Multiplicative-error models with sample selection

Koen Jochmans†

Department of Economics, Sciences Po, Paris

[Revised July 17, 2014]

Abstract. This paper presents a simple approach to deal with sample selection in models with multiplicative errors. Models for non-negative limited dependent variables such as counts fit this framework. The approach builds on a specification of the conditional mean of the outcome only and is, therefore, semiparametric in nature. GMM estimators are constructed for both cross-section data and for panel data. We derive distribution theory and present Monte Carlo evidence on the finite-sample performance of the estimators.

Keywords: count data; fixed-effect model; nonlinear model; sample selection; semiparametric inference; two-stage estimation

1. Introduction

The detrimental effects of estimating economic models from non-randomly selected samples on statistical inference are well known. While the issue has received a substantial amount of attention in the literature, the proposed solutions have been confined mostly to the linear regression model; Gronau (1973) and Heckman (1974, 1978, 1979) have provided seminal contributions, and Vella (1998) provides an overview. However, sample selection is no less of a problem with nonlinear specifications, and the literature has been rather slow in devising flexible approaches to inference for such situations. Of course, one may consider taking a full-information approach to inference, but specifying the full likelihood will require some tedious choices on the distribution of unobservables, and leads to estimators that are difficult to compute. This paper discusses relatively simple procedures to estimate nonlinear models with an additive- or multiplicative-error structure when the data is subject to sample selection.

One leading setting for which our developments are relevant is models for count data. Such models are widely used in economics, for example when studying the demand for health care (Cameron, Trivedi, Milne, and Piggott 1988) and the economics of innovation (Hausman, Hall, and Griliches 1984). Nonetheless, only Terza (1998) and Winkelmann (1998) have given some attention to the issue of sample selection with cross-sectional count data, and their approaches require a tight specification of the distribution of unobservables. In contrast, the approach advocated in this paper is semiparametric, in the sense that it

†*Address for correspondence:* Sciences Po, Department of Economics, 28 rue des Saints-Pères, 75007 Paris, France. *E-mail:* koen.jochmans@sciencespo.fr.

avoids a full specification of the distribution of unobservables, and is applicable to both cross-section data and short panel data.

More generally, this paper fits into the literature on estimation using control functions (Heckman and Robb 1985; Blundell and Powell 2003). Moreover, in the cross-sectional case, our estimator can be seen as a generalization of the estimator of Powell (1987, 2001) for linear sample-selection models. Blundell and Powell (2004) have developed a related generalization of Powell (1987, 2001) when considering endogeneity in cross-sectional binary-choice models. However, the setup they consider and the one entertained here are not nested. In addition, our approach avoids the need for nonparametric estimation of auxiliary parameters and the associated problem of selecting trimming parameters, which should be beneficial for its small-sample performance. Honoré and Powell (2005) and Jochmans (2013) contain other related work, albeit in different settings.

Another nice feature of our strategy compared to the work of Blundell and Powell (2004) is that it can be extended to fixed-effect specifications for panel data. Estimation of models with group-specific nuisance parameters is well-known to be problematic under asymptotics where the number of groups grows large while the number of observations per group remains fixed; see, e.g., Arellano and Honoré (2001). It is therefore not surprising that, besides focusing exclusively on linear models, the literature has favored a random-effect approach to inference in such cases; see, e.g., Verbeek and Nijman (1992) and Wooldridge (1995). The only fixed-effect estimator of linear sample-selection models is the one developed by Kyriazidou (1997, 2001). Her estimator is a panel-data version of Powell (1987, 2001) and, indeed, the estimator discussed here can be interpreted as the corresponding nonlinear version.

The estimators introduced in this paper are GMM estimators. They are constructed from local moment conditions that are inspired by a differencing argument introduced by Chamberlain (1992) in a different context. Under rather conventional assumptions that we spell out, these estimators are consistent and have a Gaussian limit distribution, with a variance that can be consistently estimated. To evaluate their small-sample performance we use Monte Carlo experiments to estimate count data models with sample selection. The simulations show that the techniques advocated here can provide reliable inference, even in small samples.

2. A semiparametric approach for cross-section data

2.1. The model and moment conditions

2.1.1. The model

For an integer n and i.i.d. random variables $\{y_i, x_i, u_i\}_{i=1}^n$, consider the conditional-mean model

$$\mathcal{E}[y_i|x_i, u_i] = \mu(x_i; \alpha_0) + \varphi(x_i; \beta_0)\, u_i, \tag{2.1}$$

where μ and φ are functions that are known up to the Euclidean parameter $\theta_0 \equiv (\alpha_0', \beta_0')'$. Our aim will be to infer θ_0 from a sample into which observations have self-selected. The

selection process is modelled as a threshold-crossing model for the binary selection indicator s_i, with propensity score

$$\Pr[s_i = 1|p_i] = \mathcal{E}\left[1\{p_i \geq v_i\}|\, p_i\right], \tag{2.2}$$

for $p_i = p(z_i)$ an estimable aggregator mapping observables z_i to the real line; x_i and z_i need not be disjoint, although identification will require some of the variables in z_i to be excluded from x_i. We view (u_i, v_i) as representing unobserved heterogeneity that jointly influence (y_i, s_i). These latent factors are taken to be independent of the observable characteristics (x_i, z_i), but not necessarily of each other. The sample-selection problem, then, is to perform inference on θ_0 from a random sample in which realizations of (s_i, x_i, z_i) are always observed but realizations of y_i are observed only when $s_i = 1$.[1]

Before proceeding we note that our general specification covers several models of special interest. Nonlinear models with additive unobservables, for example, can be represented as

$$\mathcal{E}[y_i|x_i, u_i] = \mu(x_i; \alpha_0) + u_i.$$

Even for such prevalent models, no off-the-shelf approach to estimation has been discussed in the literature on sample selection. Models with multiplicative unobservables are also covered. For non-negative limited dependent variables such models can be written as

$$\mathcal{E}[y_i|x_i, u_i] = \varphi(x_i; \beta_0)\, u_i,$$

where $u_i \geq 0$ and φ maps to the positive real half-line. A prototypical specification for count data—such as the Poisson and the negative binomial models, for example—would have

$$\varphi(x_i; \beta) = \exp(x_i'\beta).$$

However, we also note that exponential-regression models of this type also arise frequently in a variety of other settings. One important example is the estimation of trade flows; see Santos Silva and Tenreyro (2006). A binary-choice model where $\varphi(x_i; \beta) = G(x_i'\beta)$ for some distribution function G is equally covered here. Wooldridge (1997, Example 4.2) provides an empirical motivation for such a specification. Again, estimation of such models has received only very limited attention.

Because sampling will not provide information on the distribution of y_i given $s_i = 0$, sample selection complicates inference on θ_0. To see how the problem manifests itself, observe that

$$\mathcal{E}^*[y_i|x_i, z_i] = \mu(x_i; \alpha_0) + \varphi(x_i; \beta_0)\ \mathcal{E}^*[u_i|x_i, z_i],$$

where $\mathcal{E}^*$ refers to an expectation concerning the subpopulation for which $s_i = 1$. Given the threshold-crossing structure of the selection rule and independence of (u_i, v_i) from the observables covariates, we can further dissect the influence of sample selection by using

$$\lambda(p_i) \equiv \frac{\int_{-\infty}^{p_i}\int_{-\infty}^{+\infty} u\, f(u, v)\, \mathrm{d}u\, \mathrm{d}v}{F_v(p_i)} = \mathcal{E}^*[u_i|\, p_i], \tag{2.3}$$

[1] The analysis can easily be extended to situations in which x_i, too, is only observed when $s_i = 1$.

where f denotes the joint density of (u_i, v_i) and F_v is the marginal distribution of v_i. Indeed,

$$\mathcal{E}^*[y_i|x_i, p_i] = \mu(x_i; \alpha_0) + \varphi(x_i; \beta_0)\, \lambda(p_i). \tag{2.4}$$

If u_i and v_i are independent, λ is constant. Otherwise, λ depends on the data through the index driving the propensity score of selection. In either case, $\mathcal{E}[y_i|x_i] \neq \mathcal{E}^*[y_i|x_i]$, in general. This implies that an estimation strategy that does not account for sample selection will typically suffer from misspecification bias in the sense of Heckman (1979). Note that this result is different from the linear model, where ignoring exogenous sample selection would only affect the estimation of the intercept term. Furthermore,

$$\lambda(p_i) = -\rho\sigma_u \frac{\phi(-p_i/\sigma_v)}{1 - \Phi(-p_i/\sigma_v)} \qquad \text{if} \qquad \begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix} \right),$$

which is the correction term originally derived by Heckman (1979) in the context of the linear model. However, (2.3)–(2.4) hold without restricting f to belong to a known parametric family of density functions. They also stretch beyond the conventional linear specification, and so may be of use to construct semiparametric inference techniques for models with multiplicative errors.

2.1.2. Moment conditions and identification

Our approach is inspired by the work of Powell (1987, 2001) on pairwise differencing and builds on moment conditions that are similar in spirit to those of Chamberlain (1992) and Wooldridge (1997) in a different context. Because

$$\mathcal{E}^*\left[\tau_i(\theta_0)|x_i, p_i\right] = \lambda(p_i), \qquad \tau_i(\theta) \equiv \frac{y_i - \mu(x_i; \alpha)}{\varphi(x_i; \beta)},$$

follows from re-arranging (2.4), we have that

$$\mathcal{E}^*[\tau_i(\theta_0)|x_i, p_i] - \mathcal{E}^*[\tau_j(\theta_0)|x_j, p_j] = \lambda(p_i) - \lambda(p_j)$$

for any pair i, j. If λ is a smooth function, the right-hand side of this expression will converge to zero as $|p_i - p_j| \to 0$. This suggests an approach based on moment conditions that difference-out the selection bias. Let $\Delta_{ij} \equiv p_i - p_j$. If

$$|\lambda(p_i) - \lambda(p_j)| \leq \Lambda(p_i, p_j)\, |\Delta_{ij}|$$

for some function Λ, then

$$\left|\mathcal{E}^*\left[\tau_i(\theta_0) - \tau_j(\theta_0)|x_i, x_j, \Delta_{ij}\right]\right| \leq \mathcal{E}^*[\Lambda(p_i, p_j)|x_i, x_j, \Delta_{ij}]\, |\Delta_{ij}| \xrightarrow{|\Delta_{ij}| \downarrow 0} 0,$$

provided that $\mathcal{E}^*[\Lambda(p_i, p_j)|x_i, x_j, \Delta_{ij}]$ exists for Δ_{ij} in a neighborhood of zero. This is a fairly weak condition. For example, if λ is everywhere differentiable with derivative λ' we can take

$$\Lambda(p_i, p_j) = \sup_{p \in [p_i, p_j]} |\lambda'(p)|,$$

and a tail condition on this quantity allows the use of a dominated-convergence argument to establish the existence of the expectation. If λ' is also continuous, then λ is locally Lipschitz and, therefore, locally bounded. This would equally imply the required condition to hold for sufficiently small Δ_{ij}. By Leibniz's rule,

$$\lambda'(p_i) = \big(\mathcal{E}[u_i|v_i = p_i] - \lambda(p_i)\big)\, r(p_i),$$

where $r(p_i) \equiv f_v(p_i)/F_v(p_i)$ is the inverse Mills ratio of v_i, and so λ' will be locally bounded if $\mathcal{E}[u_i|v_i = p_i]$, $\lambda(p_i)$, and $r(p_i)$ are continuous. Although routinely done, demanding λ' to be uniformly bounded is somewhat too strong a requirement, particularly if p_i can take on values on the whole real line. For example, when (u_i, v_i) are jointly normal,

$$\lambda'(p_i) = \left(\rho\frac{\sigma_u}{\sigma_v}p_i + \rho\sigma_u\frac{\phi(-p_i/\sigma_v)/\sigma_v}{1-\Phi(-p_i/\sigma_v)}\right)\frac{\phi(-p_i/\sigma_v)/\sigma_v}{1-\Phi(-p_i/\sigma_v)}.$$

Here, the magnitude of all terms increases without bound when $p_i \to -\infty$ and, indeed, $\lambda(p_i)$ itself diverges in this case. A similar pattern arises in the more general case of log-concave symmetric densities, for which the inverse Mills ratio is known to be monotonically decreasing in p_i.

The above argument suggests a strategy based on unconditional moment conditions in which more weight is assigned to pairs of observations for which Δ_{ij} lies in a shrinking neighborhood of zero. More precisely, for some suitable transformation function ω, let $\omega(x_i, x_j)$ denote instrumental variables.[2] Let κ be a chosen symmetric kernel function and let ς be an associated bandwidth. Then, under conventional regularity conditions introduced in Assumptions 2.3 and 2.4 below,

$$\mathcal{E}^*\left[\frac{\omega(x_i, x_j)\,(\tau_i(\theta_0) - \tau_j(\theta_0))}{\varsigma}\kappa\left(\frac{\Delta_{ij}}{\varsigma}\right)\right] \xrightarrow{|\varsigma|\downarrow 0} 0. \tag{2.5}$$

Overidentification is allowed for in the sense that the number of moments, say m, can exceed $\dim\theta$. A sample counterpart of these moment conditions is readily constructed using any of a number of available first-stage estimators of the selection equation, and can be combined easily through a GMM procedure to yield a two-step semiparametric estimator of θ_0. This estimator will be the topic of the next subsection.

