the rate of convergence of the resulting estimator is $N^{2/5}$, and indicates how the estimator and regularity conditions can be modified to achieve a rate of convergence arbitrarily close to $N^{1/2}$. Horowitz (1992) used a different approach, but similar strengthening of the regularity conditions, to obtain a median estimator for binary response with a faster convergence rate. Horowitz modifies the score function of (3.5) by replacing the conditional median function $1\{x'\beta > 0\}$ by a "smoothed" version, so that an estimator of β_0 is defined as a minimizer of the criterion

$$S_n^*(\beta) \equiv \sum_{i=1}^{N} y_i K(x_i'\beta/h_N) + (1 - y_i)[1 - K(x_i'\beta/h_N)], \tag{3.8}$$

where $K(\cdot)$ is a smooth function in $[0, 1]$ with $K(u) \to 0$ or 1 as $u \to -\infty$ or ∞, and h_N is a sequence of bandwidths which tends to zero as the sample size increases (so that $K(x'\beta_0/h_N)$ approaches the binary median $1(x'\beta_0 > 0)$ as $N \to \infty$). With particular conditions on the function $K(\cdot)$ and the smoothness of the regressor distribution and with the conditional density of the errors at the median being zero, Horowitz (1992) shows how the rate of convergence of the minimizer of $S_N^*(\beta)$ over Θ can be made at least $N^{2/5}$ and arbitrarily close to $N^{1/2}$; moreover, asymptotic normality of the resulting estimator is shown (and consistent estimators of asymptotic bias and covariance terms are provided), so that normal sampling theory can be used to construct confidence regions and hypothesis tests in large samples.

When the error terms in the binary response model are assumed to satisfy the stronger assumption of independence of the errors and regressors, Cosslett (1987) showed that the semiparametric information matrix for estimation of β_0 in (3.1) (once a suitable normalization is imposed) is generally nonsingular, a necessary condition for existence of a regular root-N-consistent estimator. Its form is analogous to the parametric information matrix when the distribution function $F(\cdot)$ of the errors is known, except that the regressors x are replaced by deviations from their conditional means given the latent regression function $x'\beta_0$; that is, the best attainable asymptotic covariance matrix for a regular estimator of β_0 when ε is independent of x with unknown distribution function $F(\cdot)$ is

$$V^* \equiv \left\{ E\left[\frac{[f(x'\beta_0)]^2}{F(x'\beta)[1 - F(x'\beta_0)]} [\tilde{x} - E(\tilde{x}|x'\beta_0)][\tilde{x} - E(\tilde{x}|x'\beta_0)]' \right] \right\}^{-1}, \tag{3.9}$$

where $f(u) = dF(u)/du$ and $\tilde{x}$ is the subvector of regressors x which eliminates the

last component (whose coefficient is assumed normalized to unity to pin down the scale of β_0). Existence of the inverse in (3.9) implies that a constant term is excluded from the regression vector, and the corresponding intercept term is absorbed into the definition of the error cumulative $F(\cdot)$.

For the binary response model under an index restriction, Cosslett (1983) proposed a nonparametric maximum likelihood estimator (NPMLE) of β_0 through maximization of the average log-likelihood function $\mathscr{L}_N(\beta; F)$ simultaneously over $\beta \in \Theta$ and $F \in \mathscr{F}$, where $\mathscr{F}$ is the space of possible cumulative distributions (monotonic functions on $[0, 1]$). Computationally, given a particular trial value b of β, an estimator of F is obtained by monotonic regression of the indicator y on $x'b$, using the *pool adjacent violators* algorithm of isotonic regression; this estimator $\hat{F}$ of F is then substituted into the likelihood function, and the concentrated criterion $\mathscr{L}_N(b; \hat{F})$ is maximized over $b \in \Theta \equiv \{\beta: \|\beta\| = 1\}$. Cosslett (1983) establishes consistency of the resulting estimators of β_0 and $F(\cdot)$ through verification of the Kiefer–Wolfowitz (1956) conditions for the consistency of NPMLE, constructing a topology which ensures compactness of the parameter space $\mathscr{F}$ of possible nuisance functions $F(\cdot)$. As noted in Section 2.4 above, an asymptotic distribution for NMPLE has not yet been established.

Instead of the monotonic regression estimator $\hat{F}(\cdot)$ of $F(\cdot)$ implicit in the construction of the NPMLE, the same estimation approach can be based upon other nonparametric estimators of the error cumulative. The resulting *profile likelihood* estimator of β_0, maximizing $\mathscr{L}_N(b; \tilde{F})$ of (3.2) using a kernel regression estimator $\tilde{F}$, was considered by Severini and Wong (1987a) (for a single parameter) and Klein and Spady (1993). Because kernel regression does not impose monotonicity of the function estimator, this profile likelihood estimator is valid under a weaker index restriction on the error distribution $\Pr\{\varepsilon \leqslant u|x\} = \Pr\{\varepsilon \leqslant u|x'\beta_0\}$, which implies that $E[y|x] = F(x'\beta_0)$ for some (not necessarily monotone) function $F(\cdot)$. Theoretically, the form of the profile likelihood $\mathscr{L}_N(b; \tilde{F})$ is modified by Klein and Spady (1993) to "trim" observations with imprecise estimators of $F(\cdot)$ in order to show root-N-consistency and asymptotic normality of the resulting estimator $\hat{\beta}$. Klein and Spady show that this estimator is asymptotically efficient under the assumption of independence of the errors and regressors, since its asymptotic covariance matrix equals the best attainable value V^* of (3.9) under this restriction.

Other estimators of the parameters of the binary response model have been proposed which do not exploit the particular structure of the binary response model, but instead are based upon general properties of transformation models. If independence of the errors and regressors is assumed, the monotonicity of the structural function (3.1) in ε can be used to define a pairwise comparison estimator of β_0. Imposition of a weaker index restriction $\Pr\{\varepsilon \leqslant u|x\} = \Pr\{\varepsilon \leqslant u|x'\beta_0\}$ implies that

$$E[y|x] = G(x'\beta_0) \tag{3.10}$$

for some unknown function $G(\cdot)$, so any estimator which is based on this restriction

is applicable to the binary response model. A number of estimators proposed for this more general setup are discussed in the following section on transformation models.

Estimation of the multinomial response model (3.3) under independence and index restrictions can be based on natural extensions of the methods for the binary response model. In addition to the maximum score estimator defined by minimizing (3.7), Thompson (1989a, b) considered identification and estimation of the parameters in (3.3) assuming independence of the errors and regressors; Thompson showed how consistent estimators of $(\beta_0^1, \ldots, \beta_0^J)$ could be constructed using a least squares criterion even if only a single element y_j of the vector of choice indicators $(y_1, \ldots, y_j)$ is observed. L. Lee (1991) extended profile likelihood estimation to the multinomial response model, and obtained a similar efficiency result to Klein and Spady's (1993) result for binary response under index restrictions on the error terms. And, as for the binary response model, various pairwise comparison or index restriction estimators for multiple index models are applicable to the multinomial response model; these estimators are reviewed in the next section.

3.2. *Transformation models*

In Section 1.3 above, two general classes of transformation models were distinguished. Parametric transformation models, in which the relation between the latent and observed dependent variables is invertible and of known parametric form, are traditionally estimated assuming the errors are independent of the regressors with density function $f(\cdot;\tau)$ of known parametric form. In this setting, the average conditional log-likelihood function for the dependent variable

$$y = t(x'\beta_0 + \varepsilon; \lambda_0) \Leftrightarrow \varepsilon = t^{-1}(y; \lambda_0) - x'\beta_0 \equiv e(y, x, \beta_0, \lambda_0) \tag{3.11}$$

is

$$\mathscr{L}_N(\beta, \lambda, \tau; f) = \frac{1}{N} \sum_{i=1}^{N} (\ln[f(e(y_i, x_i, \beta, \lambda); \tau)] - \ln[|\partial e(y_i, x_i, \beta, \lambda)/\partial y|]), \tag{3.12}$$

which is maximized over $\theta \equiv (\beta, \lambda, \tau)$ to obtain estimators of the parameters β_0 and λ_0 of interest.

Given both the monotonicity of the transformation $t(\cdot)$ in the latent variable and the explicit representation function $e(\cdot)$ for the errors in terms of the observable variables and unknown parameters, these models are amenable to estimation under most of the semiparametric restrictions on the error distribution discussed in Section 2. For example, Amemiya and Powell (1981) considered nonlinear two-stage least squares (method-of-moments) estimation of β_0 and λ_0 for the Box–Cox

transformation under a conditional mean restriction on the errors ε given the regressors x, and showed how this estimator could greatly outperform (in a mean-squared-error sense) a misspecified Gaussian ML estimator over some ranges of the transformation parameter λ_0. Carroll and Ruppert (1984) and Powell (1991) discuss least absolute deviations and quantile estimators of the Box–Cox regression model, imposing independence or constant quantile restrictions on the errors. Han (1987b) also assumes independence of the errors and regressors, and constructs a pairwise difference estimator of the transformation parameter λ_0 and the slope coefficients β_0 which involves maximization of a fourth-order U-statistic; this approach is a natural generalization of the maximum rank correlation estimation method described below. Newey (1989c) constructs efficient method-of-moments estimators for the Box–Cox regression model under conditional mean, symmetry, and independence restrictions on the error terms. Though not yet considered in the econometric literature, it would be straightforward to extend the general estimation strategies described in Section 2.5 above to estimate the parameters of interest in a semilinear variant of the Box–Cox regression model.

When the form of the transformation function $t(\cdot)$ in (3.11) is not parametrically specified (i.e. the transformation itself is an infinite-dimensional nuisance parameter), estimation of β_0 becomes more problematic, since some of the semiparametric restrictions on the errors no longer suffice to identify β_0 (which is, at most, uniquely determined up to a scale normalization). For instance, since a special case is the binary response model, it is clear from the discussion of the previous section that a conditional mean restriction on ε is insufficient to identify the parameters of interest. Conversely, any dependent variable generated from an unknown (nonconstant and monotonic) transformation can be further transformed to a binary response model, so that identification of the parameters of a binary response model generally implies identification of the parameters of an analogous transformation model.

Under the assumption of independence of the errors and regressors, Han (1987a) proposed a pairwise comparison estimator, termed the *maximum rank correlation estimator*, for the model (3.11) with $t(\cdot)$ unknown but nondecreasing. Han actually considered a generalization of (3.11), the *generalized regression* model, with structural function

$$y = t[s(x'\beta_0, \varepsilon)], \tag{3.13}$$

with $t[\cdot]$ a monotone (but possibly noninvertible) function and $s(\cdot)$ smooth and invertible in both of its arguments; with continuity and unbounded support of the error distribution, this construction ensures that the support of y will not depend upon the unknown parameters β_0. Though the discussion below focusses on the special case $s(x'\beta, \varepsilon) = x'\beta + \varepsilon$, the same arguments apply to this, more general, setup.

For model (3.11), with $t(\cdot)$ unknown and ε and x assumed independent, Han proposed estimation of β_0 by maximization of

$$R_N(\beta) \equiv \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} [1(y_i > y_j)\, 1(x_i'\beta > x_j'\beta) + 1(y_i < y_j)\, 1(x_i'\beta < x_j'\beta)] \tag{3.14}$$

over a suitably-restricted parameter space Θ (e.g. normalizing one of the components of β_0 to unity). Maximization of (3.14) is equivalent to minimization of a least absolute deviations criterion for the sign of $y_i - y_j$ minus its median, the sign of $x_i'\beta - x_j'\beta$, for those observations with nonzero values of $y_i - y_j$:

$$\hat{\beta} \equiv \underset{\Theta}{\operatorname{argmax}}\, R_N(\beta) = \underset{\Theta}{\operatorname{argmin}} \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} 1(y_i \neq y_j) \,|\, 1(y_i > y_j) - 1(x_i'\beta > x_j'\beta)|. \tag{3.15}$$

In terms of the pairwise difference estimators of Section 2.4, defining

$$e_{ij}(\beta) \equiv 1(y_i \neq y_j)\operatorname{sgn}[1(y_i > y_j) - 1(x_i'\beta > x_j'\beta)],$$

identification of β_0 using the maximum rank correlation criterion is related to the conditional symmetry of

$$\begin{aligned} e_{ij}(\beta_0) - e_{ji}(\beta_0) &= 2\, e_{ij}(\beta_0) \\ &= 2\, 1(y_i \neq y_j)\operatorname{sgn}[1((x_i - x_j)'\beta_0 > \varepsilon_j - \varepsilon_i) - 1((x_i - x_j)'\beta_0 > 0)] \end{aligned}$$

about zero given x_i and x_j. The maximum rank correlation estimator defined in (3.15) does not solve a sample moment condition like (2.39) of Section 2.4 (though such estimators could easily be constructed), because the derivative of $R_N(\beta)$ is zero wherever it is well-defined; still, the estimator $\hat{\beta}$ is motivated by the same general pairwise comparison approach described in Section 2.4.