Because our approach relies on pairwise differencing, the parameters associated with regressors that are constant across observations will not be identified from (2.5). The leading example would be an intercept in $\mu(x_i;\alpha)$. This is not surprising. Even in the classical linear sample-selection model, recovering the intercept term in a semiparametric fashion requires an argument involving identification at infinity, and yields estimators with non-standard properties (Andrews and Schafgans 1998).

[2] The analysis to follow can be extended to allow ω to depend on θ. We could equally extend the definition of ω by allowing it to depend on variables other than the covariates x_i, which we do not do here for notational simplicity.

Deriving sufficient conditions for global identification is difficult in models that are specified by a set of nonlinear moment conditions (see, e.g., Hall 2005, Section 3.1, for a discussion). The general model entertained here is a case in point. However, because our analysis is based on moment conditions that are conditional on the difference between the p_i, it is intuitive that we will need sufficient variation in the x_i for given values of p_i. For example, in the standard linear model, where $\mu(x_i;\alpha) = x_i'\alpha$ and we set $\omega(x_i, x_j) = x_i - x_j$, it can be verified using (2.8) below that, if

$$\operatorname{rank} \mathcal{E}^*[\operatorname{var}^*(x_i|p_i)\, f^*(p_i)] = \dim \alpha,$$

then α_0 is globally identified. Here, f^* denotes the density of p_i given selection and $\operatorname{var}^*(x_i|p_i)$ is the conditional variance of x_i given p_i in the selected subpopulation. Therefore, as usual in control-function problems, identification requires the presence of instrumental variables in the selection equation.

Local identification is easier to study. For example, in the general nonlinear model with additive unobservables, again with $\omega(x_i, x_j) = x_i - x_j$, local identification is achieved when

$$\operatorname{rank} \mathcal{E}^*[\operatorname{cov}^*(x_i, \mu'(x_i;\alpha_0)|\, p_i)\, f^*(p_i)] = \dim \alpha,$$

where μ' is the first-derivative vector of μ with respect to α. Of course, in the linear case, this boils down to the rank condition given earlier. As an example of a model with multiplicative errors, consider the exponential regression model from above. In this case, use of (2.8) shows that, if

$$\operatorname{rank} \mathcal{E}^*[\lambda(p_i)\operatorname{var}^*(x_i|p_i)\, f^*(p_i)] = \dim \beta,$$

then β_0 is locally identified. Furthermore, because local identification is equivalent to the Jacobian of the population moments having full rank at θ_0, it can be empirically tested by applying any of a battery of rank tests to a plug-in version of this matrix (see, e.g., Kleibergen and Paap 2006).

2.2. Estimation

For conciseness I will work with a linear-index specification for the propensity score, that is, I set $p_i \equiv p(z_i) = z_i'\gamma_0$ for an unknown finite-dimensional parameter value γ_0. Flexible specifications of this form that include power transforms and interaction terms between regressors are common practice in empirical work. The distribution theory below could be extended to allow for a nonparametric selection rule at the cost of stronger smoothness requirements and more cumbersome notation.

Without loss of generality, take $\omega(x_i, x_j)$ to be antisymmetric in its arguments. Then a feasible empirical counterpart to the moment condition in (2.5) is

$$\widehat{q}_n(\theta) \equiv \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{i<j} \frac{\omega(x_i, x_j)\,(\tau_i(\theta) - \tau_j(\theta))}{\varsigma}\, \kappa\left(\frac{\widehat{p}_i - \widehat{p}_j}{\varsigma}\right) s_i s_j, \tag{2.6}$$

where $\widehat{p}_i \equiv z_i'\gamma_n$ and γ_n is a consistent estimator of γ_0. Our GMM estimator of θ_0 is the minimizer of a quadratic form in $\widehat{q}_n(\theta)$. It is given by

$$\theta_n \equiv \arg\min_{\theta\in\Theta} \widehat{q}_n(\theta)'\, V_n\, \widehat{q}_n(\theta)$$

for a chosen symmetric positive-definite matrix V_n of conformable dimension that serves to weight the moment conditions when $m > \dim\theta$, and a suitably defined parameter space Θ over which the minimization is performed. The interpretation of the kernel weight is immediate: pairs of observations for which $|\widehat{p}_i - \widehat{p}_j|$ is smaller receive a higher weight. Letting ς decrease with n ensures that, asymptotically, only observations for which this difference converges to zero are taken into account.

One attractive feature of estimation based on pairwise differencing is that the function λ need not be estimated. Alternative approaches to estimating θ_0 could be devised that replace λ in (2.4) by a nonparametric kernel or series estimator, and subsequently estimate θ_0 via a least-squares procedure. Such an approach would be in line with the work of Lee (2007) and Newey (2009). Contrary to the approach taken here, however, it does not generalize easily to the panel-data context. Furthermore, if an estimator of $\lambda(p)$ is desired, we may use

$$\lambda_n(p) \equiv \frac{\sum_{i=1}^n \tau_i(\theta_n)\,\kappa\left(\frac{\widehat{p}_i - p}{\varsigma}\right) s_i}{\sum_{i=1}^n \kappa\left(\frac{\widehat{p}_i - p}{\varsigma}\right) s_i},$$

for example. Of course, a series estimator would be equally well suited for this purpose. Under the conditions spelled out in Assumptions 2.1–2.5 below, θ_n will be $\sqrt{n}$-consistent. From this it follows that the asymptotic behavior of λ_n is not affected by the estimation noise in θ_n. So inference on λ using λ_n can be performed using standard tools from nonparametric conditional-mean estimation. The empirical distribution function of the $\lambda_n(\widehat{p}_i)$ may be of interest. For example, if $\lambda_n(\widehat{p}_i)$ is found to vary substantially across i, this provides evidence on the presence of sample selection.

We now state the conditions under which we will derive distribution theory for θ_n. The first condition imposes standard regularity conditions.

Assumption 2.1 (regularity). *The space Θ is compact and θ_0 lies in its interior. Equation (2.5) identifies θ_0, and μ and φ are twice continuously differentiable in θ. The distribution of p_i given selection is absolutely continuous.*

The compactness and smoothness requirements are conventional. The absolute-continuity assumption ensures f^* to be well defined. Note that this demands at least some of the components of z_i to be continuous.

The second assumption concerns the first-stage estimator.

Assumption 2.2 (first step). *The first-stage estimator, γ_n, is $\sqrt{n}$-consistent, and*

$$\sqrt{n}(\gamma_n - \gamma_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_i + o_p(1)$$

for independent and identically distributed random variables ψ_i that have zero mean and finite fourth-order moment.

Assumption 2.2 states that γ_n must satisfy an asymptotic-linearity condition. This is not very demanding, as most semiparametric candidates for γ_n do so; Powell (1994) provides a long list of eligible approaches.

Our third assumption collects standard restrictions on the kernel function and postulates permissible bandwidth choices.

ASSUMPTION 2.3 (KERNEL). *The kernel function κ is bounded and twice continuously differentiable with bounded derivatives κ' and κ'', symmetric, and integrates to one. For some integer k, $\int_{-\infty}^{+\infty} \varepsilon^h \kappa(\varepsilon)\,\mathrm{d}\varepsilon = 0$ for all $0 < h < k$, and $\int_{-\infty}^{+\infty} |\kappa(\varepsilon)|\,\mathrm{d}\varepsilon < +\infty$ and $\int_{-\infty}^{+\infty} |\varepsilon^k|\,|\kappa(\varepsilon)|\,\mathrm{d}\varepsilon < +\infty$. The bandwidth satisfies $\varsigma \propto n^{-r}$ for $\frac{1}{2k} < r < \frac{1}{6}$.*

Assumption 2.3 requires κ to be a higher-order kernel. An valid bandwidth sequence can be constructed as soon as $k > 3$. Higher-order kernels are used to ensure that the limit distribution of θ_n is free of asymptotic bias. They are easy to construct, especially given that κ is taken to be symmetric (see Li and Racine 2007, Section 1.11). Müller (1984) provides formulae to do so. For example, a fourth-order kernel based on the standard-normal density is

$$\kappa(\varepsilon) = \left(\frac{3}{2} - \frac{1}{2}\varepsilon^2\right)\phi(\varepsilon), \tag{2.7}$$

which is easily shown to satisfy the conditions in Assumption 2.3. We note that the use of a higher-order kernel is not required for consistency, and that $r < \frac{1}{4}$ would suffice for such a purpose.

The fourth assumption contains moment restrictions that are needed to ensure uniform convergence of the objective function to its large-sample counterpart, and are thus required for consistency. Let

$$\zeta_i(p_j;\theta) = \Big(\mathcal{E}^*[\omega(x_i,x_j)|x_i,p_j]\,\tau_i(\theta) + A_i(p_j;\theta) - B_i(p_j;\theta)\,\lambda(p_j)\Big)\,d_i(p_j), \tag{2.8}$$

where $d_i(p_j) \equiv s_i\, f^*(p_j)\, \Pr[s_i = 1]$ and

$$A_i(p_j;\theta) \equiv \mathcal{E}^*\left[\omega(x_i,x_j)\,\frac{\mu(x_j;\alpha) - \mu(x_j;\alpha_0)}{\varphi(x_j;\beta)}\,\middle|\, x_i, p_j\right],$$
$$B_i(p_j;\theta) \equiv \mathcal{E}^*\left[\omega(x_i,x_j)\,\frac{\varphi(x_j;\beta_0)}{\varphi(x_j;\beta)}\,\middle|\, x_i, p_j\right].$$

Also, let τ' denote the first derivative of τ with respect to θ. Denote the Euclidean norm and the Frobenius norm by $\|\cdot\|$.

ASSUMPTION 2.4 (FINITE MOMENTS). *For each $\theta \in \Theta$, $\mathcal{E}^*[\tau_i(\theta)^8]$ and $\mathcal{E}^*[\|\tau_i'(\theta)\|^4]$ are finite. Both $\mathcal{E}[\|\omega(x_i,x_j)\|^8]$ and $\mathcal{E}[\|z_i\|^4]$ are finite. The function $\zeta_i(p;\theta)$ is continuous in p and $\mathcal{E}[\sup_p \|\zeta_i(p;\theta)\|]$ is finite for each $\theta \in \Theta$.*

The dominance condition on $\|\zeta_i(p;\theta)\|$ is standard in nonparametric estimation. Note from the form of $\zeta_i(p;\theta)$ that it can be interpreted as a restriction on its tail behavior. This condition is needed for convergence of the kernel-weighted objective function as n diverges; see, e.g., Hansen (2008) for the application of such conditions in generic problems.

Introduce

$$\zeta_i'(p;\theta) \equiv \frac{\partial \zeta_i(p;\theta)}{\partial \theta'}, \qquad \nabla_h \zeta_i(p;\theta) \equiv \frac{\partial^h \zeta_i(p;\theta)}{\partial p^h}.$$

The fifth assumption is used to derive the limit distribution of $\sqrt{n}(\theta_n - \theta_0)$.

ASSUMPTION 2.5 (SMOOTHNESS). *For each $\theta \in \Theta$, $\mathcal{E}^*[\|\tau_i''(\theta)\|^4]$ is finite. For each $\theta \in \Theta$, the function $\zeta_i'(p;\theta)$ is continuous in p and $\mathcal{E}[\sup_p \|\zeta_i'(p;\theta)\|]$ is finite. $\nabla_h \zeta_i(p;\theta_0)$ exists and $\mathcal{E}[\sup_p \|\nabla_h \zeta_i(p;\theta_0)\|^2]$ is finite for all integers $h \leq k+1$.*

The conditions on τ'' and $\zeta_i'(p;\theta)$ are needed to establish convergence of the Jacobian of the moment conditions uniformly on Θ. The higher-order smoothness requirements are used to ensure the limit distribution of $\sqrt{n}(\theta_n - \theta_0)$ to be free of asymptotic bias. Such an approach to bias control is common in inference problems of this type. To interpret these restrictions, observe that

$$\zeta_i(p_j;\theta_0) = \mathcal{E}^*[\omega(x_i,x_j)|x_i,p_j]\,[\tau_i(\theta_0) - \lambda(p_j)]\,d_i(p_j).$$

Assumption 2.5 then requires that $\mathcal{E}^*[\omega(x_i,x_j)|x_i,p_j]$, $f^*(p_j)$, and $\lambda(p_j)$ are at least five times differentiable and also restricts the tail behavior of these quantities, as well as of their respective derivatives.

To state the limit distribution of the estimator, let $Q_0 \equiv \mathcal{E}[\zeta_i'(p_i;\theta_0)]$ and introduce

$$\sigma_i \equiv \zeta_i(p_i;\theta_0) + H_0\,\psi_i$$

for $H_0 \equiv -\mathcal{E}[\nabla_1 \zeta_j(p_j;\theta_0)\,z_j']$. Note that $\mathcal{E}[\sigma_i] = 0$ and that $\Sigma \equiv \mathcal{E}[\sigma_i \sigma_i'] < +\infty$.