Han (1987a) gave regularity conditions under which $\hat{\beta}$ is consistent for β_0; these included continuity of the error distribution and compact support for the regressors. Under similar conditions Sherman (1993) demonstrated the root-N-consistency and asymptotic normality of the maximum rank estimator; writing the estimator as the minimizer of a second-order U-process,

$$\hat{\beta} \equiv \underset{\Theta}{\operatorname{argmax}} \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p(z_i, z_j, \beta), \tag{3.16}$$

Sherman showed that the asymptotic distribution of $\hat{\beta}$ is the same as that for an M-estimator based on $N/2$ observations which maximizes the sample average of the conditional expectation $r(z_i, \beta) = E[p(z_i, z_j, \beta)|z_i]$ over the parameter space Θ,

and also showed consistency of an estimator of the asymptotic covariance matrix of $\hat{\beta}$ based upon numerical derivatives of sample estimators of this expectation function $r(z, \beta)$. Cavanagh and Sherman (1991) propose a variant of the maximum rank correlation estimator which maximizes

$$Q_N(\beta) \equiv \sum_{j=1}^{N} M(y_i)\mathscr{R}_{Ni}(\beta), \tag{3.17}$$

for $M(\cdot)$ an increasing function and $\mathscr{R}_{Ni}(\beta)$ the rank of $x_i'\beta$ in $\{x_j'\beta, j = 1, \ldots, N\}$, i.e.

$$\mathscr{R}_{Ni}(\beta) \equiv \sum_{j=1}^{N} 1(x_i'\beta \geqslant x_j'\beta). \tag{3.18}$$

They also consider a related estimator based on a criterion which replaces $M(y_i)$ in (3.17) with the rank of y_i in $\{y_j\}$, defined analogously to (3.18). For the binary response model, maximization of any of these criteria is numerically equivalent to maximization of the criterion $R_N(\beta)$ of (3.14), but the estimators differ for non-binary dependent variables. Cavanagh and Sherman (1991) demonstrate root-N-consistency and asymptotic normality of these estimators under relatively weak regularity conditions, and discuss how consistent asymptotic covariance matrix estimators can be obtained.

When the transformation function $t(\cdot)$ of (3.11) is not known to be monotonic, or when the errors are assumed to satisfy only the weaker index restriction $\mathscr{L}(\varepsilon|x) = \mathscr{L}(\varepsilon|x'\beta_0)$, the maximum rank correlation estimator and its variants will not be consistent in general (being based on monotonicity of the relation between y and $x'\beta_0$). Instead, the resulting index restriction on the dependent variable, $\mathscr{L}(y|x) = \mathscr{L}(y|x'\beta_0)$, or the implied mean index restriction $E[y|x] = E[y|x'\beta_0]$, can be used to form estimators of β_0. Some of these estimators impose strong restrictions on the distribution of the regressors, while others use nonparametric estimators to sidestep such restrictions.

A number of papers in the statistics and econometrics literature have noted that, under the index restriction, certain misspecified maximum likelihood estimators will be consistent for β_0 (up to an intercept and scale normalization) when the regressors satisfy a particular linearity condition on their conditional expectations. Chung and Goldberger (1984) show that the classical least squares regression coefficients for y on x (and a constant term) will be consistent up to scale if the joint distribution of the regressors and latent variable $y^* \equiv x'\beta_0 + \varepsilon$ satisfies

$$E[x|y^*] \equiv \mu_0 + v_0 y^* \tag{3.19}$$

for some fixed vectors μ_0 and v_0; in this case, the least squares coefficients tend to $\kappa\beta_0$, where κ is the population least squares regression coefficient of $y \equiv t(y^*)$ on

y^*. Greene (1981, 1983) derives similar results for classical least squares estimates in the special case of a censored dependent variable. Brillinger (1983) shows consistency of classical least squares estimates for the general transformation model when the regressors are jointly normally distributed, which implies that the conditional distribution of the regressors x given the index $x'\beta_0$ has the linear form

$$E[x|x'\beta_0] \equiv \mu_0 + v_0(x'\beta_0) \tag{3.20}$$

for some μ_0 and v_0. Ruud (1983) noted that condition (3.20) (with a full-rank condition on the distribution of the regressors) was sufficient for consistency (up to scale) of a misspecified maximum likelihood estimator of β_0 in a binary response model with independence of the errors and regressors; this result was extended by Ruud (1986) to include all misspecified maximum likelihood estimators for latent variable models when (3.11), (3.20) and independence of the errors and regressors are assumed. Li and Duan (1989) have recently noted this result, emphasizing the importance of convexity of the assumed likelihood function (which ensures uniqueness of the minimizer $\kappa\beta_0$ of the limiting objective function). As Ruud points out, all of these results use the fact that the least squares or misspecified ML estimators $\hat{\alpha}$ and $\hat{\gamma}$ of the intercept term and slope coefficients satisfy a sample moment condition of the form

$$0 = \sum_{i=1}^{N} r(y_i, \hat{\alpha} + x_i'\hat{\gamma}) \begin{bmatrix} 1 \\ x_i \end{bmatrix} \tag{3.21}$$

for some "quasi-residual" function $r(\cdot)$. Letting $\tilde{r}(x'\beta_0, \alpha + x'\gamma) \equiv E[r(y, \alpha + x'\gamma)|x]$ and imposing condition (3.20), the value $\gamma^* = \kappa\beta_0$ will solve the corresponding population moment condition if κ and the intercept α are chosen to satisfy the two conditions

$$0 = E[\tilde{r}(x'\beta_0, \alpha + \kappa(x'\beta_0))] = E[\tilde{r}(x'\beta_0, \alpha + \kappa(x'\beta_0))(x'\beta_0)],$$

since the population analogue of condition (3.21) then becomes

$$0 = E\left\{\tilde{r}(x'\beta_0, \alpha + \kappa(x'\beta_0))\left(\begin{bmatrix} 1 \\ \mu_0 \end{bmatrix} + \begin{bmatrix} 0 \\ v_0 \end{bmatrix}(x'\beta_0)\right)\right\}$$

under the restriction (3.20). (An analogous argument works for condition (3.19), replacing $x'\beta_0$ with y^* where appropriate; in this case, the index restriction $\mathcal{L}(y|x) = \mathcal{L}(y|x'\beta_0)$ is not necessary, though this condition may not be as easily verified as (3.20).) Conditions (3.19) and (3.20) are strong restrictions which seem unlikely to hold for observational data, but the consistency results may be useful in experimental design settings (where the distribution of the regressors can be chosen to satisfy

(3.20)), and the results suggest that the inconsistency of traditional maximum likelihood estimators may be small when the index restriction holds and (3.19) or (3.20) is approximately satisfied.

If the regressors are assumed to be jointly continuously distributed with known density function $f_x(x)$, modifications of least squares estimators can yield consistent estimators of β_0 (up to scale) even if neither (3.19) nor (3.20) holds. Ruud (1986) proposed estimation of β_0 by weighted least squares,

$$\hat{\beta} = \left[\sum_{i=1}^{N} (\phi(x_i)/f_x(x_i))(x_i - \dot{x})(x_i - \dot{x})' \right]^{-1} \sum_{i=1}^{N} (\phi(x_i)/f_x(x_i))(x_i - \dot{x})(y_i - \dot{y}), \tag{3.22}$$

where $\phi(x)$ is any density function for a random vector satisfying condition (3.20) (for example, a multivariate normal density function) and

$$\dot{x} = \frac{1}{N} \sum_{i=1}^{N} (\phi(x_i)/f_x(x_i))\, x_i, \tag{3.23}$$

with an analogous definition for $\dot{y}$. This reweighting ensures that the probability limit for the weighted least squares estimator in (3.22) is the same as the probability limit for an unweighted least squares estimator with regressors having marginal density $\phi(x)$; since this density is assumed to satisfy (3.20), the resulting estimator will be consistent for β_0 (up to scale) by the results cited above.

A different approach to use of a known regressor density was taken by Stoker (1986), who used the mean index restriction $E[y|x] = E[y|x'\beta_0] \equiv G(x'\beta_0)$ implied by the transformation model with a strong index restriction on the errors. If the nuisance function $G(\cdot)$ is assumed to be smooth, an average of the derivative of $E[y|x]$ with respect to the regressors x will be proportional to β_0:

$$E[\partial E[y|x]/\partial x] = E[\mathrm{d}G(x'\beta_0)/\mathrm{d}(x'\beta_0)]\,\beta_0 \equiv \kappa^*\beta_0. \tag{3.24}$$

Furthermore, if the regressor density $f_x(x)$ declines smoothly to zero on the boundary of its support (which is most plausible when the support is unbounded), an integration-by-parts argument yields

$$\kappa^*\beta_0 = -E\{y\,\partial \ln[f_x(x)]/\partial x\}, \tag{3.25}$$

which implies that β_0 can be consistently estimated (up to scale) by the sample average of y_i times the derivative of the log-density of the regressors, $\partial \ln[f_x(x_i)]/\partial x$. Also, using the facts that

$$E\{\partial \ln[f_x(x)]/\partial x\} = 0, \qquad E\{(\partial \ln[f_x(x)]/\partial x)\, x'\} = -I, \tag{3.26}$$

Stoker proposed an alternative estimator of $\kappa^* \beta_0$ as the slope coefficients of an instrumental variables fit of y_i on x_i using the log-density derivatives $\partial \ln[f_x(x_i)]/\partial x$, and a constant as instruments. This estimator, as well as Ruud's density-weighted least squares estimator, is easily generalized to include models which have regressor density $f_x(x;\tau_0)$ of known parametric form, by substitution of a preliminary estimator $\hat{\tau}$ for the unknown distribution parameters and accounting for the variability of this preliminary estimator in the asymptotic covariance matrix formulae, using formula (1.53) in Section 1.4 above.

When the regressors are continuously distributed with density function $f_x(x)$ of unknown form, nonparametric (kernel) estimators of this density function (and its derivatives) can be substituted into the formulae for the foregoing estimators. Although the nonparametrically-estimated components necessarily converge at a rate slower than $N^{1/2}$, the corresponding density-weighted LS and average derivative estimators will be root-N-consistent under appropriate conditions, because they involve averages of these nonparametric components across the data. Newey and Ruud (1991) give conditions which ensure that the density-weighted LS estimator (defined in (3.22) and (3.23)) is root-N-consistent and asymptotically normal when $f_x(x)$ is replaced by a kernel estimator $\hat{f}_x(x)$. These conditions include the requirement that the reweighting density $\phi(x)$ is nonzero only inside a compact set which has $f_x(x)$ bounded above zero, to guarantee that the reciprocal of the corresponding nonparametric estimator $\hat{f}_x(x)$ is well-behaved. Härdle and Stoker (1989) and Stoker (1991) considered substitution of the derivative of a kernel estimator of the log-density, $\partial \ln[\hat{f}_x(x)]/\partial x$ into a sample analogue of condition (3.26) (which deletes observations for which $\partial \ln[\hat{f}_x(x_i)]/\partial x$ is small), and gave conditions for root-N-consistency and asymptotic normality of the resulting estimator.