THEOREM 2.1 (ASYMPTOTIC DISTRIBUTION). *Let Assumptions 2.1–2.5 hold. Suppose that Σ is positive definite, that Q_0 has full column rank, and that $V_n \xrightarrow{P} V_0$ for V_0 positive definite. Then $\sqrt{n}\|\theta_n - \theta_0\| = O_P(1)$ and*

$$\sqrt{n}(\theta_n - \theta_0) \xrightarrow{L} \mathcal{N}(0,\Upsilon), \quad \Upsilon \equiv 4(Q_0'V_0Q_0)^{-1}\,(Q_0'V_0\Sigma V_0Q_0)\,(Q_0'V_0Q_0)^{-1}.$$

In particular, if $V_n \xrightarrow{P} \Sigma^{-1}$, then $\Upsilon = 4(Q_0'\Sigma^{-1}Q_0)^{-1}$.

Choosing V_n so that $V_n \xrightarrow{P} \Sigma^{-1}$ is well known to be optimal in terms of asymptotic efficiency for a given set of moment conditions (Hansen 1982).

An estimator of the asymptotic variance matrix in Theorem 2.1 is needed to perform inference. An estimator of V_0 is available from the outset in the form of V_n. Estimators of Q_0 and Σ can be constructed via the plug-in principle. Moreover,

$$Q_n \equiv \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{i<j} \frac{\omega(x_i,x_j)\,(\tau_i'(\theta_n) - \tau_j'(\theta_n))'}{\varsigma}\,\kappa\left(\frac{\widehat{p}_i - \widehat{p}_j}{\varsigma}\right) s_i s_j,$$

$$H_n \equiv \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{i<j} \frac{\omega(x_i,x_j)\,(\tau_i(\theta_n) - \tau_j(\theta_n))}{\varsigma}\,\kappa'\left(\frac{\widehat{p}_i - \widehat{p}_j}{\varsigma}\right) \frac{(z_i - z_j)'}{\varsigma}\, s_i s_j,$$

constitute consistent estimators of the matrices Q_0 and H_0, respectively. An estimator of Σ then is $\Sigma_n \equiv n^{-1}\sum_{i=1}^n \widehat{\sigma}_i\widehat{\sigma}_i'$ for

$$\widehat{\sigma}_i \equiv \widehat{\zeta}_i + H_n\,\widehat{\psi}_i, \qquad \widehat{\zeta}_i \equiv \frac{1}{n-1}\sum_{j\neq i}\frac{\omega(x_i,x_j)\,(\tau_i(\theta_n)-\tau_j(\theta_n))}{\varsigma}\,\kappa\left(\frac{\widehat{p}_i-\widehat{p}_j}{\varsigma}\right)s_i s_j,$$

where $\widehat{\psi}_i$ is an estimator of the influence function of γ_n. The precise form of this estimator will depend on the first-stage estimator used. Impose the following condition.

ASSUMPTION 2.6. $n^{-1}\sum_{i=1}^n\|\widehat{\psi}_i-\psi_i\|^2 = o_P(1)$.

This assumption ensures that the asymptotic variance of the first-stage estimator can be consistently estimated.

The following theorem permits the construction of asymptotically-valid test procedures based on the Wald principle.

THEOREM 2.2 (INFERENCE). *Let Assumptions 2.1–2.6 hold. Suppose that Σ is positive definite, that Q_0 has full column rank, and that $V_n \xrightarrow{P} V_0$ for V_n and V_0 positive definite. Then*

$$\Upsilon_n \xrightarrow{P} \Upsilon, \qquad \Upsilon_n \equiv 4(Q_n'V_nQ_n)^{-1}\,(Q_n'V_n\Sigma_nV_nQ_n)\,(Q_n'V_nQ_n)^{-1}.$$

Other consequences of Theorem 2.2 are the feasibility of the two-step GMM estimator and the fact that

$$n\,\widehat{q}_n(\theta_n)'\Sigma_n^{-1}\widehat{q}_n(\theta_n) \xrightarrow{P} \chi^2_{m-\dim\theta}$$

for the optimally-weighted estimator. This justifies the asymptotic validity of the usual overidentification tests (Sargan 1958; Hansen 1982).

We note that, under the conditions stated in Lemma 3.4 and Lemma 3.5 of Pakes and Pollard (1989), all results so far continue to hold for the continuously-updated version of θ_n (see Hansen, Heaton, and Yaron 1996).

2.3. Simulations

We next discuss the results of a Monte Carlo experiment where y_i given (x_i,u_i) is a Poisson variate with mean

$$\mathcal{E}[y_i|x_i,u_i] = \exp(c+x_i\beta_0)\,u_i,$$

where c is a constant. Let $w_i \equiv \log u_i$. We generate

$$\begin{pmatrix} w_i \\ v_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_w^2 & \rho\sigma_w \\ \rho\sigma_w & 1 \end{pmatrix}\right)$$

for $|\rho|<1$, so that the marginal distribution of u_i is log-normal and the selection equation is a conventional probit. The unconditional mean of u_i equals $\exp(\sigma_w^2/2)$. We set $c=-\sigma_w^2/2$ to recenter its distribution at one. Then

$$\mathcal{E}^*[y_i|x_i,p_i] = \exp(x_i\beta_0)\,\lambda(p_i), \qquad \lambda(p_i) = \frac{\Phi(-\rho\sigma_w+p_i)}{\Phi(p_i)};$$

see also Terza (1998). For reasons of parsimony, we set $p_i = x_i + \gamma_0 a_i$ and draw (x_i, a_i) as

$$\begin{pmatrix} x_i \\ a_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \varrho\sigma_x\sigma_a \\ \varrho\sigma_x\sigma_a & 1 \end{pmatrix} \right)$$

for $|\varrho| < 1$. This specification fixes the coefficient associated with x_i to unity, which is a convenient normalization for our first-stage estimator. It also allows us to get a closed-form expression for $\mathcal{E}^*[\lambda(p_i)|x_i]$. Let

$$\mu_p \equiv \frac{-\rho\sigma_w}{\sqrt{1 + (|\gamma_0|\,(1-\varrho)\,\sigma_a)^2}}, \qquad \sigma_p \equiv \frac{1 + \gamma_0\varrho\,\sigma_a/\sigma_x}{\sqrt{1 + (|\gamma_0|\,(1-\varrho)\,\sigma_a)^2}}.$$

Then, after some algebra, we find that

$$\mathcal{E}^*[y_i|x_i] = \exp(x_i\beta_0)\,\frac{\Phi\left(\mu_p + \sigma_p x_i\right)}{\Phi\left(\sigma_p x_i\right)};$$

the calculation underlying this result equally reveals that $\Pr[s_i = 1|x_i] = \Phi(\sigma_p x_i)$. Thus, indeed, $\mathcal{E}^*[y_i|x_i] \neq \mathcal{E}[y_i|x_i]$ and an exponential-regression estimator will be inconsistent unless $\mu_p = 0$, which holds only when either $\rho = 0$ or $\sigma_w = 0$. Below we consider inference on β_0 for different choices of the parameter values based on the vector of empirical moments

$$\frac{1}{\varsigma}\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{i<j}\begin{pmatrix} x_i - x_j \\ a_i - a_j \end{pmatrix}\left(\frac{y_i}{\exp(x_i\beta)} - \frac{y_j}{\exp(x_j\beta)}\right)\kappa\left(\frac{\widehat{p}_i - \widehat{p}_j}{\varsigma}\right)s_i s_j. \tag{2.9}$$

In this case β_0 is overidentified. We consider both one-step and two-step versions of our estimator; the one-step estimator minimizes the Euclidean norm of the empirical moments, the two-step estimator uses the one-step estimator to form a plug-in estimator of the optimal metric.

We use the maximum rank-correlation estimator of Han (1987) to construct estimates of the p_i. Although it is known not to reach the semiparametric efficiency bound for the binary-choice model, this estimator is a simple and robust choice that does not require the choice of tuning parameters. It is consistent and asymptotically-linear under weak conditions; see Sherman (1993). For the implementation of our weighting procedure we take κ to equal the fourth-order kernel given in (2.7) and set $\varsigma = h_n\, n^{-1/7}$ for some scalar constant h_n; we will consider several choices for h_n. In any case, it is good practice to standardize the argument of the kernel function by its empirical standard deviation of $\widehat{p}_i$, and this is done throughout. To get an idea of the severity of the sample selection in each design we also consider one-step and two-step versions of the naive GMM estimator that is based on an unweighted version of (2.9).

Simulation results for three designs are presented. The designs differ only in the severity of the sample-selection issue, by varying ρ. A full description of the designs is given below Tables 1 and 2. For each design we evaluate the performance of the estimators for $n = 250$ and $n = 500$. Note that $\Pr[s_i = 1] = \frac{1}{2}$, so the effective sample sizes are only 125 and

Table 1. One-step estimators

ρ	n	bias						standard deviation					
	h_n	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$
0	250	.017	.020	.019	.018	.018	.018	.188	.299	.288	.236	.214	.271
0	500	.009	.016	.016	.015	.015	.004	.127	.205	.201	.175	.159	.173
-.5	250	-.202	-.010	-.016	-.057	-.084	.000	.167	.274	.264	.217	.195	.255
-.5	500	-.199	.007	.003	-.033	-.060	.002	.117	.193	.189	.163	.147	.169
.5	250	.281	.038	.047	.105	.139	.030	.205	.307	.296	.247	.230	.232
.5	500	.262	.036	.040	.080	.111	.010	.141	.212	.208	.179	.164	.130
ρ	n	standard error to standard deviation						rejection frequency					
	h_n	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$
0	250	.908	1.072	1.092	1.056	1.013	1.211	.067	.033	.020	.033	.052	.033
0	500	.948	1.071	1.065	1.065	1.036	1.260	.065	.030	.022	.019	.031	.038
-.5	250	.911	1.122	1.157	1.132	1.046	1.217	.320	.037	.028	.060	.085	.031
-.5	500	.931	1.079	1.079	1.121	1.055	1.228	.453	.044	.036	.046	.067	.034
.5	250	.936	1.069	1.132	1.061	.993	1.428	.295	.032	.019	.046	.075	.021
.5	500	.961	1.042	1.061	1.102	1.036	1.690	.486	.038	.036	.053	.081	.032

Parameters: $\varrho = -.5$, $\sigma_x = \sqrt{.5}$, $\sigma_a = \sqrt{.5}$, $\sigma_w = \sqrt{.5}$; $\beta_0 = 1$; $\gamma_0 = -1$.

250, on average, which is fairly small for semiparametric techniques. We computed β_n using four different but fixed values for h_n to evaluate the sensitivity of the results to the bandwidth choice, and also using one automated choice. The latter bandwidth is obtained by minimizing

$$\widehat{q}_n(\beta, h_n)' V_n \, \widehat{q}_n(\beta, h_n)$$

jointly with respect to (β, h_n), using obvious notation. That is, we treat h_n as a parameter to be estimated. This is similar to the proposal of Härdle, Hall, and Ichimura (1993) in the semiparametric least-squares context, where it is known to possess certain optimality properties. Although we make no claim that such optimality carries over to the current setting we will find that this approach works quite well in our simulations. Tables 1 and 2 contain the bias, the standard deviation, the ratio of the average estimated standard error to the standard deviation, and the empirical rejection rate of 95%-confidence intervals for β_0 for the one-step and the two-step GMM estimators, respectively, obtained over $1,000$ Monte Carlo replications. The columns with $h_n = \sim$ refer to the naive estimator. The columns with $h_n = \widehat{h}_n$ relate to the estimator constructed by estimating the bandwidth as just described.

Both tables show that the naive estimator does well when sample selection is exogenous but is heavily biased when it is not. Our estimator has much smaller bias in such cases for all designs considered. Like in standard nonparametric regression, the choice of h_n affects both the bias and the variance of the estimator. The larger h_n, the smaller the variance but the larger the bias, and vice versa. Nonetheless, the plug-in estimator of the asymptotic variance does quite well in capturing the variability of the estimator across the Monte Carlo replications for all choices of h_n. Similarly, the empirical rejection frequencies are fairly close to their nominal value of .05 while the bias in the naive estimator implies poor coverage of the confidence intervals constructed from it.