A "density-weighted" variant on the average derivative estimator was proposed by Powell et al. (1989), using the fact that

$$E[f_x(x)\partial E[y|x]/\partial x] = E[f(x)\,\mathrm{d}G(x'\beta_0)/\mathrm{d}(x'\beta_0)]\,\beta_0 \equiv \kappa^+ \beta_0 = -2E\{y\,\partial f_x(x)/\partial x\}, \tag{3.27}$$

where the last inequality follows from a similar integration-by-parts argument as used to derive (3.25). The resulting estimator $\hat{\delta}$ of $\delta_0 \equiv \kappa^+ \beta_0$,

$$\hat{\delta} = \frac{1}{N} \sum_{i=1}^{N} -2(\partial \hat{f}_x(x_i)/\partial x)\, y_i, \tag{3.28}$$

was shown to have lth component of the form

$$\hat{\delta}_l = \binom{N}{2}^{-1} \sum_{i=1}^{N} \sum_{j=i+1}^{N-1} \omega_{lN}(x_i - x_j)\left[\frac{y_i - y_j}{x_{il} - x_{jl}}\right], \tag{3.29}$$

with weights $\omega_{lN}(x_i - x_j)$ which tend to zero as $\|x_i - x_j\|$ increases, and, for fixed

$\|x_i - x_j\|$, which tend to zero as N increases. Thus, the estimator implicitly uses finite difference ratios to approximate derivatives, averaging over those difference ratios for which the denominator is small. An instrumental variables version of the estimator, which uses $\partial \hat{f}_x(x_i)/\partial x$ as instrumental variables for a linear fit of y_i on x_i, was also proposed, using the integration-by-parts condition

$$-2E\{(\partial f_x(x)/\partial x)x'\} = E[f_x(x)]I. \tag{3.30}$$

Because the estimator $\hat{\delta}$ (and the components of its instrumental variables version) is of the U-statistic form considered in Section 1.4, its root-N-consistency and asymptotic normality are relatively simple to establish under appropriate conditions; Powell et al. (1989) showed its influence function is

$$\psi(y, x, \delta_0) = 2\{(f_x(x)\partial E[y|x]/\partial x - \delta_0) + (y - E[y|x])\partial f_x(x)/\partial x\}. \tag{3.31}$$

As in the Härdle and Stoker (1989) and Newey and Ruud (1991) papers, the conditions imposed on the density-weighted average derivative estimator $\hat{\delta}$ include particular rates of convergence of a "bandwidth" sequence (governing the degree of smoothing in the nonparametric estimators) to zero, and the use of "higher-order bias reducing" kernels to ensure that the asymptotic bias of the estimator is of $o(N^{-1/2})$.

If some components of the regressor vector x are not continuously distributed, neither the density-weighted LS nor average derivative estimation approaches are applicable, since they need a regressor density $f_x(x)$ which is well-defined (with respect to the Lebesgue measure). In this general setting, Ichimura (1992) proposed a semiparametric M-estimator for β_0 (up to scale) under an index restriction on the errors, using the conditional mean formulation

$$E[y|x] = G(x'\beta_0) = E[y|x'\beta_0] = E[y|x'b]|_{b=\beta_0}. \tag{3.32}$$

This estimator, a specialization of the Friedman and Stuetzle (1981) projection pursuit approach, uses kernel regression to estimate $E[y|x'b]$ as a function of b, then chooses $\hat{\beta}$ to maximize

$$S_N(b) = \sum_{i=1}^{N} \hat{w}(x_i)(y_i - \hat{E}[y_i|x_i'b])^2, \tag{3.33}$$

where the weights $\hat{w}(x)$ are constructed to equal zero where the nonparametric estimator of the conditional expectation is imprecise. Ichimura (1992) gave conditions for the identification of β_0 and the root-N-consistency and asymptotic normality of $\hat{\beta}$. The formula for the asymptotic covariance of this estimator is similar to the analogous formula for a weighted nonlinear least squares estimator with known expectation function $G(\cdot)$, except that the regression vector x, where it would appear

separately from the index $x'\beta_0$ in that formula, is replaced by the deviation of the regressors from their conditional mean given the index, $x - E[x|x'\beta_0]$. Newey and Stoker (1993) derived the semiparametric efficiency bound for estimation of β_0 (up to a scale normalization on one coefficient) under condition (3.32), which has a similar form to the semiparametric efficiency bound for estimation under exclusion restrictions given by Chamberlain (1992), as described in Section 2.5 above.

3.3. *Censored and truncated regression models*

A general notation for censored regression models which covers fixed and random censoring takes the dependent variable y and an observable indicator variable d to be generated as

$$y = \min\{x'\beta_0 + \varepsilon, u\}, \qquad d = 1\{y < u\}. \tag{3.34}$$

This notation covers the censored Tobit model with the dependent variable censored below at zero (with $u \equiv 0$ and a sign change on the dependent and explanatory variables) and the accelerated failure time model (y equals log failure time) with either fixed (u always observable) or random censoring times. Given a parametric density $f(\varepsilon; \tau_0)$ for the error terms (assumed independent of x), estimation of the resulting parametric model can be based upon maximization of the likelihood function

$$\mathscr{L}_N^C(\beta, \tau; F) = \frac{1}{N}\sum_{i=1}^{N} (d_i \ln[f(y_i - x_i'\beta; \tau)] + (1 - d_i)\ln[1 - F(u_i - x_i'\beta; \tau)]) \tag{3.35}$$

over possible values for β_0 and τ_0, where $F(\cdot)$ is the c.d.f. of ε (i.e. the antiderivative of the density $f(\cdot)$). This likelihood is actually the conditional likelihood of y_i, d_i given the regressors $\{x_i\}$ and the censoring points $\{u_i\}$ for all observations (assuming u_i is independent of y_i and x_i), but since it only involves the censoring point u_i for those observations which are censored, maximization of the likelihood in (3.35) is equally feasible for fixed or random censoring. For truncated data (i.e. sampling conditional on $d = 1$), the likelihood function becomes

$$\mathscr{L}_N^T(\beta, \tau; F) = \frac{1}{N}\sum_{i=1}^{N} \ln[f(y_i - x_i'\beta; \tau)/F(u_i - x_i'\beta; \tau)]; \tag{3.36}$$

here the truncation points must be known for all observations.

When the error density is Gaussian (or in a more general exponential family), the first-order conditions for the maximum likelihood estimator of β_0 with censored data can be interpreted in terms of the "EM algorithm" (Dempster et al. (1977), as

a solution to

$$0 = \frac{1}{N}\sum_{i=1}^{N} (y_i^+(\hat{\beta}, \hat{\tau}) - x_i'\beta)x_i, \tag{3.37}$$

where

$$\begin{aligned} y_i^+(\beta_0, \tau_0) &= d_i y_i + (1 - d_i)E[x_i'\beta_0 + \varepsilon_i | d_i = 0, x_i, u_i] \\ &= d_i y_i + (1 - d_i)\left[x_i'\beta_0 + \int_{u_i - x_i'\beta_0}^{\infty} \varepsilon f(\varepsilon; \tau_0)\, d\varepsilon \right], \end{aligned} \tag{3.38}$$

with a similar expression for the nuisance parameter estimator $\hat{\tau}$. Related formulae for the conditional mean of y given x and u,

$$E[y|x,u] = [1 - F(u - x'\beta_0)]u + \int_{-\infty}^{u - x'\beta_0} [x'\beta_0 + \varepsilon] f(\varepsilon; \tau_0)\, d\varepsilon, \tag{3.39}$$

or for the conditional mean of y given x and u and with $d = 1$,

$$E[y|x,u,d=1] = [F(u - x'\beta_0)]^{-1} \int_{-\infty}^{u - x'\beta_0} [x'\beta_0 + \varepsilon] f(\varepsilon; \tau_0)\, d\varepsilon, \tag{3.40}$$

can be used to define nonlinear least squares estimators for censored data (or for truncated data using (3.40)) in a fully parametric model.

As discussed in Section 2.1 above, the parameters of interest β_0 for the censored regression model (3.34) will not in general be identified if the error terms are assumed only to satisfy a constant conditional mean restriction, because the structural function is not invertible in the error terms. However, the monotonicity of the censoring transformation in ε for fixed x and u implies that the constant conditional quantile restrictions discussed in Section 2.2 will be useful in identifying and consistently estimating β_0. For fixed censoring (at zero), Powell (1984) proposed a least absolute deviations estimator for β_0 under the assumption that the error terms had conditional median zero; in the notation of model (3.34), this estimator $\hat{\beta}$ would be defined as

$$\hat{\beta} = \underset{\Theta}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} |y_i - \min\{x_i'\beta, u_i\}| \tag{3.41}$$

where Θ is the (compact) parameter space. Since the conditional median of y given x and u depends on the censoring value u for all observations (even if y is uncensored), the estimator is not directly applicable to random censoring models. Demonstration

of the root-N-consistency and asymptotic normality of this estimator follows the steps outlined in Section 2.3. The asymptotic covariance matrix of $\sqrt{N}(\hat{\beta} - \beta_0)$ for this model will be $H_0^{-1} V_0 H_0^{-1}$, with

$$H_0 \equiv 2E[f(0|x)\, 1\{x'\beta_0 < u\} xx'] \quad \text{and} \quad V_0 \equiv E[1\{x'\beta_0 < u\} xx'].$$

$f(0|x)$ is the conditional density of the error term ε at its median, zero.

This approach was extended to the model with a general constant quantile restriction by Powell (1986a), which derived analogous conditions for consistency and asymptotic normality. Under the stronger restriction that the error terms are independent of the regressors, this paper showed how more efficient estimators of the slope coefficients in β_0 could be obtained by combining coefficients estimated at different quantiles, and how the assumption of independent errors could be tested by testing convergence of the differences in quantile slope estimators to zero, as proposed by Koenker and Bassett (1982) for the linear model. Nawata (1990) proposed a two-step estimator for β_0 which calculates a nonparametric estimator of the conditional median of y given x, in the first step, by grouping the regressors into cells and computing the within-cell medians of the dependent variable. The second step treats these cell medians $\tilde{y}_j$ and the corresponding cell averages of the regressors $\tilde{x}_j$ as raw data in a Gaussian version of the likelihood function (3.35) and weights these quasi-observations by the cell frequencies (which would be optimal if the conditional density of the errors at the median were constant). Nawata gives conditions for the consistency of this estimator, and shows how its asymptotic distribution approaches the distribution of the censored least absolute deviations estimator (defined in (3.41)) as the regressor cells become small. And, as mentioned in Section 3.2, Newey and Powell (1990) showed that an efficient estimator of β_0, under a quantile restriction on the errors, is a weighted quantile estimator with weights proportional to $f(0|x)$, the conditional density of the errors at their conditional quantile, and proposed a feasible one-step version of this estimator which is asymptotically efficient.

When the censoring value u is observed only for censored observations, with u independently distributed from (y, x), Ying et al. (1991) propose a quantile estimator for β_0 under the restriction $\Pr\{\varepsilon \leqslant 0|x\} \equiv \pi \in (0, 1)$ using the implied relation

$$\begin{aligned} \Pr\{y > x'\beta_0|x\} &= \Pr\{x'\beta_0 < u \quad \text{and} \quad \varepsilon > 0|x\} \\ &= \Pr\{x'\beta_0 < u|x\} \Pr\{\varepsilon > 0|x\} \\ &= H(x'\beta_0)(1 - \pi), \end{aligned} \tag{3.42}$$

where $H(c) \equiv \Pr\{u > c\}$ is the survivor function of the random variable u. The unknown function $H(\cdot)$ can be consistently estimated using the Kaplan and Meier (1958) product-limit estimator for the distribution function for censored data. The resulting consistent estimator $\hat{H}(\cdot)$ uses only the dependent variables $\{y_i\}$ and the

censoring indicators $\{d_i\}$. Ying et al. (1991) define a quantile estimator $\hat{\beta}$ as a solution to estimating equations of the form

$$0 \cong \frac{1}{N}\sum_{i=1}^{N} [[\hat{H}(x_i'\hat{\beta})]^{-1} 1\{y_i > x_i'\hat{\beta}\} - (1-\pi)]x_i, \tag{3.43}$$

based on the conditional moment restriction (3.42) and give conditions for the root-N-consistency and asymptotic normality of this estimator. Since $H(x'\beta_0) = \hat{H}(x'\beta_0) = 1\{x'\beta_0 \leqslant u_0\}$ when the censoring points u_i are constant at some value u_0 with probability one, these equations are not well-defined for fixed censoring (say, at zero) except in the special case $\Pr\{x'\beta_0 \leqslant u_0\} \equiv 1$. A modification of the sample moment conditions defined in (3.43),

$$0 \cong \frac{1}{N}\sum_{i=1}^{N} [1\{y_i > x_i'\hat{\beta}\} - [\hat{H}(x_i'\hat{\beta})](1-\pi)]x_i, \tag{3.44}$$

would allow a constant censoring value, and when $\pi = \frac{1}{2}$ would reduce to the subgradient condition for the minimization problem (3.41) in this case. Unfortunately, this condition may have a continuum of inconsistent roots, if $\hat{\beta}$ can be chosen so that $x_i'\hat{\beta} > u_i$ for all observations. It is not immediately clear whether an antiderivative of the right-hand side of (3.44) would yield a minimand which could be used to consistently estimate β_0 under random censoring, as it does (yielding (3.41) for $\pi = \frac{1}{2}$) for fixed censoring.

Because the conditional median (and other quantiles) of the dependent variable y depend explicitly on the error distribution when the dependent variable is truncated, quantile restrictions are not helpful in identifying β_0 for truncated samples. With a stronger restriction of conditional symmetry of the errors about a constant (zero), the "symmetric trimming" idea mentioned in Section 2.3 can be used to construct consistent estimators for both censored and truncated samples. Powell (1986b) proposed a symmetrically truncated least squares estimator of β_0 for a truncated sample. The estimator exploited the moment condition

$$E[1\{y > 2x'\beta_0 - u\}(y - x'\beta_0)|x, y < u] = E[1\{\varepsilon > x'\beta_0 - u\}\varepsilon|x, \varepsilon < u - x'\beta_0] = 0 \tag{3.45}$$

which holds for the truncated model under conditional symmetry given x and u. The resulting estimator is defined to minimize

$$T_N(\beta) \equiv \frac{1}{N}\sum_{i=1}^{N} \left(y_i - \min\left\{x_i'\beta, \frac{y_i}{2} - u_i\right\}\right)^2, \tag{3.46}$$

which yields a sample analogue to (3.45) as an approximate first-order condition.

Similarly, a symmetrically censored least squares estimator for the censored regression model (3.34) will solve a sample moment condition based upon the condition

$$E[\max\{y, 2x'\beta_0 - u\} - x'\beta_0 | x] = E[\max\{\min(\varepsilon, u - x'\beta_0), x'\beta_0 - u\} | x] = 0. \tag{3.47}$$

The root-N-consistency and asymptotic normality of these estimators were established by Powell (1986b). In addition to conditional symmetry and a full-rank condition on the matrix $V_0 \equiv E[1\{x'\beta_0 < u\}xx']$, a unimodality condition on the error distribution was imposed in the truncated case. A variant on the symmetric trimming approach was proposed by M. Lee (1993a, b) which, for a fixed scalar $w > 0$, constructed estimators for truncated and censored samples based on the moment conditions

$$\begin{aligned} &E[1\{u - x'\beta_0 > w\}\, 1\{|y - x'\beta_0| < w\}(y - x'\beta_0) | x, y < u] \\ &\quad = E[1\{u - x'\beta_0 > w\}\, 1\{|\varepsilon| < w\}\varepsilon | x] = 0 \end{aligned} \tag{3.48}$$

and

$$\begin{aligned} &E[1\{u - x'\beta_0 > w\} \min\{|y - x'\beta_0|, w\} \operatorname{sgn}\{y - x'\beta_0\} | x] \\ &\quad = E[1\{u - x'\beta_0 > w\} \min\{|\varepsilon|, w\} \operatorname{sgn}\{\varepsilon\} | x] = 0, \end{aligned} \tag{3.49}$$

respectively. Newey (1989a) derives the semiparametric efficiency bounds for estimation of β_0 under conditional symmetry with censored and truncated samples, noting that the symmetrically truncated least squares estimator attains that efficiency bound in the special case where the unknown error distribution is, in fact, Gaussian (the analogous result does not hold, though, for the symmetrically censored estimator).

As described at the end of Section 2.2, conditional mode restrictions can be used to identify β_0 for truncated data, and an estimator proposed by M. Lee (1992) exploits this restriction. This estimator solves a sample analogue to the characterization of β_0 as the solution to the minimization problem

$$\beta_0 = \underset{\Theta}{\operatorname{argmin}} \Pr\{|y - \min\{u_i + \omega, x_i'b\}| > \omega\}, \tag{3.50}$$

as long as the modal interval of length 2ω for the untruncated error distribution is assumed to be centered at zero. M. Lee (1992) showed the $N^{1/3}$-consistency of this estimator and considered its robustness properties.

Most of the literature on semiparametric estimation for censored and truncated regression in both statistics and econometrics has been based upon independence restrictions. Early estimators of β_0 for random censoring models which relaxed the assumed parametric form of the error distribution (but maintained independence

of the censoring times and the latent dependent variable) were proposed by Buckley and James (1979) and Koul et al. (1981). The Buckley–James estimator uses the Kaplan–Meier (1958) nonparametric estimator for the error distribution, applied using residuals $\hat{\varepsilon} \equiv y - x'\hat{\beta}$ and their censoring points $u - x'\hat{\beta}$, to obtain nonparametric estimators of the conditional expectation $E[y|x, d = 0]$ in (3.38) above. Then, (3.37) and (3.38) are used iteratively to obtain a semiparametric analogue of the EM algorithm. Although Buckley and James did not rigorously establish consistency of this estimator, they demonstrated that it was well-behaved in practice, and Ritov (1990) showed how a modification of this approach yields a root-N-consistent and asymptotically normal estimator. Koul et al. (1981) proposed estimation of β_0 using a weighted least squares regression of the uncensored dependent variables on their corresponding regressors, using the inverse of the estimated survival function for u evaluated at y, $[\hat{H}(y)]^{-1}$, as weights. Using the fact that

$$E[d|x, \varepsilon] = \Pr\{x'\beta_0 + \varepsilon < u\} = H(x'\beta_0 + \varepsilon), \tag{3.51}$$

this estimator exploits the moment condition

$$E[d \cdot [H(y)]^{-1}(y - x'\beta_0)|x] = E[E[d|x, \varepsilon][H(x'\beta_0 + \varepsilon)]^{-1}\varepsilon|x] = E[\varepsilon|x] = 0. \tag{3.52}$$

Thus, only a conditional mean restriction is required for consistency of the resulting estimator; however, the upper limit of the support of the censoring variable u must be larger than the upper limit of the support of the latent variable $y^* = x'\beta_0 + \varepsilon$, which rules out a fixed censoring point (unless censoring never occurs).

In the econometrics literature, where the censoring value u is assumed to be fixed at zero, Duncan (1986) and Fernandez (1986) proposed semiparametric profile likelihood estimators of β_0 by replacing the unknown error density and cumulative by nonparametric estimators, using different smoothing techniques. Horowitz (1986) showed consistency of a nonlinear least squares estimator for β_0 using an integration-by-parts formula for the conditional mean of $y = \min\{x'\beta_0 + \varepsilon, 0\}$ given x:

$$E[y|x] = \int_{-\infty}^{-x'\beta_0} (e + x'\beta_0) f(e)\,de = -\int_{-\infty}^{-x'\beta_0} F(e)\,de, \tag{3.53}$$

where $f(\cdot)$ and $F(\cdot)$ are the error density and cumulative. To obtain a feasible estimator, the unknown error cumulative $F(\cdot)$ is replaced by its Kaplan–Meier estimator based upon residuals, as for the Buckley–James estimator. Horowitz (1988) constructed a more efficient nonlinear weighted least squares version of this estimator and showed its root-N-consistency and asymptotic normality. A similar approach, based on the analogous expression for the conditional mean of y given x and with $d = 1$, was proposed by Moon (1989).

Pairwise difference estimators for the censored and truncated regression models have also been constructed by Honoré and Powell (1991). For model (3.34) with fixed censoring, and using the notation of Section 2.4, these estimators were based upon the transformation

$$e_{ij}(\theta) \equiv e(z_i, z_j, \beta) = \min\{y_i - x_i'\beta, \quad u_i - x_j'\beta\}, \tag{3.54}$$

which satisfies

$$e_{ij}(\theta_0) = \min\{\min\{\varepsilon_i, u_i - x_i'\beta_0\}, u_i - x_j'\beta_0\} = \min\{\varepsilon_i, u_i - x_i'\beta_0, u_i - x_j'\beta_0\},$$

so that $e_{ij}(\theta_0)$ and $e_{ji}(\theta_0)$ are clearly independently and identically distributed given x_i and x_j. Again choosing $l(x_i, x_j, \theta) = x_i - x_j$, the pairwise difference estimator for the censored regression model was given as a solution to the sample moment condition (2.39) of Section 2.4 above. These estimating equations were shown to have a unique solution, since they correspond to first-order conditions for a convex minimization problem. Honoré and Powell (1991) also considered estimation of the truncated regression model, in which y_i and x_i are observed only if y_i is positive; that is, if $y_i = x_i'\beta_0 + v_i$, where v_i has the conditional distribution of ε_i given $\varepsilon_i > -x_i'\beta_0$, then $\mathscr{L}(v_i|x_i) = \mathscr{L}(\varepsilon_i|x_i, \varepsilon_i > -x_i'\beta_0)$. Again assuming the untruncated errors ε_i are i.i.d. and independent of the regressors x_i, a pairwise difference estimator of β_0 was defined using the transformation

$$e(z_i, z_j, \beta) \equiv (y_i - x_i'\beta)\, 1(y_i - x_i'\beta > -x_j'\beta)\, 1(y_j - x_j'\beta > -x_i'\beta). \tag{3.55}$$

When evaluated at the true value β_0, the difference

$$e_{ij}(\beta_0) - e_{ji}(\beta_0) = (v_i - v_j)\, 1(v_i > -x_j'\beta)\, 1(v_j > -x_i'\beta) \tag{3.56}$$

is symmetrically distributed around zero given x_i and x_j. As for the censored case, the estimator $\hat{\beta}$ for this model was defined using $l(x_i, x_j, \theta) = (x_i - x_j)$ and (2.39) through (2.40) above. When the function $\xi(d) = \text{sgn}(d)$, the solution to (2.39) for this model was proposed by Bhattacharya et al. (1983) as an estimator of β_0 for this model under the assumption that x_i is a scalar. The general theory derived for minimizers of mth-order U-statistics (discussed in Section 1.3) was applied to show root-N-consistency and to obtain the large-sample distributions of the pairwise difference estimators for the censored and truncated regression models.

3.4. *Selection models*

Rewriting the censored selection model of (1.21) and (1.22) as

$$\begin{aligned} d &= 1\{x_1'\delta_0 + \eta > 0\}, \\ y &= d[x_2'\beta_0 + \varepsilon] \end{aligned} \tag{3.57}$$

(for $y_1 \equiv d$, $y_2 \equiv y$, $\beta_0^1 \equiv \delta_0$, and $\beta_0^2 \equiv \beta_0$), a fully parametric model would specify the functional form of the joint density $f(\varepsilon, \eta; \tau_0)$ of the error terms. Then the maximization of the average log-likelihood function

$$\mathscr{L}_N(\beta, \delta, \tau; f) = \frac{1}{N} \sum_{i=1}^{N} \left[d_i \ln \left[\int_{-x_{1i}'\delta}^{\infty} f(y_i - x_{2i}'\beta, \eta; \tau)\, d\eta \right] + (1 - d_i) \ln \left[\int_{-\infty}^{\infty} \int_{-x_{1i}'\delta}^{\infty} f(\varepsilon, \eta; \tau)\, d\eta\, d\varepsilon \right] \right] \tag{3.58}$$

over β, δ, and τ in the parameter space. An alternative estimation method, proposed by Heckman (1976), can be based upon the conditional mean of y given x and with $d = 1$:

$$E[y|x, d = 1] = x_2'\beta_0 + \left[\int_{-\infty}^{\infty} \int_{-x_1'\delta_0}^{\infty} f(\varepsilon, \eta; \tau_0)\, d\eta\, d\varepsilon \right]^{-1} \times \left[\int_{-\infty}^{\infty} \int_{-x_1'\delta_0}^{\infty} \varepsilon f(\varepsilon, \eta; \tau_0)\, d\eta\, d\varepsilon \right] \equiv x_2'\beta_0 + \lambda(x_1'\delta_0; \tau_0). \tag{3.59}$$

When the "selection correction function" $\lambda(x_1'\delta; \tau)$ is linear in the distributional parameters τ (as is the case for bivariate Gaussian densities), a two-step estimator of β_0 can be constructed using linear least squares, after inserting a consistent first-step estimator $\tilde{\delta}$ of δ_0 (using the indicator d and regressors x_1 in the binary log-likelihood of (3.2)) into the selection correction function. Alternatively, a nonlinear least squares estimator of the parameters can be constructed using (3.59), which is also applicable for truncated data (i.e. for y and x being observed conditional on $d = 1$).