Table 2. Two-step estimators

ρ	n	bias						standard deviation					
	h_n	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$
0	250	.026	.021	.019	.027	.026	.026	.185	.301	.303	.212	.193	.307
0	500	.013	.016	.015	.016	.015	.006	.124	.203	.204	.152	.131	.207
-.5	250	-.163	-.009	-.004	-.109	-.125	.001	.166	.275	.284	.209	.176	.275
-.5	500	-.161	.006	.010	-.085	-.134	-.002	.115	.192	.194	.171	.122	.190
.5	250	.271	.039	.032	.196	.191	.044	.204	.312	.317	.233	.211	.315
.5	500	.245	.037	.032	.188	.209	.017	.140	.212	.214	.182	.145	.207
ρ	n	standard error to standard deviation						rejection frequency					
	h_n	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$
0	250	.889	.990	.940	1.014	.997	.997	.081	.050	.050	.050	.057	.056
0	500	.935	1.041	.993	1.029	1.056	1.013	.074	.036	.038	.028	.042	.045
-.5	250	.886	1.025	.966	.944	.996	1.026	.277	.055	.062	.124	.140	.046
-.5	500	.913	1.035	1.000	.872	1.011	1.040	.375	.049	.059	.152	.217	.043
.5	250	.905	.974	.924	1.012	1.012	.987	.297	.044	.058	.099	.116	.045
.5	500	.929	1.002	.970	.950	1.055	1.023	.457	.048	.055	.184	.230	.047

Parameters: $\varrho = -.5$, $\sigma_x = \sqrt{.5}$, $\sigma_a = \sqrt{.5}$, $\sigma_w = \sqrt{.5}$; $\beta_0 = 1$; $\gamma_0 = -1$.

3. A semiparametric approach for panel data

3.1. The model and moment conditions

Now consider a semiparametric model for group-level data with stratum-specific nuisance parameters. For independent groups $i = 1, 2 \ldots, n$, let $y_i \equiv (y_{i1}, y_{i2})$ be outcomes whose conditional mean given observables $x_i \equiv (x_{i1}, x_{i2})$, unobservables $u_i \equiv (u_{i1}, u_{i2})$, and a group-specific fixed effect η_i is given by

$$\mathcal{E}[y_{ij}|x_i, u_i, \eta_i] = \mu(x_{ij}; \alpha_0) + \eta_i \, \varphi(x_{ij}; \beta_0) \, u_{ij}. \tag{3.1}$$

The focus on two datapoints per group will simplify the subsequent exposition in terms of notational burden but is without loss of generality. A panel analog of the cross-sectional selection rule from above has

$$\Pr[s_{ij} = 1|p_i, \iota_i] = \mathcal{E}[1\{p_{ij} + \iota_i \geq v_{ij}\}|p_i, \iota_i], \tag{3.2}$$

where $p_i \equiv (p_{i1}, p_{i2})$ with $p_{ij} \equiv z_{ij}'\gamma_0$ for regressors $z_i \equiv (z_{i1}, z_{i2})'$, and ι_i a fixed effect. Interest again lies in consistently estimating the finite-dimensional parameter $\theta_0 = (\alpha_0', \beta_0')'$ under asymptotics where the number of groups, n, diverges.

Similar to before, let $\tau_{ij}(\theta) \equiv (y_{ij} - \mu(x_{ij}; \alpha))/\varphi(x_{ij}; \beta)$. Suppose that (u_i, v_i) are jointly independent of (x_i, z_i) conditional on (η_i, ι_i). Then, if the $\{(u_{i1}, v_{i1}), (u_{i2}, v_{i2})\}$ are exchangeable conditional on (η_i, ι_i),

$$\mathcal{E}^*[\tau_{i1}(\theta_0) - \tau_{i2}(\theta_0)|x_i, p_i, \eta_i, \iota_i] = \eta_i \, \lambda_i(p_{i1} + \iota_i, p_{i2} + \iota_i) - \eta_i \, \lambda_i(p_{i2} + \iota_i, p_{i1} + \iota_i), \tag{3.3}$$

where, now, the superscript on the expectations operator is a shorthand for the conditioning event $\{s_{i1} = 1, s_{i2} = 1\}$ holding, and

$$\lambda_i(p_{i1} + \iota_i, p_{i2} + \iota_i) \equiv \mathcal{E}^*[u_{i1}|v_{i1} \leq p_{i1} + \iota_i, v_{i2} \leq p_{i2} + \iota_i, \eta_i, \iota_i],$$
$$\lambda_i(p_{i2} + \iota_i, p_{i1} + \iota_i) \equiv \mathcal{E}^*[u_{i2}|v_{i1} \leq p_{i1} + \iota_i, v_{i2} \leq p_{i2} + \iota_i, \eta_i, \iota_i].$$

The i subscript on λ stresses that the function can be heterogenous because the distribution of (u_i, v_i) can vary with i beyond its dependence on (η_i, ι_i). For example, if (u_{ij}, v_{ij}) were independent of (η_i, ι_i) and i.i.d. within groups, λ_i would simplify to

$$\lambda_i(p_{ij} + \iota_i) = \frac{\int_{-\infty}^{p_{ij}+\iota_i} \int_{-\infty}^{+\infty} u \; f_i(u, v) \, \mathrm{d}u \, \mathrm{d}v}{F_i(p_{ij} + \iota_i)}$$

for f_i the joint distribution of (u_{ij}, v_{ij}) and F_i the marginal distribution of the v_{ij}. This function varies with i because ι_i and f_i do.

In (3.3), the differencing is done within groups, not across groups. Indeed, the additional heterogeneity across i that is allowed for here, as compared to the cross-sectional model, invalidates any approach based on the pairwise comparison of observations along the cross-sectional dimension of the panel. However, exchangeability implies that, for $\Delta_i \equiv p_{i1} - p_{i2}$,

$$|\lambda_i(p_{i1} + \iota_i, p_{i2} + \iota_i) - \lambda_i(p_{i2} + \iota_i, p_{i1} + \iota_i)| \xrightarrow{|\Delta_i| \downarrow 0} 0$$

provided λ_i is a smooth function in the same sense as discussed above. This motivates a strategy that aims to recover θ_0 from moment conditions of the form

$$\mathcal{E}^* \left[\frac{\omega(x_{i1}, x_{i2}) \left(\tau_{i1}(\theta_0) - \tau_{i2}(\theta_0) \right)}{\varsigma} \kappa\left(\frac{\Delta_i}{\varsigma} \right) \right] \xrightarrow{|\varsigma| \downarrow 0} 0, \tag{3.4}$$

where we recycle notation for the kernel function κ and bandwidth ς, and retain $\omega(x_{i1}, x_{i2})$ to indicate a vector of instrumental variables. We will provide distribution theory for estimators based on (3.4) in the next subsection.

Like in the cross-sectional case, these moment conditions can be linked to work that deals with sample selection in the linear model, namely Kyriazidou (1997, 2001); see also the work of Honoré and Kyriazidou (2000) for a related approach to a different problem. Indeed, Kyriazidou (1997, 2001) can be interpreted as a fixed-effect version of Powell (1987). It is, perhaps, useful to stress that the presence of fixed effects makes it difficult to extend the approaches of Blundell and Powell (2004) and Newey (2009) to models for panel data. Indeed, an operational version of (3.4) requires only a consistent estimator of γ_0, not of the ι_i or the λ_i, nor of functions thereof. Such estimators cannot be constructed under asymptotics where the number of observations per group is treated as fixed.

3.2. Estimation

Similar to before, (3.4) suggests estimating θ_0 by a GMM estimator of the form

$$\theta_n = \arg\min_{\theta \in \Theta} \widehat{q}_n(\theta)' V_n \widehat{q}_n(\theta),$$

where V_n is a weight matrix and, now,

$$\widehat{q}_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} \frac{\omega(x_{i1}, x_{i2}) \left(\tau_{i1}(\theta) - \tau_{i2}(\theta) \right)}{\varsigma} \kappa\left(\frac{\widehat{\Delta}_i}{\varsigma} \right) s_{i1} s_{i2}$$

for $\widehat{\Delta}_i \equiv \widehat{p}_{i1} - \widehat{p}_{i2}$, $\widehat{p}_{ij} \equiv z_{ij}'\gamma_n$, and γ_n a first-stage estimator of γ_0. Although this estimator looks similar to the one introduced for the cross-sectional model above, its asymptotic behavior is quite different. Most importantly, it will converge at the nonparametric rate $1/\sqrt{n\varsigma}$ rather than at the parametric rate $1/\sqrt{n}$. The reason for this is the need to resort to within-group differences rather than between-group differences. Indeed, here, $\widehat{q}_n(\theta)$ has the form of (the numerator of) a kernel-based nonparametric conditional-mean estimator, while it was a U-statistic of order two in the cross-sectional case. The smoothing that arises from the additional averaging in the U-statistic is what leads to $1/\sqrt{n}$ convergence in that case.

The conditions under which we will derive distribution theory for θ_n are provided next. The first assumption is again standard.

ASSUMPTION 3.1 (REGULARITY). *The space Θ is compact and θ_0 lies in its interior. Equation (3.4) identifies θ_0, and μ and φ are twice continuously differentiable in θ. The distribution of Δ_i given selection is absolutely continuous and the associated density, f^*, is strictly positive in a neighborhood of zero.*

Assumption 3.1 does not require stationarity of the data, but identification clearly requires that the support of p_{i1} and the support of p_{i2} overlap to some extent.

The next two assumptions deal with the first-stage estimator, and with the kernel and bandwidth, respectively.

ASSUMPTION 3.2 (FIRST STEP). $\|\gamma_n - \gamma_0\| = O_p(n^{-s})$ *for* $s \in (2/5, 1/2]$.

ASSUMPTION 3.3 (KERNEL). *The kernel function κ is bounded and twice continuously differentiable with bounded derivatives κ' and κ'', symmetric, and integrates to one. For some integer k, $\int_{-\infty}^{+\infty} \varepsilon^h \kappa(\varepsilon)\,\mathrm{d}\varepsilon = 0$ for all $0 < h < k$, and $\int_{-\infty}^{+\infty} |\kappa(\varepsilon)|\,\mathrm{d}\varepsilon < +\infty$ and $\int_{-\infty}^{+\infty} |\varepsilon^k|\,|\kappa(\varepsilon)|\,\mathrm{d}\varepsilon < +\infty$. Also, $\int_{-\infty}^{+\infty} |\kappa(\varepsilon)|^2\,\mathrm{d}\varepsilon < +\infty$. The bandwidth satisfies $\varsigma \propto n^{-r}$ for $\max\{\frac{1}{1+2k}, 1-2s\} < r < \frac{s}{2}$.*

Assumptions 3.2 and 3.3 are different from Assumptions 2.2 and 2.3 in some important respects. First, fixed-effect estimation of binary-choice models is known to be difficult, and possible at the parametric rate only in a logit specification (Chamberlain 2010). Thus, we need to allow for estimators that converge at a slower rate. Second, the permissible convergence rates of the bandwidth depend on the first-stage estimator. Under Assumptions 3.2 and 3.3, $\sqrt{n\varsigma}\,\|\gamma_n - \gamma_0\| = o_P(1)$, so θ_n will converge slower than γ_n, and the asymptotic variance of the former will not depend on the estimation noise in the latter. The rates in Assumption 3.2 allow for estimation by smoothed maximum score (Horowitz 1992; Charlier, Melenberg, and van Soest 1995) but rule out the original maximum-score estimator (Manski 1975, 1985, 1987), which converges at the rate $n^{-1/3}$ (Kim and Pollard 1990).

To move on to the formulation of moment requirements and smoothness conditions, let

$$\zeta(\Delta_i;\theta) \equiv \mathcal{E}^*\left[\{A_{i1}(\theta) - A_{i2}(\theta)\} + \eta_i \left\{\lambda_i(p_{i1}+\iota_i)B_{i1}(\theta) - \lambda_i(p_{i2}+\iota_i)B_{i2}(\theta)\right\} |\Delta_i\right] d(\Delta_i),$$

where, recalling that f^* is the density of Δ_i given selection, $d(\Delta_i) \equiv f^*(\Delta_i)\Pr[s_{i1}s_{i2}=1]$ and

$$A_{ij}(\theta) \equiv \omega(x_{i1},x_{i2})\,\frac{\mu(x_{ij};\alpha_0)-\mu(x_{ij};\alpha)}{\varphi(x_{ij};\beta)}, \qquad B_{ij}(\theta) \equiv \omega(x_{i1},x_{i2})\,\frac{\varphi(x_{ij};\beta_0)}{\varphi(x_{ij};\beta)}.$$

Consistency will follow from the following restrictions. The motivation for these conditions is as before.

Assumption 3.4 (finite moments). *For each $\theta\in\Theta$, both $\mathcal{E}^*[|\tau_{i1}(\theta)-\tau_{i2}(\theta)|^6]$ and $\mathcal{E}^*[\|\tau'_{i1}(\theta)-\tau'_{i2}(\theta)\|^4]$ are finite. Both $\mathcal{E}[\|\omega(x_{i1},x_{i2})\|^6]$ and $\mathcal{E}[\|z_{i1}-z_{i2}\|^4]$ are finite. The function $\zeta(\Delta;\theta)$ is continuous in Δ in a neighborhood of zero and $\sup_\Delta\|\zeta(\Delta;\theta)\|$ is finite for each $\theta\in\Theta$.*

The moment conditions validate the use of laws of large numbers. The requirements on $\zeta(\Delta_i;\theta)$ allow the use of a bounded-convergence argument to establish uniform convergence of the empirical moment. It can again be seen as a tail condition on the various conditional expectations involving $A_{ij}(\theta)$, $B_{ij}(\theta)$, η_i, and λ_i that appear in it.