To date, semiparametric modelling of the selection model (3.57) has imposed independence or index restrictions on the error terms (ε, η). Chamberlain (1986a) derived the semiparametric efficiency bound for estimation of β_0 and δ_0 in (3.57) when the errors are independent of the regressors with unknown error density. The form of the efficiency bound is a simple modification of the parametric efficiency bound for this problem when the error density is known, with the regression vectors x_1 and x_2 being replaced by their deviations from their conditional means, given the selection index, $x_1 - E[x_1|x_1'\delta_0]$ and $x_2 - E[x_2|x_1'\delta_0]$, except for terms which involve the index $x_1'\delta_0$. Chamberlain notes that, in general, nonsingularity of the semiparametric information matrix will require an exclusion restriction on x_2 (i.e. some component of x_1 with nonzero coefficient in δ_0 is excluded from x_2), as well as a normalization restriction on δ_0. The efficiency bound, which was derived imposing independence of the errors and regressors, apparently holds more generally when the joint distribution of the errors in (3.57), given the regressors, depends only upon the index $x_1'\delta_0$ appearing in the selection equation.

Under this index restriction, the conditional mean of y given $d = 1$ and x will have the same form as in (3.59), but with a selection correction function of unknown form. More generally, conditional on $d = 1$, the dependent variable y has the linear representation $y = x_2'\beta_0 + \varepsilon$, where ε satisfies the distributional index restriction

$$\mathscr{L}(\varepsilon | d = 1, x) = \mathscr{L}(\varepsilon | d = 1, x_1'\delta_0) \quad \text{a.s.}, \tag{3.60}$$

so that other estimation methods for distributional index restrictions (discussed in Section 2.5) are applicable here. So far, though, the econometric literature has exploited only the weaker mean index restriction

$$E(\varepsilon | d = 1, x) = E(\varepsilon | d = 1, x_1'\delta_0). \tag{3.61}$$

A semiparametric analogue of Heckman's two-step estimator was constructed by Cosslett (1991), assuming independence of the errors and regressors. In the first step of this approach, a consistent estimator of the selectivity parameter δ_0 is obtained using Cosslett's (1983) NPMLE for the binary response model, described in Section 3.1 above. In this first step, the concomitant estimator $\hat{F}(\cdot)$ of the marginal c.d.f. of the selection error η is a step function, constant on a finite number J of intervals $\{\hat{I} \equiv (\hat{c}_{j-1}, \hat{c}_j), j = 1, \ldots, J\}$ with $c_0 \equiv -\infty$ and $c_J \equiv \infty$. The second-step estimator of β_0 approximates the selection correction function $\lambda(\cdot)$ by a piecewise-constant function on those intervals. That is, writing

$$y \equiv x_2'\beta_0 + \sum_{j=1}^{J} \lambda_j 1\{x_1'\delta_0 \in \hat{I}_j\} + \hat{e}, \tag{3.62}$$

the estimator $\hat{\beta}$ is constructed from a linear least squares regression of y on x_2 and the J indicator variables $\{1\{x_1'\hat{\delta} \in \hat{I}_j\}\}$. Cosslett (1991) showed consistency of the resulting estimator, using the fact that the number of intervals, J, increases slowly to infinity as the sample size increases so that the piecewise linear function could approximate the true selection function $\lambda(\cdot)$ to an arbitrary degree. An important identifying assumption was the requirement that some component of the regression vector x_1 for the selection equation was excluded from the regressors x_2 in the equation for y, as discussed by Chamberlain (1986a).

Although independence of the errors and regressors was imposed by Cosslett (1991), this was primarily used to ensure consistency of the NPML estimator of the selection coefficient vector δ_0. The same approach to approximation of the selection correction function will work under an index restriction on the errors, provided the first-step estimator of δ_0 only requires this index restriction. In a parametric context, L. Lee (1982) proposed estimation of β_0 using a flexible parametrization of the selection correction function $\lambda(\cdot)$ in (3.59). For the semiparametric model Newey (1988) proposed a similar two-step estimator, which in the second step used a series

approximation to the selection correction function to obtain the approximate model

$$y \equiv x_2'\beta_0 + \sum_{j=1}^{J} \lambda_j \rho_j(x_1'\delta_0) + e, \tag{3.63}$$

which was estimated (substituting a preliminary estimator $\hat{\delta}$ for δ_0) by least squares to obtain an estimator of β_0. Here the functions $\{\rho_j(\cdot)\}$ were a series of functions whose linear combination could be used to approximate (in a mean-squared-error sense) the function $\lambda(\cdot)$ arbitrarily well as $J \to \infty$. Newey (1988) gave conditions (including a particular rate of growth of the number J of series components) under which the estimator $\hat{\beta}$ of β_0 was root-N-consistent and asymptotically normal, and also discussed how efficient estimators of the parameters could be constructed.

As discussed in Section 2.5, weighted versions of the pairwise-difference estimation approach can be used under the index restriction of (3.61). Assuming a preliminary, root-N-consistent estimator $\hat{\delta}$ of δ_0 is available, Powell (1987) considers a pairwise-difference estimator of the form (2.55) when $\xi(d) = d$, $e_{ij}(\theta) = y_i - x_{i2}'\beta$ and $l(x_i, x_j, \theta) = x_{i2} - x_{j2}$, yielding the explicit estimator

$$\hat{\beta} = \left[\sum_{i<j} w_N((x_{i1} - x_{j1})'\hat{\delta})(x_{i2} - x_{j2})(x_{i2} - x_{j2})' \right]^{-1} \times \left[\sum_{i<j} w_N((x_{i1} - x_{j1})'\hat{\delta})(x_{i2} - x_{j2})(y_{i2} - y_{j2}) \right]. \tag{3.64}$$

Conditions were given in Powell (1987) on the data generating process, the weighting functions $w_N(\cdot)$, and the preliminary estimator $\hat{\delta}$ which ensured the root-N-consistency and asymptotic normality of $\hat{\beta}$. The dependence of this asymptotic distribution on the large-sample behavior of $\hat{\delta}$ was explicitly derived, along with a consistent estimator of the asymptotic covariance matrix. The approach was also extended to permit endogeneity of some components of x_{i2} using an instrumental variables version of the estimator. L. Lee (1991) considers system identification of semiparametric selection models with endogenous regressors and proposes efficient estimators of the unknown parameters under an independence assumption on the errors.

When the errors in (3.57) are assumed independent of the regressors, and the support of the selection error η is the entire real line, the assumption of a known parametric form $x_1'\delta_0$ of the regression function in the selection equation can be relaxed. In this case, the dependent variable y given $d = 1$ has the linear representation $y_i = x_i'\beta_0 + \varepsilon_i$, where the error term ε satisfies the distributional index restriction

$$\mathscr{L}(\varepsilon | d = 1, x) = \mathscr{L}(\varepsilon | d = 1, p(x_1)) \quad \text{a.s.}, \tag{3.65}$$

where now the single index $p(x_1)$ is the "propensity score" (Rosenbaum and Rubin

(1983)), defined as

$$p(x_1) = E[d|x_1] = E[d|x]. \tag{3.66}$$

Given a nonparametric estimator $\hat{p}(x_1)$ of the conditional mean $p(x_1)$ of the selection indicator, it is straightforward to modify the estimation methods above to accommodate this new index restriction, by replacing the estimated linear index $x_1'\hat{\delta}$ by the nonparametric index $\hat{p}(x_1)$ throughout. Choi (1990) proposed a series estimator of β_0 based on (3.63) with this substitution, while Ahn and Powell (1993) modified the weighted pairwise-difference estimator in (3.64) along these lines. Both papers used a nonparametric kernel estimator to construct $\hat{p}(x_1)$, and both gave conditions on the model, the first-step nonparametric estimator and the degree of smoothing in the second step which guaranteed root-N-consistency and asymptotic normality of the resulting estimators of β_0. The influence functions for these estimators depend upon the conditional variability of the errors ε and the deviations of the selection indicator from its conditional mean, $d - p(x_1)$. Newey and Powell (1993) calculate the semiparametric efficiency bounds for β_0 under the distributional index restriction (3.65) and its mean index analogue, while Newey and Powell (1991) discuss construction of semiparametric M-estimators which will attain these efficiency bounds.

For the truncated selection model (sampling from (3.57) conditional on $d = 1$), identification and estimation of the unknown parameters is much more difficult. Ichimura and Lee (1991) consider a semiparametric version of a nonlinear least squares estimator using the form of the truncated conditional mean function

$$E[y|x, d = 1] = x_2'\beta_0 + \lambda(x_1'\delta_0) \tag{3.67}$$

from (3.59) with $\lambda(\cdot)$ unknown, following the definition of Ichimura's (1992) estimator in (3.33) above. Besides giving conditions for identification of the parameters and root-N-consistency of their estimators, Ichimura and Lee (1991) consider a generalization of this model in which the nonparametric component depends upon several linear indices. If the linear index restriction (3.61) is replaced by the nonparametric index restriction (3.65), identification and consistent estimation of β_0 requires the functional independence of x_1 and x_2, in which case the estimator proposed by Robinson (1988), discussed in Section 2.5 above, will be applicable. Chamberlain (1992) derives the efficiency bound for estimation of the parameters of the truncated regression model under the index restriction (3.65).

Just as eliminating the information provided by the selection variable d makes identification and estimation of β_0 harder, a strengthening of the information in the selection variable makes estimation easier, and permits identification using other semiparametric restrictions on the errors. Honoré et al. (1992) consider a model in which the binary selection variable d is replaced by a censored dependent variable

y_1, so that the model becomes

$$y_1 = \max\{0, x_1'\delta_0 + \eta\},$$
$$y_2 = 1\{y_1 > 0\}[x_2'\beta_0 + \varepsilon]. \tag{3.68}$$

This model is called the "Type 3 Tobit" model by Amemiya (1985). Assuming conditional symmetry of the errors (ε, η) about zero given x (as defined in Section 2.3), the authors note that δ_0 can be consistently estimated using the quantile or symmetric trimming estimators for censored regression models discussed in Section 3.3, and, furthermore, by symmetrically trimming the dependent variable y_2 using the trimming function

$$h(y_1, y_2, x_1, x_2, \delta, \beta) \equiv 1\{0 < y_1 < 2x_1'\delta\}(y_2 - x_2'\beta), \tag{3.69}$$

the function $h(\cdot)$ satisfies the conditional moment restriction

$$E[h(y_1, y_2, x_1, x_2, \delta_0, \beta_0)|x] \equiv E[1\{-x_1'\delta_0 < \eta < x_1'\delta_0\}\varepsilon|x] = 0 \tag{3.70}$$

because of the joint conditional symmetry of the errors. By constructing a sample analogue of (3.70) (possibly based on other odd functions of $y_2 - x_2'\beta$) and inserting the preliminary estimator $\hat{\delta}$, Honoré et al. (1992) show the resulting estimator $\hat{\beta}$ to be root-N-consistent and asymptotically normal under relatively weak conditions on the model. Thus, with the additional information on the latent variable $x_1'\delta_0 + \eta$ provided by the censored variable y_2, it is possible to consistently estimate β_0 without obtaining explicit nonparametric estimators of infinite-dimensional nuisance functions.

3.5. *Nonlinear panel data models*

For panel data versions of the latent variable models considered above, with

$$y_s = t(\eta + x_s'\beta_0 + \varepsilon_s, \tau_0), \qquad s = 1, \ldots, T, \tag{3.71}$$

derivation of log-likelihood functions like the ones above is straightforward if the individual-specific intercept η is assumed independent of x (or its dependence is parametrically specified) with a distribution of known parametric form. The conditional density of $y \equiv (y_1, \ldots, y_T)$ given x for each individual can be obtained from the joint density of the convolution $u \equiv (\eta + \varepsilon_1, \ldots, \eta + \varepsilon_T)$, which, for special (e.g. Gaussian) choices of error distribution is of simple form. Maximum likelihood estimators of β_0 for these nonlinear "random effect" models have the usual optimality properties, but their consistency depends on proper specification of both the error

terms $\varepsilon \equiv (\varepsilon_1, \ldots, \varepsilon_T)$ and the random effect η. When the individual-specific intercepts are treated as unknown parameters ("fixed effects"), the corresponding log-likelihoods for the parameters β_0 and the vector of intercept terms $(\eta_1, \ldots, \eta_i, \ldots, \eta_N)$ are even simpler to derive, being of the same general forms as given above when the errors ε_s are assumed to be i.i.d. across individuals and time. However, because the vector of unknown intercept terms increases with the sample size, maximum likelihood estimators of these fixed effects will be inconsistent unless the number of time periods T also increases to infinity; moreover, the inconsistency of the fixed effect estimators leads to inconsistency of the estimators of the parameters of interest, β_0, as a consequence of the notorious "incidental parameters" problem (Neyman and Scott (1948)).