The next assumption is used to obtain asymptotic normality and zero asymptotic bias. To state it, again let

$$\zeta'(\Delta;\theta) \equiv \frac{\partial\zeta(\Delta;\theta)}{\partial\theta'}, \qquad \nabla_h\zeta(\Delta;\theta) \equiv \frac{\partial^h\zeta(\Delta;\theta)}{\partial\Delta^h},$$

$Q_0 \equiv \zeta'(0;\theta_0)$, and τ'' for the second-derivative matrix of τ. The proof to asymptotic normality shows that, under Assumptions 3.1–3.5,

$$\sqrt{n\varsigma}\,\widehat{q}_n(\theta_0) \xrightarrow{L} \mathcal{N}(0,\Sigma),$$

where $\Sigma \equiv \Sigma(0;\theta_0)\int_{-\infty}^{+\infty}\kappa(\varepsilon)^2\,\mathrm{d}\varepsilon$ for

$$\Sigma(\Delta_i;\theta_0) \equiv \mathcal{E}^*[\omega(x_{i1},x_{i2})\omega(x_{i1},x_{i2})'(\tau_{i1}(\theta_0)-\tau_{i2}(\theta_0))^2|\Delta_i]\,d(\Delta_i).$$

The expression for the asymptotic variance cannot be simplified further given that our model does not restrict the conditional variance of the y_{ij} and does not rule out serial correlation in the (u_{ij},v_{ij}). Observe that, indeed, Σ does not depend on the asymptotic variance of the first-stage estimator γ_n. We will need a technical restriction on the matrix

$$H(\Delta_i) \equiv \mathcal{E}^*[\omega(x_{i1},x_{i2})\,(z_{i1}-z_{i2})'\,\eta_i\,\{\lambda_i(p_{i1}+\iota_i)-\lambda_i(p_{i2}+\iota_i)\}|\Delta_i]\,d(\Delta_i)$$

to justify this formally. This matrix arises from an expansion of the empirical moment around γ_0.

Assumption 3.5 (smoothness). *For each $\theta\in\Theta$, $\mathcal{E}^*[\|\tau''_{i1}(\theta)-\tau''_{i2}(\theta)\|^4]$ is finite and $\zeta'(\Delta;\theta)$ is continuous in Δ, and $\sup_\Delta\|\zeta'(\Delta;\theta)\|$ is finite. $\Sigma(\Delta;\theta_0)$ is continuous in Δ in a neighborhood of zero, $\sup_\Delta\|\Sigma(\Delta;\theta_0)\|$ is finite, and $H(\Delta)$ is continuously-differentiable in Δ and the derivative is bounded. The functions $\nabla_h\zeta(\Delta;\theta_0)$ exist and $\sup_\Delta\|\nabla_h\zeta(\Delta;\theta_0)\|$ is finite for all integers $h\leq k$.*

Note that

$$\zeta(\Delta_i;\theta_0) = \mathcal{E}^*\left[\eta_i\,\omega(x_{i1},x_{i2})\left\{\lambda_i(p_{i1}+\iota_i)-\lambda_i(p_{i2}+\iota_i)\right\}|\Delta_i\right]\,d(\Delta_i)$$

and so, indeed, $\zeta(0;\theta_0)=0$.

The delta method then yields the asymptotic distribution of θ_n.

THEOREM 3.1 (ASYMPTOTIC DISTRIBUTION). *Let Assumptions 3.1–3.5 hold. Suppose that Σ is positive definite, that Q_0 has full column rank, and that $V_n \xrightarrow{P} V_0$ for V_0 positive definite. Then $\sqrt{n\varsigma}\|\theta_n-\theta_0\| = O_P(1)$ and*

$$\sqrt{n\varsigma}(\theta_n-\theta_0) \xrightarrow{L} \mathcal{N}(0,\Upsilon), \quad \Upsilon \equiv (Q_0'V_0Q_0)^{-1}\,(Q_0'V_0\Sigma V_0Q_0)\,(Q_0'V_0Q_0)^{-1}.$$

In particular, if $V_n \xrightarrow{P} \Sigma^{-1}$, then $\Upsilon = (Q_0'\Sigma^{-1}Q_0)^{-1}$.

The asymptotic variance can be estimated using the plug-in estimators

$$Q_n \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{\omega(x_{i1},x_{i2})(\tau_{i1}'(\theta_n)-\tau_{i2}'(\theta_n))'}{\varsigma}\kappa\left(\frac{\widehat{\Delta}_i}{\varsigma}\right)\,s_{i1}s_{i2},$$

$$\Sigma_n \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{\omega(x_{i1},x_{i2})\omega(x_{i1},x_{i2})'\,(\tau_{i1}(\theta_n)-\tau_{i2}(\theta_n))^2}{\varsigma}\kappa\left(\frac{\widehat{\Delta}_i}{\varsigma}\right)^2\,s_{i1}s_{i2},$$

for Q_0 and Σ, respectively.

THEOREM 3.2 (INFERENCE). *Let Assumptions 3.1–3.5 hold. Suppose that Σ is positive definite, that Q_0 has full column rank, and that $V_n \xrightarrow{P} V_0$ for V_n and V_0 positive definite. Then*

$$\Upsilon_n \xrightarrow{P} \Upsilon, \qquad \Upsilon_n \equiv (Q_n'V_nQ_n)^{-1}\,(Q_n'V_n\Sigma_nV_nQ_n)\,(Q_n'V_nQ_n)^{-1}.$$

Theorem 3.2 allows to conduct hypothesis tests on θ_0 or transformations thereof, and also again implies validity of overidentification tests based on the optimally-weighted GMM criterion.

3.3. Simulations

As a numerical illustration we apply the fixed-effect estimator to a panel-data version of the Poisson model we used in the cross-sectional simulation exercise. To maximize the comparison between both sampling situations the data was generated in exactly the same way as before. Thus, for each of n groups, two observations were generated from a Poisson distribution with conditional mean

$$\mathcal{E}[y_{ij}|x_i,u_i] = \exp(c+x_{ij}\beta_0)\,u_i,$$

and all conditioning variables were drawn as before. The sample-selection process, too, was designed in the same way. The only design change we consider, relative to the cross-sectional

Table 3. One-step estimators

ρ	n	bias						standard deviation					
	h_n	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$
0	500	.008	.071	.035	.020	.010	.051	.176	.744	.473	.336	.210	.383
0	1000	.006	.038	.012	.004	.001	.030	.122	.474	.315	.241	.156	.253
-0.5	500	-.184	.075	.046	.001	-.090	.049	.155	.679	.432	.306	.198	.446
-0.5	1000	-.198	.027	.008	-.013	-.093	.025	.112	.450	.301	.250	.154	.326
0.5	500	.257	.073	.051	.078	.168	.037	.195	.760	.475	.323	.225	.330
0.5	1000	.254	.058	.047	.064	.146	.025	.136	.522	.337	.260	.168	.244
ρ	n	standard error to standard deviation						rejection frequency					
	h_n	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$
0	500	.956	1.226	.875	.921	.971	1.115	.059	.076	.062	.068	.050	.049
0	1000	.979	.943	.954	.967	.987	1.149	.046	.057	.060	.056	.053	.034
-0.5	500	.971	.934	.934	.952	.956	1.103	.263	.064	.061	.061	.102	.059
-0.5	1000	.944	.937	.941	.883	.933	.929	.488	.070	.064	.059	.130	.042
0.5	500	.957	1.064	.875	.957	.952	1.260	.255	.075	.069	.061	.108	.043
0.5	1000	.969	1.262	.924	.927	.965	1.326	.477	.064	.056	.055	.129	.038

Parameters: $\varrho = -.5$, $\sigma_x = \sqrt{.5}$, $\sigma_a = \sqrt{.5}$, $\sigma_w = \sqrt{.5}$; $\beta_0 = 1$; $\gamma_0 = -1$.

model, is the sample size. Indeed, because $\Pr[s_{i1}s_{i2} = 1] = \Pr[s_{i1} = 1]\Pr[s_{i2} = 1] = \frac{1}{4}$, we double n so as to keep the effective sample size as before, on average. Smoothed maximum score was used to estimate the parameters of the selection equation. The maximum-score estimator was implemented using a standard-normal distribution function to smooth the indicator function and a bandwidth parameter set to $1.06\, n^{-1/4}$. Experimentation with different bandwidths yielded very similar results. We again compare our estimator to the naive unweighted estimator based on the same set of moment conditions. Note that this estimator is $\sqrt{n}$-consistent when sample selection is exogenous, as opposed to the two-stage approach introduced in this paper.

Tables 3 and 4 have the same layout as before. The performance of the estimator is broadly in line with the cross-sectional case and so we can be brief in our description of the tables. The naive estimator is again heavily biased when selection into the sample is endogenous. Our approach tends to deliver estimators with relatively small bias. Although the effective sample size is quite small, the ratio of the estimated standard error to the standard deviation is reasonably close to unity across the designs, so that inference based on the asymptotic distribution theory derived above may be reliable. A look at the empirical rejection frequencies confirms this. These findings are reassuring, as they suggest that the nonparametric convergence rate need not pose a serious problem in applications with a reasonable cross-sectional sample size.

Acknowledgments

A previous version of this paper circulated under the title 'Simple estimators for count data models with sample selection'. I am grateful to Han Hong, an associate editor, and two referees for very constructive comments. Financial support from Sciences Po's SAB is gratefully acknowledged.

Table 4. Two-step estimators

ρ	n	bias						standard deviation					
	h_n	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$
0	500	.013	.162	.064	.029	.012	.123	.171	.769	.464	.317	.206	.638
0	1000	.006	.089	.036	.017	.005	.061	.117	.481	.311	.223	.150	.462
-0.5	500	-.148	.145	.059	-.033	-.115	.101	.151	.698	.435	.285	.192	.636
-0.5	1000	-.162	.059	.011	-.050	-.129	.036	.111	.444	.304	.237	.146	.451
0.5	500	.235	.173	.122	.155	.208	.123	.188	.733	.478	.327	.225	.585
0.5	1000	.230	.118	.099	.137	.199	.069	.132	.510	.334	.253	.168	.431
ρ	n	standard error to standard deviation						rejection frequency					
	h_n	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$	$\sim$	1	1.5	2.5	3	$\widehat{h}_n$
0	500	.940	1.100	.837	.891	.949	.781	.070	.086	.071	.069	.046	.073
0	1000	.982	.877	.927	.959	.981	.748	.044	.066	.063	.056	.049	.066
-0.5	500	.962	.817	.859	.916	.935	.695	.211	.080	.075	.073	.131	.113
-0.5	1000	.930	.908	.894	.849	.921	.743	.396	.071	.067	.080	.203	.097
0.5	500	.938	1.051	.843	.898	.936	.857	.254	.082	.082	.080	.143	.097
0.5	1000	.956	1.733	.905	.902	.943	.810	.436	.067	.065	.080	.232	.088

Parameters: $\varrho = -.5$, $\sigma_x = \sqrt{.5}$, $\sigma_a = \sqrt{.5}$, $\sigma_w = \sqrt{.5}$; $\beta_0 = 1$; $\gamma_0 = -1$.

Appendix A. Proof of Theorems 2.1 and 2.2

Notation. The following notation will be used. Let $\xi_{ij}(\theta) \equiv \omega(x_i, x_j)\,(\tau_i(\theta) - \tau_j(\theta))$. Then

$$q_n(\theta) \equiv \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{i<j} \frac{\xi_{ij}(\theta)}{\varsigma} \kappa\left(\frac{p_i - p_j}{\varsigma}\right) s_i s_j.$$

Also, $q_0(\theta) \equiv \mathcal{E}[\zeta_i(p_i;\theta)]$. Note that $q_n(\theta)$ is an infeasible version of $\widehat{q}_n(\theta)$ where the kernel weight depends on the true p_i and that, under our assumptions, $q_0(\theta)$ is the limit of this empirical moment condition, i.e, $q_0(\theta) = \lim_{n\uparrow+\infty} \mathcal{E}[q_n(\theta)]$, as will be verified below. We let $\xi'_{ij}(\theta) \equiv \omega(x_i, x_j)\,(\tau'_i(\theta) - \tau'_j(\theta))'$ and write $\widehat{Q}_n(\theta)$, $Q_n(\theta)$, and $Q_0(\theta)$ for the Jacobian matrices associated with $\widehat{q}_n(\theta)$, $q_n(\theta)$, and $q_0(\theta)$, respectively.

Consistency. Given the regularity conditions in Assumption 2.1 and the fact that $V_n \xrightarrow{P} V_0$ by construction, it suffices to show that

$$\sup_{\theta\in\Theta} \|\widehat{q}_n(\theta) - q_0(\theta)\| = o_P(1).$$

Consistency will then follow from Theorem 2.1 of Newey and McFadden (1994). Note that the moment conditions in Assumption 2.4 imply that there exist $a > 0$ and $C_n = O_P(1)$ such that, for all pairs $\theta_1, \theta_2 \in \Theta$, $s_i s_j \|\xi_{ij}(\theta_1) - \xi_{ij}(\theta_2)\| \leq C_n \|\theta_1 - \theta_2\|^a$. Together with the boundedness of the kernel function, this implies that $\|\widehat{q}_n(\theta_1) - \widehat{q}_n(\theta_2)\| \leq O_P(1) \|\theta_1 - \theta_2\|^a$. Therefore, the uniform convergence result will follow if we can show that $\widehat{q}_n(\theta) \xrightarrow{P} q_0(\theta)$ for all $\theta \in \Theta$; see Lemma 2.9 in Newey and McFadden (1994). To do so, we use the triangle inequality to get the bound

$$\|\widehat{q}_n(\theta) - q_0(\theta)\| \leq \|\widehat{q}_n(\theta) - q_n(\theta)\| + \|q_n(\theta) - q_0(\theta)\| \tag{A.1}$$

for each $\theta \in \Theta$, and establish the pointwise convergence in probability of each of the terms on the right-hand side.

To tackle the estimation noise in the $\widehat{p}_i$, use the continuity of the kernel function and boundedness of its derivative stated in Assumption 2.3 to validate a mean-value expansion around γ_0. This shows that $\|\widehat{q}_n(\theta) - q_n(\theta)\|$ is bounded by

$$\frac{\sup_\varepsilon |\kappa'(\varepsilon)| \binom{n}{2}^{-1} \sum_{i=1}^n \sum_{j\neq i} s_i s_j \, \|\xi_{ij}(\theta)\| \, \|z_i\|}{\varsigma^2} \|\gamma_n - \gamma_0\| = O_P\left(\frac{1}{\varsigma^2\sqrt{n}}\right). \tag{A.2}$$

The first transition follows from the symmetry of the functions $\xi_{ij}(\theta)$ and κ, and the second transition follows from the finite-moment requirements in Assumption 2.4 and by the $\sqrt{n}$-consistency of γ_n stated in Assumption 2.2. Assumption 2.3 then ensures the bandwidth sequence to be so that this term is $o_P(1)$.

To see that $\|q_n(\theta) - q_0(\theta)\| = o_P(1)$, first note that $\|q_n - \mathcal{E}[q_n(\theta)]\| = o_P(1)$. Indeed, by the Cauchy-Schwarz inequality, $\mathcal{E}^*[\|\xi_{ij}(\theta)\|^2] \leq \sqrt{2\,\mathcal{E}^*[\|\omega(x_i,x_j)\|^4]\;\mathcal{E}^*[|\tau_i(\theta)|^4]}$, which is $O_P(1)$ by Assumption 2.3. Furthermore, by the boundedness of the kernel function stated in Assumption 2.4, $\mathcal{E}[q_n(\theta)]$ exists and

$$\mathcal{E}^*\left[\left\|\frac{\xi_{ij}(\theta)}{\varsigma}\,\kappa\left(\frac{p_i - p_j}{\varsigma}\right)\right\|^2\right] \leq \frac{\mathcal{E}^*[\|\xi_{ij}(\theta)\|^2]\,\sup_\varepsilon|\kappa(\varepsilon)|}{\varsigma^2} = O_P\left(\frac{1}{\varsigma^2}\right) = o(n),$$

with the last transition following from the rate conditions on the bandwidth sequence. Lemma 3.1 of Powell, Stock, and Stoker (1989) then provides the appropriate law of large numbers. Next,

$$\mathcal{E}\left[\frac{\xi_{ij}(\theta)}{\varsigma}\,\kappa\left(\frac{p_i - p_j}{\varsigma}\right) s_i s_j\right] = \mathcal{E}\left[\int_{-\infty}^{+\infty} \frac{\zeta_i(p;\theta)}{\varsigma}\,\kappa\left(\frac{p_i - p}{\varsigma}\right) \mathrm{d}p\right] \to \mathcal{E}[\zeta_i(p_i;\theta)],$$

by a dominated-convergence argument validated through the dominance condition on $\zeta_i(p;\theta)$ in Assumption 2.4, and so we have that $\|\mathcal{E}[q_n(\theta)] - q_0(\theta)\| = o_(1)$. By the triangle inequality,

$$\|q_n(\theta) - q_0(\theta)\| \leq \|q_n - \mathcal{E}[q_n(\theta)]\| + \|\mathcal{E}[q_n(\theta)] - q_0(\theta)\|, \tag{A.3}$$

and the result follows.

Combining (A.1) through (A.3) then implies that $\|\widehat{q}_n(\theta) - q_0(\theta)\| = o_P(1)$ for each $\theta \in \Theta$, which is what we wanted to show. □

Projection of the empirical moment. The main step in deriving the limit distribution of θ_n is showing that

$$\widehat{q}_n(\theta_0) = \frac{2}{n}\sum_{i=1}^n \sigma_i + o_P(1/\sqrt{n}). \tag{A.4}$$

To do so, note that a second-order expansion of $\widehat{q}_n(\theta)$ around the first-stage estimator gives

$$\widehat{q}_n(\theta) - q_n(\theta) = \binom{n}{2}^{-1} \sum_{i=1}^n \sum_{j\neq i} \frac{\xi_{ij}(\theta)}{\varsigma}\,\kappa'\left(\frac{(z_i - z_j)'\gamma_0}{\varsigma}\right) \frac{z_i'(\gamma_n - \gamma_0)}{\varsigma}\, s_i s_j + R_n \tag{A.5}$$

for a remainder term R_n which will be shown to be asymptotically negligible. To show (A.4) we will proceed in two steps. The first is to work out the expression for $\widehat{q}_n(\theta_0) - q_n(\theta_0)$ above up to $o_P(1/\sqrt{n})$, the second is to approximate the U-statistic $q_n(\theta_0)$ by its projection and to show that the approximation error is $o_P(1/\sqrt{n})$.

To quantify the impact of first-stage estimation error, first note that R_n is bounded by

$$\frac{\frac{1}{2}\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j\neq i} s_i s_j \, \|\xi_{ij}(\theta)\| \, \|z_i\|^2 \, \sup_\varepsilon |\kappa''(\varepsilon)|}{\varsigma^3} \|\gamma_n - \gamma_0\|^2 = O_P\left(\frac{1}{\varsigma^3 n}\right)$$

by the conditions on the kernel in Assumption 2.3, the moment conditions in Assumption 2.4, and the $\sqrt{n}$-convergence rate of $\|\gamma_n - \gamma_0\|$ in Assumption 2.2. Furthermore, because $\varsigma \propto n^{-r}$ for some $r < \frac{1}{6}$, this term is $o_P(1/\sqrt{n})$ and, thus, is asymptotically negligible. Replacing $\gamma_n - \gamma_0$ by its influence-function expression in (B.2) gives

$$\widehat{q}_n(\theta) - q_n(\theta) = \frac{1}{3}\binom{n}{3}^{-1}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\neq i,j} \frac{\xi_{ij}(\theta)\, s_i s_j}{\varsigma}\, \kappa'\left(\frac{(z_i - z_j)'\gamma_0}{\varsigma}\right)\frac{z_i'\psi_k}{\varsigma} + o_P\left(\frac{1}{\sqrt{n}}\right),$$

where we ignore terms for which either $k = i$ or $k = j$, as they are asymptotically negligible, and rescale appropriately; the effect of this rescaling is asymptotically negligible. Indeed, the contribution of terms with $k = i$, for example, is bounded by

$$\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{n}\left\|\frac{\xi_{ij}(\theta)\, s_i s_j}{\varsigma}\, \kappa'\left(\frac{(z_i - z_j)'\gamma_0}{\varsigma}\right)\frac{z_i'\psi_i}{\varsigma}\right\| = O_P\left(\frac{1}{\varsigma^2 n}\right) = o_P\left(\frac{1}{\sqrt{n}}\right),$$

which follows from Assumptions 2.2, 2.3, and 2.4. A symmetrization argument then allows writing (B.2) as a third-order U-statistic whose second moment is $o(n)$. This can be verified using the same arguments as were used to show consistency. Therefore, by Lemma 3.1 of Powell, Stock, and Stoker (1989), $\widehat{q}_n(\theta) - q_n(\theta)$ differs from its projection by a term that is $o_P(1/\sqrt{n})$. The projection itself equals

$$\frac{2}{3}\,\mathcal{E}[h_i(\theta)] - \frac{2}{n}\sum_{i=1}^{n}\{h_i(\theta) - \mathcal{E}[h_i(\theta)]\}, \qquad h_i(\theta) \equiv \mathcal{E}\left[\frac{\xi_{jk}(\theta)\, z_j'}{\varsigma^2}\, \kappa'\left(\frac{p_j - p_k}{\varsigma}\right) s_j s_k\right]\psi_i.$$

Now, by a change-of-variable in the first step and integration by parts in the second step,

$$\begin{aligned} h_i(\theta) &= \mathcal{E}\left[\frac{\int_{-\infty}^{+\infty}\zeta_j(p_j - \varsigma\eta;\theta)\, z_j'\, \kappa'(\eta)\, \mathrm{d}\eta}{\varsigma}\right]\psi_i \\ &= \mathcal{E}\left[\frac{\zeta_j(p_j - \varsigma\eta;\theta)\, z_j'\, \kappa(\eta)|_{-\infty}^{+\infty}}{\varsigma}\right]\psi_i - \mathcal{E}\left[\int_{-\infty}^{+\infty}\nabla_1\zeta_j(p_j - \varsigma\eta;\theta)\kappa(\eta)\,\mathrm{d}\eta\, z_j'\right]\psi_i. \end{aligned}$$

On evaluating in θ_0 the first right-hand side term is zero, being bounded in magnitude by

$$\mathcal{E}\left[\left\|\frac{\zeta_j(p_j - \varsigma\eta;\theta_0)\, z_j'\, \kappa(\eta)|_{-\infty}^{+\infty}}{\varsigma}\right\|\right]\|\psi_i\| \leq \frac{\sqrt{\mathcal{E}[\|z_j\|^2]\;\mathcal{E}[\sup_p\|\zeta_j(p;\theta_0)\|^2]}\;\kappa(\eta)|_{-\infty}^{+\infty}}{\varsigma}\|\psi_i\| = 0,$$

because the relevant moments exist and $\kappa(\varepsilon)-\kappa(-\varepsilon)=0$ for any $\varepsilon>0$. Further, because κ is a kth-order kernel, a kth-order Taylor expansion of $\nabla_1\zeta_j(p_j-\varsigma\eta;\theta_0)$ around $\eta=0$ yields

$$\mathcal{E}\left[\int_{-\infty}^{+\infty}\nabla_1\zeta_j(p_j-\varsigma\eta;\theta)\kappa(\eta)\,\mathrm{d}\eta\,z_j'\right]=\mathcal{E}\left[\nabla_1\zeta_j(p_j;\theta_0)\,z_j'+\frac{\int_{-\infty}^{+\infty}\nabla_{k+1}\zeta_j(*;\theta_0)\,\eta^k\,\kappa(\eta)\,\mathrm{d}\eta\,z_j'}{\varsigma^{-k}\,k!}\right],$$

where $*$ lies between $p_j-\varsigma\eta$ and p_j. Invoking Assumptions 2.3 and 2.5 and applying a dominated-convergence argument to the remainder term gives

$$\left\|\mathcal{E}\left[\frac{\int_{-\infty}^{+\infty}\nabla_{k+1}\zeta_j(*;\theta_0)\,\eta^k\,\kappa(\eta)\,\mathrm{d}\eta\,z_j'}{k!}\right]\right\|\leq C\,\frac{\sqrt{\mathcal{E}[\|z_j\|^2]\;\mathcal{E}[\sup_p\|\nabla_{k+1}\zeta_j(p;\theta_0)\|^2]}}{k!}=O_P(1)$$

for $C\equiv\int_{-\infty}^{+\infty}|\varepsilon^k|\,|\kappa(\varepsilon)|\,\mathrm{d}\varepsilon$. Because Assumption 2.3 implies that $\varsigma^k=o_P(1/\sqrt{n})$, we obtain

$$h_i(\theta_0)=-\,\mathcal{E}\left[\nabla_1\zeta_j(p_j;\theta_0)z_j'\right]\,\psi_i+o_P(1/\sqrt{n}).$$

Moreover,

$$\widehat{q}_n(\theta_0)-q_n(\theta_0)=-\frac{2}{n}\sum_{i=1}^n\mathcal{E}\left[\nabla_1\zeta_j(p_j;\theta_0)z_j'\right]\psi_i+o_P\left(\frac{1}{\sqrt{n}}\right),\tag{A.6}$$

because $\mathcal{E}[\psi_i]=0$.

Now turn to the projection of $q_n(\theta_0)$. By the arguments in the proof of consistency, we may again invoke Lemma 3.1 of Powell, Stock, and Stoker (1989) to justify that $q_n(\theta_0)$ equals its projection

$$\mathcal{E}[g_i(\theta_0)]+\frac{2}{n}\sum_{i=1}^n\{g_i(\theta_0)-\mathcal{E}[g_i(\theta_0)]\},\qquad g_i(\theta)\equiv\mathcal{E}\left[\frac{\xi_{ij}(\theta)}{\varsigma}\,\kappa\left(\frac{p_i-p_j}{\varsigma}\right)s_is_j\,\middle|\,y_i,x_i,p_i,s_i\right],$$

up to a term that is $o_P(1/\sqrt{n})$. By using a kth-order Taylor expansion and proceeding as before, Assumption 2.5 yields $g_i(\theta_0)=\zeta_i(p_i;\theta_0)+o_P(1/\sqrt{n})$, and so

$$q_n(\theta_0)=\frac{2}{n}\sum_{i=1}^n\zeta_i(p_i;\theta_0)+o_P\left(\frac{1}{\sqrt{n}}\right)\tag{A.7}$$

because $\mathcal{E}[g_i(\theta_0)]=o_P(1/\sqrt{n})$.