For some special parametric discrete response models, consistent estimators of β_0 with fixed effects can be obtained by maximizing a "conditional likelihood" function, which conditions on a fixed sum of the discrete dependent variable across time for each individual. In the special case $T = 2$, this is the same as maximizing the conditional likelihood given that $y_1 \neq y_2$ and that the estimation method is the analogue to estimation using pairwise differences (over time) for linear panel data models. Models for which a version of pairwise differencing can be used to eliminate the fixed effect in panel data include the binary logit model (Andersen (1970)), the Poisson regression model (Hausman et al. (1984)) and certain duration models (Chamberlain (1984)); however, these results require a particular (exponential) structure to the likelihood which does not hold in general.

For the binary, censored, and truncated regression models with fixed effects, estimators have been proposed under the assumption that the time-specific errors $\{\varepsilon_s\}$ are identically distributed across time periods s given the regressors x. Manski (1987) shows that, with $T = 2$ time periods, the conditional median of the difference $y_2 - y_1$ of the binary variables $y_s = 1\{x_s'\beta_0 + \varepsilon_s \geq 0\}$, given that $y_1 \neq y_2$, is $1\{(x_2 - x_1)'\beta_0 > 0\}$, so that a consistent estimator for β_0 will be

$$\hat{\beta} = \underset{\Theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} 1\{y_{i2} \neq y_{i1}\} |(y_{i2} - y_{i1}) - 1\{(x_2 - x_1)'\beta_0 > 0\}|, \tag{3.72}$$

which will be consistent under conditions on $(x_{i2} - x_{i1})$, etc., similar to those for consistency of the maximum score estimator. Honoré (1992) considered pairwise-difference estimators for censored and truncated regression models with fixed effects using the approach described in Section 3.3. Specifically, using the transformations given in (3.54) and (3.55) for the censored and truncated cases, respectively, estimators of the parameter vector β_0 in both cases were defined as solutions to minimization problems which generate a first-order condition of the form

$$0 \cong \sum_{i=1}^{N} \xi[e(z_{i2}, z_{i1}, \hat{\beta}) - e(z_{i1}, z_{i2}, \hat{\beta})](x_{i2} - x_{i1}). \tag{3.73}$$

As discussed at the end of Section 2.4, the expectation of the right-hand side of (3.73) will be zero when evaluated at β_0, even in the presence of a fixed effect. As for Manski's binary panel data estimator, this estimation approach can be generalized to allow for more than $T = 2$ time periods.

4. Summary and conclusions

As the previous section indicates, the theoretical analysis of the properties of estimators under various semiparametric restrictions is quite extensive, at least for the latent variable models considered above. The following table gives a general summary of the state of the econometric literature on estimation of several semiparametric models.

	Mean	Median	Mode	Index	Symmetry	Independence
Linear	3	3	1	0+	3	3
Transformed	3	3	0	0+	3	3
Censored	0	3	0	0+	3	3
Truncated	0	0	1	0	3	3
Binary	0	1	0	3	1	3
Monotone	0	1	0	2	1	2
Semilinear	3	2	?	3	2	3
Selection	0	?	?	3	2	3
Binary panel	0	?	?	?	?	1
Censored panel	0	?	?	?	?	2

Key: 0 – Not identified (0+ – Identified only up to scale); 1 – Parameter identified/consistent estimator; 2 – $\sqrt{N}$-consistent, asymptotically normal estimator; 3 – Efficient estimator.

Of course, this table should be viewed with caution, as some of its entries are ambiguous (for instance, the entry under "symmetry" for the "selection" row refers to the "Type 3 Tobit" model with a censored regression model as the selection equation, while the other columns presume a binary selection equation). Nevertheless, the table should be suggestive of areas where more research is needed.

The literature on the empirical application of semiparametric methods (apart from estimation of invertible models under conditional mean restrictions) is much less extensive. When applied to relatively small data sets (roughly 100 observations per parameter), the potential bias from misspecification of the parametric model has proven to be less important than the additional imprecision induced when parametric restrictions are relaxed. For example, Horowitz and Neumann (1987) and McFadden and Han (1987) estimate the parameters of an employment duration data set imposing independence and quantile restrictions, but for these data even maximum likelihood estimates are imprecise (in terms of their asymptotic standard errors). A similar outcome was obtained by Newey et al. (1990), which reanalyzed data on married women's labor supply originally studied (in a parametric context)

by Mroz (1987). For these data, estimates based upon semiparametric restrictions were fairly comparable to their parametric counterparts, with differences in the estimates having large standard errors. On the other hand, for larger data sets (with relatively few parameters), the bias due to distributional misspecification is more likely to be evident. Chamberlain (1990) and Buchinski (1991b) apply quantile methods to estimate the returns to education for a large, right-censored data set, and find these estimates to be quite precise. Other empirical papers which use semiparametric methods, with mixed success, include those by Deaton and Irish (1984), Newey (1987), Das (1991), Horowitz (1993), Bult (1992a, b), Horowitz and Markatou (1993), Deaton and Ng (1993) and Melenberg and van Soest (1993).

Besides the possible imprecision due to weakening of semiparametric restrictions, an obstacle to routine use of some of the estimators described in Section 3 is their dependence upon a choice of type and degree of "smoothing" imposed for estimators which depend explicitly upon nonparametric components of the model. Though this question has been widely studied in the literature on nonparametrics, the results are different when the nonparametric component is a nuisance parameter. Some early results on the proper degree of smoothing are available for some special cases of estimators for censored regression (Hall and Horowitz (1990)) or upon index restrictions (Hall and Marron (1987), Powell and Stoker (1991), Härdle et al. (1992)), but more theoretical results are needed to narrow the choice of possible estimators which depend upon nonparametrically-estimated components.

References

Ahn, H. and C.F. Manski (1993) "Distribution Theory for the Analysis of Binary Choice Under Uncertainty with Nonparametric Estimation of Expectations", *Journal of Econometrics*, forthcoming.

Ahn, H. and J.L. Powell (1993) "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism", *Journal of Econometrics*, forthcoming.

Amemiya, T. (1974) "The Nonlinear Two-Stage Least-Squares Estimator", *Journal of Econometrics*, 2, 105–110.

Amemiya, T. (1977) "The Maximum Likelihood and Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equations Model", *Econometrica*, 45, 955–968.

Amemiya, T. (1982) "Two Stage Least Absolute Deviations Estimators", *Econometrica*, 50, 689–711.

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

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

Andersen, E.B. (1970) "Asymptotic Properties of Conditional Maximum Likelihood Estimators", *Journal of the Royal Statistical Society*, Series B, 32, 283–301.

Andrews, D.W.K. (1987) "Consistency in Nonlinear Econometric Models, A Generic Uniform Law of Large Numbers", *Econometrica*, 55, 1465–1471.

Andrews, D.W.K. (1990a) "Asymptotics for Semiparametric Econometric Models, I. Estimation and Testing", Cowles Foundation, Yale University, Discussion Paper No. 908R.

Andrews, D.W.K. (1990b) "Asymptotics for Semiparametric Econometric Models, II. Stochastic Equicontinuity and Nonparametric Kernel Estimation", Cowles Foundation, Yale University, Discussion Paper No. 909R.

Andrews, D.W.K. (1991) "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models", *Econometrica*, 59, 307–345.

Arabmazar, A. and P. Schmidt (1981) "Further Evidence on the Robustness of the Tobit Estimator to Heteroscedasticity", *Journal of Econometrics*, 17, 253–258.
Arabmazar, A. and P. Schmidt (1982) "An Investigation of the Robustness of the Tobit Estimator to Non-Normality", *Econometrica*, 50, 1055–1063.
Bassett, G.S. and R. Koenker (1978) "Asymptotic Theory of Least Absolute Error Regression", *Journal of the American Statistical Association*, 73, 667–677.
Begun, J., W. Hall, W. Huang and J. Wellner (1983) "Information and Asymptotic Efficiency in Parametric-Nonparametric Models", *Annals of Statistics*, 11, 432–452.
Bhattacharya, P.K., H. Chernoff and S.S. Yang (1983) "Nonparametric Estimation of the Slope of a Truncated Regression", *Annals of Statistics*, 11, 505–514.
Bickel, P.J. (1982) "On Adaptive Estimation", *Annals of Statistics*, 10, 647–671.
Bickel, P.J. and K.A. Doksum (1981) "An Analysis of Transformations Revisited", *Journal of the American Statistical Association*, 76, 296–311.
Bickel, P.J., C.A.J. Klaasen, Y. Ritov and J.A. Wellner (1993) *Efficient and Adaptive Inference in Semiparametric Models*. Washington: Johns Hopkins University Press, forthcoming.
Bierens, H.J. (1987) "Kernel Estimators of Regression Functions", in: T.F. Bewley, ed., *Advances in Econometrics, Fifth World Congress*, vol. 1. Cambridge: Cambridge University Press.
Bloomfield, P. and W.L. Steiger (1983) *Least Absolute Deviations: Theory, Applications, and Algorithms*. Boston: Birkhauser.
Box, G.E.P. and D.R. Cox (1964) "An Analysis of Transformations", *Journal of the Royal Statistical Society*, Series B, 26, 211–252.
Brillinger, D.R. (1983) "A Generalized Linear Model with 'Gaussian' Regressor Variables", in: P.J. Bickel, K.A. Doksum and J.L. Hodges, eds., *A Festschrift for Erich L. Lehmann*. Belmont, CA: Woodsworth International Group.
Buchinsky, M. (1991a) "A Monte Carlo Study of the Asymptotic Covariance Estimators for Quantile Regression Coefficients", manuscript, Harvard University, January.
Buchinsky, M. (1991b) "Changes in the U.S. Wage Structure 1963–1987: Applications of Quantile Regression", manuscript, University of Chicago.
Buchinsky, M. (1993) "How Did Women's 'Return to Education' Evolve in the U.S.? Exploration by Quantile Regression Analysis with Nonparametric Correction for Sample Selection Bias", manuscript, Yale University.
Buckley, J. and I. James (1979) "Linear Regression with Censored Data", *Biometrika*, 66, 429–436.
Bult, J.R. (1992a) "Target Selection for Direct Marketing: Semiparametric versus Parametric Discrete Choice Models", Faculty of Economics, University of Groningen, Research Memorandum No. 468.
Bult, J.R. (1992b) "Semiparametric versus Parametric Classification Models: An Application to Direct Marketing", manuscript, University of Groningen.
Burguete, J., R. Gallant and G. Souza (1982) "On Unification of the Asymptotic Theory of Nonlinear Econometric Models", *Econometric Reviews*, 1, 151–190.
Carroll, R.J. (1982) "Adapting for Heteroskedasticity in Linear Models", *Annals of Statistics*, 10, 1224–1233.
Carroll, R.J. and D. Ruppert (1982) "Robust Estimation in Heteroskedastic Linear Models", *Annals of Statistics*, 10, 429–443.
Carroll, R.J. and D. Ruppert (1984) "Power Transformations When Fitting Theoretical Models to Data", *Journal of the American Statistical Association*, 79, 321–328.
Cavanagh, C. and R. Sherman (1991) "Rank Estimators for Monotonic Regression Models", manuscript, Bellcore.
Chamberlain, G. (1984) "Panel Data", in: Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics*, Vol. 2. Amsterdam: North-Holland.
Chamberlain, G. (1986a) "Asymptotic Efficiency in Semiparametric Models with Censoring", *Journal of Econometrics*, 32, 189–218.
Chamberlain, G. (1986b) "Notes on Semiparametric Regression", manuscript, Department of Economics, University of Wisconsin–Madison.
Chamberlain, G. (1987) "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions", *Journal of Econometrics*, 34, 305–334.
Chamberlain, G. (1990) "Quantile Regression, Censoring, and the Structure of Wages", manuscript, Harvard University.
Chamberlain, G. (1992) "Efficiency Bounds for Semiparametric Regression", *Econometrica*, 567–596.