Combine (A.6) and (A.7) to see that they imply (A.4). □

Convergence of the Jacobian matrix. The same steps as those used to prove consistency will yield pointwise convergence of the Jacobian matrix, i.e.,

$$\|\widehat{Q}_n(\theta)-Q_0(\theta)\|\leq\|\widehat{Q}_n(\theta)-Q_n(\theta)\|+\|Q_n(\theta)-Q_0(\theta)\|=o_P(1)\tag{A.8}$$

for al $\theta\in\Theta$. Also, $\widehat{Q}_n(\theta)$ is continuous in θ and $s_is_j\,\|\widehat{Q}_n(\theta_1)-\widehat{Q}_n(\theta_2)\|\leq O_P(1)\,\|\theta_1-\theta_2\|^a$ for some $a>0$, and so Lemma 2.9 in Newey and McFadden (1994) will again imply the

uniform-convergence result sought for. For the first right-hand side term in (A.8), observe that

$$\|\widehat{Q}_n(\theta) - Q_n(\theta)\| \leq \frac{1}{\varsigma^2}\binom{n}{2}^{-1} \sum_{i=1}^{n}\sum_{i<j} s_i s_j \, \|\xi'_{ij}(\theta)\| \, \|z_i\| \, \|\gamma_n - \gamma_0\| = O_P\left(\frac{1}{\varsigma^2\sqrt{n}}\right) = o_P(1).$$

For the second right-hand side term, use the triangle inequality to get

$$\|Q_n(\theta) - Q_0(\theta)\| \leq \|Q_n(\theta) - \mathcal{E}[Q_n(\theta)]\| + \|\mathcal{E}[Q_n(\theta)] - Q_0(\theta)\| = o_P(1).$$

Because $\mathcal{E}[Q_n(\theta)]$ exists and

$$\mathcal{E}^*\left[\left\|\frac{\xi'(\theta)}{\varsigma}\,\kappa\left(\frac{p_i - p_j}{\varsigma}\right)\right\|^2\right] = o(n),$$

Lemma 3.1 of Powell, Stock, and Stoker (1989) establishes that $\|Q_n(\theta) - \mathcal{E}[Q_n(\theta)]\| = o_P(1)$. Also

$$\mathcal{E}\left[\frac{\xi'(\theta)}{\varsigma}\,\kappa\left(\frac{p_i - p_j}{\varsigma}\right) s_i s_j\right] = \mathcal{E}\left[\int_{-\infty}^{+\infty} \frac{\zeta'_i(p;\theta)}{\varsigma}\,\kappa\left(\frac{p_i - p}{\varsigma}\right)\,\mathrm{d}p\right] \to \mathcal{E}[\zeta'_i(p_i;\theta)],$$

and so $\|\mathcal{E}[Q_n(\theta)] - Q_0(\theta)\| = o_P(1)$. Put together this verifies (A.8) and yields uniform convergence. □

Asymptotic distribution. By continuity of $\widehat{q}_n(\theta)$ in θ, in tandem with (A.4) and the uniform convergence of $\widehat{Q}_n(\theta)$ to $Q_0(\theta)$ on Θ, an expansion of the first-order conditions of the GMM minimization problem yields

$$\sqrt{n}(\theta_n - \theta_0) = -(Q_0' V_0 Q_0)^{-1} Q_0' V_0 \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \sigma_i + o_P(1);$$

where, recall $V_n \xrightarrow{P} V_0$. Now, Assumptions 2.2 and 2.5 imply that $\mathrm{var}[\sigma_i] = \Sigma < +\infty$ while, clearly, $\mathcal{E}[\sigma_i] = 0$. Hence, θ_n is asymptotically linear, and $\sqrt{n}(\theta_n - \theta_0) \xrightarrow{L} \mathcal{N}(0, \Upsilon)$. When V_n is so that $V_0 = \Sigma^{-1}$, a calculation verifies that $\Upsilon = 4(Q_0'\Sigma^{-1}Q_0)^{-1}$. □

Inference. Because $V_n \xrightarrow{P} V_0$ by assumption, it suffices to show that (i) $\|Q_n - Q_0\| = o_P(1)$ and that (ii) $\|\Sigma_n - \Sigma\| = o_P(1)$.

To see that (i) holds, note that we have $\|\widehat{Q}_n(\theta_n) - \widehat{Q}_n(\theta_0)\| = O_P(1)\,\|\theta_n - \theta_0\|^a$ for some $a > 0$ from the argument above, and observe that $Q_n = \widehat{Q}_n(\theta_n)$. Further, because we have shown that $\|\theta_n - \theta_0\| = O_P(1/\sqrt{n})$, $Q_n = \widehat{Q}_n(\theta_0) + o_P(1)$. The results then follows from the pointwise convergence result in (A.8).

To see that (ii) holds, first note that it is implied by $n^{-1}\sum_{i=1}^{n}\|\widehat{\sigma}_i - \sigma_i\|^2 = o_P(1)$. Now,

$$\frac{1}{n}\sum_{i=1}^{n}\|\widehat{\sigma}_i - \sigma_i\|^2 \leq \frac{1}{n}\sum_{i=1}^{n}\left\{\|\widehat{\zeta}_i - \zeta_i\|^2 + \|H_0\|^2\,\|\widehat{\psi}_i - \psi_i\|^2 + \|\psi_i\|^2\,\|H_n - H_0\|^2\right\} + R_n,$$

where R_n captures lower-order terms and $\zeta_i \equiv \zeta_i(p_i;\theta_0)$. All the dominant right-hand side contributions will be $o_P(1)$ provided we can show that $\|H_n - H_0\| = o_P(1)$ and $n^{-1}\sum_{i=1}^n \|\widehat{\zeta}_i - \zeta_i\|^2 = o_P(1)$, as we assume that $\widehat{\psi}_i$ is so that $n^{-1}\sum_{i=1}^n \|\widehat{\psi}_i - \psi_i\|^2 = o_P(1)$. Start with the convergence of H_n. By an analogous reasoning as for Q_n, and a further expansion around γ_0,

$$H_n = \binom{n}{2}^{-1} \sum_{i=1}^n \sum_{j\neq i} \frac{\xi_{ij}(\theta_0)\, s_i s_j}{\varsigma}\, \kappa'\left(\frac{p_i - p_j}{\varsigma}\right) \frac{z_i'}{\varsigma} + o_P(1).$$

The summand on the right-hand side is already known to satisfy the conditions for the law of large numbers to apply. Further,

$$\mathcal{E}\left[\frac{\xi_{ij}(\theta_0)\, s_i s_j}{\varsigma}\, \kappa'\left(\frac{p_i - p_j}{\varsigma}\right) \frac{z_i'}{\varsigma}\right] \to H_0,$$

as was shown in the derivation of (A.6), and so $\|H_n - H_0\| = o_P(1)$. Also, again using similar arguments,

$$\frac{1}{n}\sum_{i=1}^n \|\widehat{\zeta}_i - \widetilde{\zeta}_i\|^2 = o_P(1), \quad \widetilde{\zeta}_i \equiv \frac{1}{n-1}\sum_{j\neq i} \frac{\omega(x_i,x_j)\,(\tau_i(\theta_0) - \tau_j(\theta_0))}{\varsigma}\, \kappa\left(\frac{p_i - p_j}{\varsigma}\right) s_i s_j,$$

is established. Finally, from the proof of Theorem 3.4 in Powell, Stock, and Stoker (1989) we immediately have

$$\mathcal{E}[\|\widetilde{\zeta}_i - \overline{\zeta}_i\|^2] = O\left(\frac{1}{\varsigma^3 n}\right) = o(1), \quad \overline{\zeta}_i \equiv \int_{-\infty}^{+\infty} \frac{\zeta_i(p;\theta_0)}{\varsigma}\, \kappa\left(\frac{p_i - p}{\varsigma}\right) \mathrm{d}p,$$

while the smoothness and dominance conditions in Assumption 2.5 imply that $\mathcal{E}[\|\overline{\zeta}_i - \zeta_i\|^2] = o(1)$. Therefore, $n^{-1}\sum_{i=1}^n \|\widetilde{\zeta}_i - \overline{\zeta}_i\|^2 = o_P(1)$ and $n^{-1}\sum_{i=1}^n \|\overline{\zeta}_i - \zeta_i\|^2 = o_P(1)$ by the law of large numbers, and $n^{-1}\sum_{i=1}^n \|\widetilde{\zeta}_i - \zeta_i\|^2 = o_P(1)$ follows. Therefore, (ii) has been shown and the proof is complete. □

Appendix B. Proof of Theorems 3.1 and 3.2

Notation. The following notation will be used. Let $\xi_i(\theta) \equiv \omega(x_{i1},x_{i2})\,(\tau_{i1}(\theta) - \tau_{i2}(\theta))$ and $s_i \equiv s_{i1}s_{i2}$. Then

$$q_n(\theta) \equiv \frac{1}{n}\sum_{i=1}^n \frac{\xi_i(\theta)}{\varsigma}\kappa\left(\frac{\Delta_i}{\varsigma}\right) s_i$$

Also, $q_0(\theta) \equiv \zeta(0;\theta)$. Note that $q_n(\theta)$ is the empirical moment condition that takes γ_0 as known, and that $q_0(\theta)$ is the large-n limit of this function. Similarly, we again let $\xi_i'(\theta) \equiv \omega(x_{i1},x_{i2})(\tau_{i1}'(\theta) - \tau_{i2}'(\theta))'$ and write $\widehat{Q}_n(\theta)$, $Q_n(\theta)$, and $Q_0(\theta)$ for the Jacobian matrices associated with $\widehat{q}_n(\theta)$, $q_n(\theta)$, and $q_0(\theta)$, respectively.

Consistency. We follow the same steps as in Appendix A to establish consistency of θ_n in the panel-data case. For any fixed $\theta \in \Theta$, by a mean-value expansion around γ_0, we have that

$$\|\widehat{q}_n(\theta) - q_n(\theta)\| \leq \frac{\sup_\varepsilon |\kappa'(\varepsilon)| \, \frac{1}{n}\sum_{i=1}^n s_i \, \|\xi_i(\theta)\| \, \|z_{i1} - z_{i2}\|}{\varsigma^2} \, \|\gamma_n - \gamma_0\| = O_P\left(\frac{1}{n^{s-2r}}\right),$$

by the the moment conditions in Assumption 3.4 and the convergence rate of the first-stage estimator. Because we set $r < s/2$, this term converges to zero in probability. Also, by a standard law of large numbers, $\|q_n(\theta) - \mathcal{E}[q_n(\theta)]\|$ is

$$\left\| \frac{1}{n}\sum_{i=1}^n \frac{\xi_i(\theta)}{\varsigma} \kappa\left(\frac{p_{i1} - p_{i2}}{\varsigma}\right) s_i - \mathcal{E}\left[\frac{\xi_i(\theta)}{\varsigma} \kappa\left(\frac{p_{i1} - p_{i2}}{\varsigma}\right) s_i\right] \right\| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Further, $\|\mathcal{E}[q_n(\theta) - q_0(\theta)]\| = o(1)$ because

$$\mathcal{E}[q_n(\theta)] = \mathcal{E}\left[\frac{\xi_i(\theta)}{\varsigma} \kappa\left(\frac{p_{i1} - p_{i2}}{\varsigma}\right) s_i\right] = \int_{-\infty}^{+\infty} \frac{\zeta(\Delta;\theta)}{\varsigma} \, \kappa\left(\frac{\Delta}{\varsigma}\right) \mathrm{d}\Delta \to q_0(\theta)$$

by a standard bounded-convergence argument, validated by the conditions in Assumption 3.4. Therefore, $\|\widehat{q}_n(\theta) - q_0(\theta)\| = o_p(1)$. The regularity conditions in Assumption 3.1 then ensure the remaining conditions for Lemma 2.9 in Newey and McFadden (1994) to apply, and consistency follows. □