Choi, K. (1990) "The Semiparametric Estimation of the Sample Selection Model Using Series Expansion and the Propensity Score", manuscript, University of Chicago.
Chung, C.-F. and A.S. Goldberger (1984) "Proportional Projections in Limited Dependent Variable Models", *Econometrica*, 52, 531–534.
Cosslett, S.R. (1981) "Maximum Likelihood Estimation for Choice-Based Samples", *Econometrica*, 49, 1289–1316.
Cosslett, S.R. (1983) "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model", *Econometrica*, 51, 765–782.
Cosslett, S.R. (1987) "Efficiency Bounds for Distribution-Free Estimators of the Binary Choice and the Censored Regression Models", *Econometrica*, 55, 559–587.
Cosslett, S.R. (1991) "Distribution-Free Estimator of a Regression Model with Sample Selectivity", in: W.A. Barnett, J.L. Powell and G. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge: Cambridge University Press.
Cox, D.R. (1972) "Regression Models and Life Tables", *Journal of the Royal Statistical Society*, Series B, 34, 187–220.
Cox, D.R. (1975) "Partial Likelihood", *Biometrika*, 62, 269–276.
Cragg, J.G. (1983) "More Efficient Estimation in the Presence of Heteroscedasticity of Unknown Form", *Econometrica*, 51, 751–764.
Das, S. (1991) "A Semiparametric Structural Analysis of the Idling of Cement Kilns", *Journal of Econometrics*, 50, 235–256.
Deaton, A. and M. Irish (1984) "Statistical Models for Zero Expenditures in Household Budgets", *Journal of Public Economics*, 23, 59–80.
Deaton, A. and S. Ng (1993) "Parametric and Non-parametric Approaches to Price and Tax Reform", manuscript, Princeton University.
Delgado, M.A. (1992) "Semiparametric Generalized Least Squares in the Multivariate Nonlinear Regression Model", *Econometric Theory*, 8, 203–222.
Dempster, A.P., N.M. Laird and D.B. Rubin (1977) "Maximum Likelihood from Incomplete Data via the E–M Algorithm", *Journal of the Royal Statistical Society*, Series B, 1–38.
Duncan, G.M. (1986) "A Semiparametric Censored Regression Estimator", *Journal of Econometrics*, 32, 5–34.
Elbadawi, I., A.R. Gallant and G. Souza (1983) "An Elasticity Can be Estimated Consistently Without *A Priori* Knowledge of its Functional Form", *Econometrica*, 51, 1731–1751.
Engle, R.F., C.W.J. Granger, J. Rice and A. Weiss (1986) "Semiparametric Estimates of the Relation Between Weather and Electricity Sales", *Journal of the American Statistical Association*, 81, 310–320.
Ferguson, T.S. (1967) *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press.
Fernandez, L. (1986) "Nonparametric Maximum Likelihood Estimation of Censored Regression Models", *Journal of Econometrics*, 32, 35–57.
Friedman, J.H. and W. Stuetzle (1981) "Projection Pursuit Regression", *Journal of the American Statistical Association*, 76, 817–823.
Gallant, A.R. (1980) "Explicit Estimators of Parametric Functions in Nonlinear Regression", *Journal of the American Statistical Association*, 75, 182–193.
Gallant, A.R. (1981) "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form, The Fourier Flexible Form", *Journal of Econometrics*, 15, 211–245.
Gallant, A.R. (1987) "Identification and Consistency in Nonparametric Regression", in: T.F. Bewley, ed., *Advances in Econometrics, Fifth World Congress*. Cambridge: Cambridge University Press.
Gallant, A.R. and D.W. Nychka (1987) "Semi-nonparametric Maximum Likelihood Estimation", *Econometrica*, 55, 363–390.
Goldberger, A.S. (1983) "Abnormal Selection Bias", in: S. Kărlin, T. Amemiya and L. Goodman, eds., *Studies in Econometrics, Time Series, and Multivariate Statistics*, New York: Academic Press.
Greene, W.H. (1981) "On the Asymptotic Bias of the Ordinary Least Squares Estimator of the Tobit Model", *Econometrica*, 49, 505–514.
Greene, W.H. (1983) "Estimation of Limited Dependent Variable Models by Ordinary Least Squares and the Method of Moments", *Journal of Econometrics*, 21, 195–212.
Grenander, U. (1981) *Abstract Inference*. New York: Wiley.
Hall, P. and J.L. Horowitz (1990) "Bandwidth Selection in Semiparametric Estimation of Censored Linear Regression Models", *Econometric Theory*, 6, 123–150.

Hall, P. and J.S. Marron (1987) "Estimation of Integrated Squared Density Derivatives", *Statistics and Probability Letters*, 6, 109–115.
Han, A.K. (1987a) "Non-Parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator", *Journal of Econometrics*, 35, 303–316.
Han, A.K. (1987b) "A Non-Parametric Analysis of Transformations", *Journal of Econometrics*, 35, 191–209.
Hansen, L.P. (1982) "Large Sample Properties of Generalized Method of Moment Estimators", *Econometrica*, 50, 1029–1054.
Härdle, W. (1991) *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.
Härdle, W. and T.M. Stoker (1989) "Investigating Smooth Multiple Regression by the Method of Average Derivatives", *Journal of the American Statistical Association*, forthcoming.
Härdle, W., J. Hart, J.S. Marron and A.B. Tsybakov (1992) "Bandwidth Choice for Average Derivative Estimation", *Journal of the American Statistical Association*, 87, 227–233.
Hausman, J., B.H. Hall and Z. Griliches (1984) "Econometric Models for Count Data with an Application to the Patents-R&D Relationship", *Econometrica*, 52, 909–938.
Heckman, J.J. (1976) "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models", *Annals of Economic and Social Measurement*, 5, 475–492.
Heckman, J.J. and T.E. MaCurdy (1980) "A Life-Cycle Model of Female Labor Supply", *Review of Economic Studies*, 47, 47–74.
Heckman, J.J. and B. Singer (1984) "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data", *Econometrica*, 52, 271–320.
Heckman, N.E. (1986) "Spline Smoothing in a Partly Linear Model", *Journal of the Royal Statistical Society*, Series B, 48, 244–248.
Hoeffding, W. (1948) "A Class of Statistics with Asymptotically Normal Distribution", *Annals of Mathematical Statistics*, 19, 293–325.
Honoré, B.E. (1986) "Estimation of Proportional Hazards Models in the Presence of Unobserved Heterogeneity", manuscript, University of Chicago, November.
Honoré, B.E. (1992) "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects", *Econometrica*, 60, 533–565.
Honoré, B.E. and J.L. Powell (1991) "Pairwise Difference Estimators of Linear, Censored, and Truncated Regression Models", manuscript, Department of Economics, Princeton University, November.
Honoré, B.E., E. Kyriazidou and C. Udry (1992) "Estimation of Type 3 Tobit Models Using Symmetric Trimming and Pairwise Comparisons", manuscript, Department of Economics, Northwestern University.
Horowitz, J.L. (1986) "A Distribution-Free Least Squares Estimator for Censored Linear Regression Models", *Journal of Econometrics*, 32, 59–84.
Horowitz, J.L. (1988a) "Semiparametric M-Estimation of Censored Linear Regression Models", *Advances in Econometrics*, 7, 45–83.
Horowitz, J.L. (1988b) "The Asymptotic Efficiency of Semiparametric Estimators for Censored Linear Regression Models", *Empirical Economics*, 13, 123–140.
Horowitz, J.L. (1992) "A Smoothed Maximum Score Estimator for the Binary Response Model", *Econometrica*, 60, 505–531.
Horowitz, J.L. (1993) "Semiparametric Estimation of a Work Trip Mode Choice Model", *Journal of Econometrics*, forthcoming.
Horowitz, J.L. and M. Markatou (1993) "Semiparametric Estimation of Regression Models for Panel Data", Department of Economics, University of Iowa, Working Paper No. 93–14.
Horowitz, J.L. and G. Neumann (1987) "Semiparametric Estimation of Employment Duration Models", with discussion, *Econometric Reviews*, 6, 5–40.
Hsieh, D. and C. Manski (1987) "Monte–Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression", *Annals of Statistics*, 15, 541–551.
Huber, P.J. (1967) "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions", *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 4, 221–233.
Huber, P.J. (1981) *Robust Statistics*. New York: Wiley.
Huber, P.J. (1984) "Projection Pursuit", with discussion, *Annals of Statistics*, 13, 435–525.
Hurd, M. (1979) "Estimation in Truncated Samples When There is Heteroskedasticity", *Journal of Econometrics*, 11, 247–258.

Ichimura, H. (1992) "Semiparametric Least Squares Estimation of Single Index Models", *Journal of Econometrics*, forthcoming.
Ichimura, H. and L.-F. Lee (1991) "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation", in W.A. Barnett, J.L. Powell and G. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge: Cambridge University Press.
Imbens, G.W. (1992) "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling", *Econometrica*, 60, 1187–1214.
Jaeckel, L.A. (1972) "Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals", *Annals of Mathematical Statistics*, 43, 1449–1458.
Jurečková, J. (1971) "Nonparametric Estimate of Regression Coefficients", *Annals of Mathematical Statistics*, 42, 1328–1338.
Kaplan, E.L. and P. Meier (1958) "Nonparametric Estimation from Incomplete Data", *Journal of the American Statistical Association*, 53, 457–481.
Kiefer, J. and J. Wolfowitz (1956) "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters", *Annals of Mathematical Statistics*, 27, 887–906.
Kim, J. and D. Pollard (1990) "Cube Root Asymptotics", *Annals of Statistics*, 18, 191–219.
Klein, R.W. and R.H. Spady (1993) "An Efficient Semiparametric Estimator for Discrete Choice Models", *Econometrica*, 61, 387–421.
Koenker, R. and G.S. Bassett Jr. (1978) "Regression Quantiles", *Econometrica*, 46, 33–50.
Koenker, R. and G.S. Bassett Jr. (1982) "Robust Tests for Heteroscedasticity Based on Regression Quantiles", *Econometrica*, 50, 43–61.
Koul, H., V. Suslara, and J. Van Ryzin (1981) "Regression Analysis with Randomly Right Censored Data", *Annals of Statistics*, 9, 1276–1288.
Lancaster, T. (1990) *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press.
Laplace, P.S. (1793) "Sur Quelques Points du Systems du Monde", *Memoires de l'Academie Royale des Sciences de Paris, Annee 1789*, 1–87.
Lee, L.F. (1982) "Some Approaches to the Correction of Selectivity Bias", *Review of Economic Studies*, 49, 355–372.
Lee, L.F. (1991) "Semiparametric Instrumental Variables Estimation of Simultaneous Equation Sample Selection Models", manuscript, Department of Economics, University of Minnesota.
Lee, L.F. (1992) "Semiparametric Nonlinear Least-Squares Estimation of Truncated Regression Models", *Econometric Theory*, 8, 52–94.
Lee, M.J. (1989) "Mode Regression", *Journal of Econometrics*, 42, 337–349.
Lee, M.J. (1992) "Median Regression for Ordered Discrete Response", *Journal of Econometrics*, 51, 59–77.
Lee, M.J. (1993a) "Windsorized Mean Estimator for Censored Regression Model", *Econometric Theory*, forthcoming.
Lee, M.J. (1993b) "Quadratic Mode Regression", *Journal of Econometrics*, forthcoming.
Levit, B.Y. (1975) "On the Efficiency of a Class of Nonparametric Estimates", *Theory of Probability and Its Applications*, 20, 723–740.
Li, K.C. and N. Duan (1989) "Regression Analysis Under Link Violation", *Annals of Statistics*, 17, 1009–1052.
Linton, O.B. (1991) "Second Order Approximation in Semiparametric Regression Models", manuscript, Nuffield College, Oxford University.
Linton, O.B. (1992) "Second Order Approximation in a Linear Regression with Heteroskedasticity of Unknown Form", manuscript, Nuffield College, Oxford University.
MaCurdy, T.E. (1982) "Using Information on the Moments of the Disturbance to Increase the Efficiency of Estimation", Stanford University, manuscript.
Manski, C.F. (1975) "Maximum Score Estimation of the Stochastic Utility Model of Choice", *Journal of Econometrics*, 3, 205–228.
Manski, C. (1983) "Closest Empirical Distribution Estimation", *Econometrica*, 51, 305–319.
Manski, C. (1984) "Adaptive Estimation of Nonlinear Regression Models", *Econometric Reviews*, 3, 145–194.
Manski, C.F. (1985) "Semiparametric Analysis of Discrete Response, Asymptotic Properties of the Maximum Score Estimator", *Journal of Econometrics*, 27, 205–228.
Manski, C.F. (1987) "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data", *Econometrica*, 55, 357–362.

Manski, C.F. (1988a) "Identification of Binary Response Models", *Journal of the American Statistical Association*, 83, 729–738.
Manski, C.F. (1988b) *Analog Estimation Methods in Econometrics*. New York: Chapman and Hall.
Manski, C.F. and S. Lerman (1977) "The Estimation of Choice Probabilities from Choice-Based Samples", *Econometrica*, 45, 1977–1988.
Manski, C.F. and D.F. McFadden (1981) "Alternative Estimators and Sample Designs for Discrete Choice Analysis", in: C. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge: MIT Press.
Manski, C.F. and T.S. Thompson (1986) "Operational Characteristics of Maximum Score Estimation", *Journal of Econometrics*, 32, 85–108.
McFadden, D.F. (1985) "Specification of Econometric Models", Presidential Address, Fifth World Congress of the Econometric Society.
McFadden, D.F. and A. Han (1987) "Comment on Joel Horowitz and George Neumann 'Semiparametric Estimation of Employment Duration Models' ", *Econometric Reviews*, 6, 257–270.
Melenberg, B. and A. Van Soest (1993) "Semi-parametric Estimation of the Sample Selection Model", manuscript, Department of Econometrics, Tilburg University.
Meyer, B (1987) "Semiparametric Estimation of Duration Models", Ph.D. dissertation, Department of Economics, MIT.
Moon, C.-G. (1989) "A Monte Carlo Comparison of Semiparametric Tobit Estimators", *Journal of Applied Econometrics*, 4, 361–382.
Mroz, T.A. (1987) "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions", *Econometrica*, 55, 765–799.
Nawata, K. (1990) "Robust Estimation Based on Grouped-Adjusted Data in Censored Regression Models," *Journal of Econometrics*, 43, 337–362.
Nawata, K. (1992) "Semiparametric Estimation of Binary Choice Models Based on Medians of Grouped Data", University of Tokyo, manuscript:
Newey, W.K. (1984) "Nearly Efficient Moment Restriction Estimation of Regression Models with Nonnormal Disturbances", Princeton University, Econometric Research Program Memo. No. 315.
Newey, W.K. (1985) "Semiparametric Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables", *Annales de l'Insee*, 59/60, 219–236.
Newey, W.K. (1987a) "Efficient Estimation of Models with Conditional Moment Restrictions", Princeton University, manuscript.
Newey, W.K. (1987b) "Interval Moment Estimation of the Truncated Regression Model", manuscript, Department of Economics, Princeton University, June.
Newey, W.K. (1987c) "Specification Tests for Distributional Assumptions in the Tobit Model", *Journal of Econometrics*, 34, 125–145.
Newey, W.K. (1988a) "Adaptive Estimation of Regression Models Via Moment Restrictions", *Journal of Econometrics*, 38, 301–339.
Newey, W.K. (1988b) "Efficient Estimation of Semiparametric Models Via Moment Restrictions", Princeton University, manuscript.
Newey, W.K. (1988c) "Two-Step Series Estimation of Sample Selection Models", Princeton University, manuscript.
Newey, W.K. (1989a) "Efficient Estimation of Tobit Models Under Symmetry", in: W.A. Barnett, J.L. Powell and G. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge: Cambridge University Press.
Newey, W.K. (1989b) "Efficiency in Univariate Limited Dependent Variable Models Under Conditional Moment Restrictions", Princeton University, manuscript.
Newey, W.K. (1989c) "Efficient Instrumental Variables Estimation of Nonlinear Models", mimeo, Princeton University.
Newey, W.K. (1989d) "Uniform Convergence in Probability and Uniform Stochastic Equicontinuity", mimeo, Department of Economics, Princeton University.
Newey, W.K. (1990a) "Semiparametric Efficiency Bounds", *Journal of Applied Econometrics*, 5, 99–135.
Newey, W.K. (1990b) "Efficient Instrumental Variables Estimation of Nonlinear Models", *Econometrica*, 58, 809–837.
Newey, W.K. (1991) "The Asymptotic Variance of Semiparametric Estimators", Working Paper No. 583, Department of Economics, MIT, revised July.

Newey, W.K. and J.L. Powell (1990) "Efficient Estimation of Linear and Type I Censored Regression Models Under Conditional Quantile Restrictions", *Econometric Theory*, 6: 295–317.
Newey, W.K. and J.L. Powell (1991) "Two-Step Estimation, Optimal Moment Conditions, and Sample Selection Models", manuscript, Department of Economics, MIT, October.
Newey, W.K. and J.L. Powell (1993) "Efficiency Bounds for Some Semiparametric Selection Models", *Journal of Econometrics*, forthcoming.
Newey, W.K. and P. Ruud (1991) "Density Weighted Least Squares Estimation", manuscript, Department of Economics, MIT.
Newey, W.K. and T. Stoker (1989) "Efficiency Properties of Average Derivative Estimators", manuscript, Sloan School of Management, MIT.
Newey, W.K. and T.M. Stoker (1993) "Efficiency of Weighted Average Derivative Estimators and Index Models", *Econometrica*, 61, 1199–1223.
Newey, W.K., J.L. Powell and J.M. Walker (1990) "Semiparametric Estimation of Selection Models: Some Empirical Results", *American Economic Review Papers and Proceedings*, 80, 324–328.
Neyman, J. and E.L. Scott (1948) "Consistent Estimates Based on Partially Consistent Observations", *Econometrica*, 16, 1–32.
Nolan, D. and D. Pollard (1987) "U-Processes, Rates of Convergence", *Annals of Statistics*, 15, 780–799.
Nolan, D. and D. Pollard (1988) "Functional Central Limit Theorems for U-Processes", *Annals of Probability*, 16, 1291–1298.
Oakes, D. (1981) "Survival Times: Aspects of Partial Likelihood", *International Statistical Review*, 49, 235–264.
Obenhofer, W. (1982) "The Consistency of Nonlinear Regression Minimizing the L1 Norm", *Annals of Statistics*, 10, 316–319.
Pakes, A. and D. Pollard (1989) "Simulation and the Asymptotics of Optimization Estimators", *Econometrica*, 57, 1027–1058.
Pollard, D. (1985) "New Ways to Prove Central Limit Theorems", *Econometric Theory*, 1, 295–314.
Powell, J.L. (1983) "The Asymptotic Normality of Two-Stage Least Absolute Deviations Estimators", *Econometrica*, 51, 1569–1575.
Powell, J.L. (1984) "Least Absolute Deviations Estimation for the Censored Regression Model", *Journal of Econometrics*, 25, 303–325.
Powell, J.L. (1986a) "Censored Regression Quantiles", *Journal of Econometrics*, 32, 143–155.
Powell, J.L. (1986b) "Symmetrically Trimmed Least Squares Estimation of Tobit Models", *Econometrica*, 54, 1435–1460.
Powell, J.L. (1987) "Semiparametric Estimation of Bivariate Latent Variable Models", Social Systems Research Institute, University of Wisconsin–Madison, Working Paper No. 8704.
Powell, J.L. (1991) "Estimation of Monotonic Regression Models Under Quantile Restrictions", in: W.A. Barnett, J.L. Powell and G. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge: Cambridge University Press.
Powell, J.L. and T.M. Stoker (1991) "Optimal Bandwidth Choice for Density-Weighted Averages", manuscript, Department of Economics, Princeton University, December.
Powell, J.L., J.H. Stock and T.M. Stoker (1989) "Semiparametric Estimation of Weighted Average Derivatives", *Econometrica*, 57, 1403–1430.
Prakasa Rao, B.L.S. (1983) *Nonparametric Functional Estimation*. New York: Academic Press.
Rice, J. (1986) "Convergence Rates for Partially Splined Estimates", *Statistics and Probability Letters*, 4, 203–208.
Rilstone, P. (1989) "Semiparametric Estimation of Missing Data Models", mimeo, Department of Economics, Laval University.
Ritov, Y. (1990) "Estimation in a Linear Regression Model with Censored Data", *Annals of Statistics*, 18, 303–328.
Robinson, P. (1987) "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form", *Econometrica*, 55, 875–891.
Robinson, P. (1988a) "Semiparametric Econometrics, A Survey", *Journal of Applied Econometrics*, 3, 35–51.
Robinson, P. (1988b) "Root-N-Consistent Semiparametric Regression", *Econometrica*, 56, 931–954.
Rosenbaum, P.R. and D.B. Rubin (1983) "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41–55.

Ruud, P. (1983) "Sufficient Conditions for Consistency of Maximum Likelihood Estimation Despite Misspecification of Distribution", *Econometrica*, 51, 225–228.
Ruud, P. (1986) "Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution", *Journal of Econometrics*, 32, 157–187.
Schick, A. (1986) "On Asymptotically Efficient Estimation in Semiparametric Models", *Annals of Statistics*, 14, 1139–1151.
Serfling, R.J. (1980) *Approximation Theorems of Mathematical Statistics*, New York: Wiley.
Severini, T.A. and W.H. Wong (1987a) "Profile Likelihood and Semiparametric Models", manuscript, University of Chicago.
Severini, T.A. and W.H. Wong (1987b) "Convergence Rates of Maximum Likelihood and Related Estimates in General Parameter Spaces", Technical Report No. 207, Department of Statistics, University of Chicago, Chicago, IL.
Sherman, R.P. (1990a) "The Limiting Distribution of the Maximum Rank Correlation Estimator", manuscript, Bell Communications Research.
Sherman, R.P. (1990b) "Maximal Inequalities for Degenerate U-Processes with Applications to Optimization Estimators", manuscript, Bell Communications Research.
Sherman, R.P. (1993) "The Limiting Distribution of the Maximum Rank Correlation Estimator", *Econometrica*, 61, 123–137.
Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
Stein, C. (1956) "Efficient Nonparametric Testing and Estimation", *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley, University of California Press.
Stock, J.H. (1989) "Nonparametric Policy Analysis", *Journal of the American Statistical Association*, 84, 1461–1481.
Stoker, T.M. (1986) "Consistent Estimation of Scaled Coefficients", *Econometrica*, 54, 1461–1481.
Stoker, T.M. (1991) "Equivalence of Direct, Indirect, and Slope Estimators of Average Derivatives", in: W.A. Barnett, J.L. Powell and G. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge: Cambridge University Press.
Stoker, T.M. (1992) *Lectures on Semiparametric Econometrics*. Louvain–La–Neuve, Belgium: CORE Lecture Series.
Thompson, T.S. (1989a) "Identification of Semiparametric Discrete Choice Models", manuscript, Department of Economics, University of Minnesota.
Thompson, T.S. (1989b) "Least Squares Estimation of Semiparametric Discrete Choice Models," manuscript, Department of Economics, University of Minnesota.
Tobin, J. (1956) "Estimation of Relationships for Limited Dependent Variables", *Econometrica*, 26, 24–36.
Wahba, G. (1984) "Partial Spline Models for the Semiparametric Estimation of Functions of Several Variables", in *Statistical Analysis of Time Series*. Tokyo, Institute of Statistical Mathematics.
White, H. (1982) "Maximum Likelihood Estimation of Misspecified Models", *Econometrica*, 50, 1–26.
Ying, Z., S.H. Jung and L.J. Wei (1991) "Survival Analysis with Median Regression Models", manuscript, Department of Statistics, University of Illinois.
Zheng, Z. (1992) "Efficiency Bounds for the Binary Choice and Sample Selection Models under Symmetry", in *Topics in Nonparametric and Semiparametric Analysis*, Ph.D. dissertation, Princeton University.