Expansion of the empirical moment. The limit distribution of θ_n is the same as the limit distribution of the infeasible estimator $\arg\max_{\theta\in\Theta} \, q_n(\theta)' V_n \, q_n(\theta)$. To show this it suffices to show that

$$\sqrt{n\varsigma} \, \{\widehat{q}_n(\theta_0) - q_n(\theta_0)\} = o_P(1), \tag{B.1}$$

that is, that the contribution of the estimation noise introduced by the first-stage estimator of γ_0 is asymptotically negligible. To do so, consider a second-order expansion around γ_0 to get

$$\begin{aligned} \widehat{q}_n(\theta_0) - q_n(\theta_0) = {} & \frac{1}{n}\sum_{i=1}^n s_i \, \frac{\xi_i(\theta_0)}{\varsigma} \, \kappa'\left(\frac{\Delta_i}{\varsigma}\right) \frac{(z_{i1} - z_{i2})'}{\varsigma} \, (\gamma_n - \gamma_0) \\ & + \frac{1}{n}\sum_{i=1}^n s_i \, \frac{\xi_i(\theta_0)}{\varsigma} \, \kappa''(*) \, \frac{(z_{i1} - z_{i2})'}{\varsigma} \, (\gamma_n - \gamma_0)(\gamma_n - \gamma_0)' \, \frac{(z_{i1} - z_{i2})}{\varsigma}, \end{aligned} \tag{B.2}$$

where $*$ is as before. As shown in the proof to consistency above, the expectation of the first term on the right-hand side in (B.2) exists. With $H(\Delta_i) = \mathcal{E}^*[\xi_i(\theta_0) \, (z_{i1} - z_{i2})' | \Delta_i] \, f^*(\Delta_i)$,

$$\mathcal{E}^*\left[\frac{\xi_i(\theta_0) \, (z_{i1} - z_{i2})'}{\varsigma^2} \, \kappa'\left(\frac{\Delta_i}{\varsigma}\right)\right] = \int_{-\infty}^{+\infty} \frac{H(\Delta)}{\varsigma^2} \, \kappa'\left(\frac{\Delta}{\varsigma}\right) \mathrm{d}\Delta$$

which, by a change-of-variable argument, can be shown to converge to $\partial H(\Delta)/\partial\Delta|_{\Delta=0}$, which is $O(1)$. Because $\sqrt{n\varsigma} \, (\gamma_n - \gamma_0) = o_P(1)$, this implies that the first term in (B.2) is

$o_P(1/\sqrt{n\varsigma})$. Similarly, the second term is bounded by

$$\frac{\sup_\varepsilon|\kappa''(\varepsilon)|\,\frac{1}{n}\sum_{i=1}^n s_i\,\|\xi_i(\theta_0)\|\,\|z_{i1}-z_{i2}\|^2}{\varsigma^3}\,\|\gamma_n-\gamma_0\|^2 = O_P\left(n^{3r}\right)\,O_P\left(n^{-2s}\right) = o_P\left(\frac{1}{\sqrt{n\varsigma}}\right),$$

and so, too, can be ignored asymptotically. This establishes (B.1). □

Asymptotic behavior of the infeasible moment. We next show that $\sqrt{n\varsigma}\,q_n(\theta_0) \xrightarrow{L} \mathcal{N}(0,\Sigma)$. Add and subtract $\mathcal{E}[q_n(\theta_0)]$ to write

$$\sqrt{n\varsigma}\,q_n(\theta_0) = \sqrt{n\varsigma}\,\mathcal{E}[q_n(\theta_0)] + \sqrt{n\varsigma}\,\{q_n(\theta_0)-\mathcal{E}[q_n(\theta_0)]\}. \tag{B.3}$$

The first term constitutes bias and is asymptotically negligible. Indeed,

$$\|\mathcal{E}[q_n(\theta_0)]\| = \left\|\int_{-\infty}^{+\infty}\frac{\zeta(\Delta;\theta_0)}{\varsigma}\,\kappa\left(\frac{\Delta_i}{\varsigma}\right)\,\mathrm{d}\Delta\right\| \le \varsigma^k\frac{\sup_\Delta\|\nabla_k\zeta(\Delta;\theta_0)\|\int_{-\infty}^{+\infty}|\eta^k|\,|\kappa(\eta)|\,\mathrm{d}\eta}{k!}.$$

Because the constant is bounded by Assumptions 3.3 and 3.5, $\sqrt{n\varsigma}\,\mathcal{E}[q_n(\theta_0)] = \sqrt{n\varsigma}O(\varsigma^k)$, which is $o(1)$ as we require that $r > \frac{1}{1+2k}$.

The second term in (B.3) is the dominant term. To deal with it we verify the conditions for Lyapunov's central limit theorem for triangular arrays. Write

$$\sqrt{n\varsigma}\,\{q_n(\theta_0)-\mathcal{E}[q_n(\theta_0)]\} = \frac{1}{\sqrt{n}}\sum_{i=1}^n\{\sigma_i-\mathcal{E}[\sigma_i]\},\ \sigma_i \equiv \frac{\xi_i(\theta_0)}{\sqrt{\varsigma}}\,\kappa\left(\frac{\Delta_i}{\varsigma}\right)s_i.$$

Then it suffices to show that (i) $\mathrm{var}[\sigma_i] < +\infty$ and $\lim_{n\uparrow+\infty}\mathrm{var}[\sigma_i] = \Sigma$, and that (ii) $\sum_{i=1}^n\mathcal{E}[\|\sigma_i/\sqrt{n}\|^3] = o(1)$. To show (i), recall that $\mathrm{var}[\sigma_i] = \mathcal{E}[\sigma_i\sigma_i'] - \mathcal{E}[\sigma_i]\,\mathcal{E}[\sigma_i']$, and that

$$\mathcal{E}[\sigma_i] = \sqrt{\varsigma}\int_{-\infty}^{+\infty}\frac{\zeta(\Delta;\theta_0)}{\varsigma}\,\kappa\left(\frac{\Delta}{\varsigma}\right)\,\mathrm{d}\Delta, \qquad \mathcal{E}[\sigma_i\sigma_i'] = \int_{-\infty}^{+\infty}\frac{\Sigma(\Delta;\theta_0)}{\varsigma}\,\kappa\left(\frac{\Delta}{\varsigma}\right)^2\,\mathrm{d}\Delta.$$

From above, $\mathcal{E}[\sigma_i]$ exists and is $o(1)$, while

$$\mathcal{E}[\sigma_i\sigma_i'] = \int_{-\infty}^{+\infty}\Sigma(\varsigma\eta;\theta_0)\,\kappa(\eta)^2\,\mathrm{d}\eta \to \Sigma(0;\theta_0)\int\kappa(\eta)^2\,\mathrm{d}\eta = \Sigma$$

by bounded convergence because $\sup_\Delta\|\Sigma(\Delta;\theta_0)\| < +\infty$ and $\Sigma(\Delta;\theta_0)$ is continuous in Δ in a neighborhood of zero. Thus, (i) is satisfied. To verify (ii), observe that

$$\sum_{i=1}^n\mathcal{E}\left[\left\|\frac{\sigma_i}{\sqrt{n}}\right\|^3\right] = \sum_{i=1}^n\mathcal{E}\left[\left\|\frac{\xi_i(\theta_0)}{\sqrt{n\varsigma}}\,\kappa\left(\frac{\Delta_i}{\varsigma}\right)\right\|^3\right] \le \frac{1}{\sqrt{n\varsigma}}\int_{-\infty}^{+\infty}\frac{g(\Delta)}{\varsigma}\left|\kappa\left(\frac{\Delta}{\varsigma}\right)\right|^3\,\mathrm{d}\Delta,$$

which is $O(1/\sqrt{n\varsigma})$ because $g(\Delta_i) \equiv \mathcal{E}^*[\|\xi_i(\theta_0)\|^3|\Delta_i]\,f^*(\Delta_i)\Pr[s_i=1]$ is bounded. Hence, (ii) holds and

$$\sqrt{n\varsigma}\,q_n(\theta_0) \xrightarrow{L} \mathcal{N}(0,\Sigma) \tag{B.4}$$

has been shown. □

Convergence of the Jacobian matrix. We use the same approach as used when establishing consistency. Fix $\theta \in \Theta$. An expansion gives

$$\|\widehat{Q}_n(\theta) - Q_n(\theta)\| \leq \frac{\sup_\varepsilon |\kappa'(\varepsilon)| \, \frac{1}{n} \sum_{i=1}^n s_i \, \|\xi_i'(\theta)\| \, \|z_{i1} - z_{i2}\|}{\varsigma^2} \, \|\gamma_n - \gamma_0\| = o_P(1).$$

Also, $\|Q_n(\theta) - \mathcal{E}[Q_n(\theta)]\| = o_P(1)$ by the law of large numbers, and

$$\mathcal{E}[Q_n(\theta)] = \int_{-\infty}^{+\infty} \frac{\zeta'(\Delta;\theta)}{\varsigma} \, \kappa\left(\frac{\Delta}{\varsigma}\right) \mathrm{d}\Delta \to Q_0(\theta)$$

by bounded convergence. Pointwise convergence of the Jacobian has been established. Uniform convergence, i.e.,

$$\sup_{\theta \in \Theta} \|\widehat{Q}_n(\theta) - Q_0(\theta)\| = o_P(1) \tag{B.5}$$

follows from Lemma 2.9 in Newey and McFadden (1994). □

Asymptotic distribution. Using the results obtained so far the asymptotic distribution of θ_n is readily established. An expansion of the first-order conditions to the GMM minimization problem gives

$$\sqrt{n\varsigma}(\theta_n - \theta_0) = -(Q_0' V_0 Q_0)^{-1} \, Q_0' V_0 \, \sqrt{n\varsigma} \, q_n(\theta_0) + o_P(1),$$

on invoking $\|\theta_n - \theta_0\| = o_P(1)$ and appealing to (B.1)–(B.5). The delta method then yields

$$\sqrt{n\varsigma}(\theta_n - \theta_0) \xrightarrow{L} \mathcal{N}(0, \Upsilon),$$

where $\Upsilon = (Q_0' V_0 Q_0)^{-1} (Q_0' V_0 \Sigma V_0 Q_0)(Q_0' V_0 Q_0)^{-1}$ for a generic positive definite V_0, and $\Upsilon = (Q_0' \Sigma^{-1} Q_0)^{-1}$ for the optimally-chosen V_0. □

Inference. Given that $V_n \xrightarrow{P} V_0$ by construction, it suffices to consider consistency of Q_n and Σ_n. $Q_n \xrightarrow{P} Q_0$ follows from $\|\theta_n - \theta_0\| = o_P(1)$ and $\sup_{\theta \in \Theta} \|\widehat{Q}_n(\theta) - Q_0(\theta)\| = o_P(1)$, because $Q_n = \widehat{Q}_n(\theta_n)$. The argument for $\Sigma_n \xrightarrow{P} \Sigma$ is similar and is omitted. □

References

Andrews, D. W. and M. M. A. Schafgans (1998). Semiparametric estimation of the intercept of a sample selection model. *Review of Economic Studies 65*, 497–517.

Arellano, M. and B. E. Honoré (2001). Panel data models: Some recent developments. In J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume V, Chapter 53, pp. 3229–3329. Elsevier.

Blundell, R. W. and J. L. Powell (2003). Endogeneity in nonparametric and semiparametric regression models. In M. Dewatripont, L. P. Hansen, and S. J. Turnovsky (Eds.), *Advances In Economics and Econometrics*, Volume II, Chapter 8, pp. 312–357. Cambridge University Press.

Blundell, R. W. and J. L. Powell (2004). Endogeneity in semiparametric binary response models. *Review of Economic Studies 71*, 655–679.

Cameron, A. C., P. K. Trivedi, F. Milne, and J. Piggott (1988). A microeconometric model of the demand for health care insurance in Australia. *Review of Economic Studies 55*, 85–106.

Chamberlain, G. (1992). Comment: Sequential moment restrictions in panel data. *Journal of Business and Economic Statistics 10*, 20–26.

Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica 78*, 159–168.

Charlier, E., B. Melenberg, and A. H. O. van Soest (1995). A smoothed maximum score estimator for the binary choice panel data model with an application to labour force participation. *Statistica Nederlandica 49*, 324–342.

Gronau, R. (1973). The intrafamily allocation of time: The value of housewives' time. *American Economic Review 63*, 634–651.

Hall, A. R. (2005). *Generalized Method of Moments*. Advanced Texts in Econometrics. Oxford University Press.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics 35*, 303–316.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory 24*, 726–748.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*, 1029–1054.

Hansen, L. P., J. Heaton, and A. Yaron (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics 14*, 262–280.

Härdle, W., P. Hall, and H. Ichimura (1993). Optimal smoothing in single-index models. *Annals of Statistics 21*, 157–178.

Hausman, J. A., 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. (1974). Shadow prices, market wages, and labor supply. *Econometrica 42*, 679–694.

Heckman, J. J. (1978). Dummy endogeneous variables in a simultaneous equation system. *Econometrica 46*, 931–959.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*, 153–161.

Heckman, J. J. and R. Robb (1985). Alternative methods for evaluating the impact of interventions. In J. J. Heckman and B. Singer (Eds.), *Longitudinal Analysis of Labor Market Data*, Chapter 4, pp. 156–245. Cambridge University Press.

Honoré, B. E. and E. Kyriazidou (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica 68*, 839–874.

Honoré, B. E. and J. L. Powell (2005). Pairwise difference estimators for nonlinear models. In D. W. Andrews and J. H. Stock (Eds.), *Identification and Inference for Econometric Models, Essays in Honor of Thomas Rothenberg*, pp. 520– 553. Cambridge University Press.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica 60*, 505–531.

Jochmans, K. (2013). Pairwise-comparison estimation with non-parametric controls. *Econometrics Journal 16*, 340–372.

Kim, J. and D. Pollard (1990). Cube root asymptotics. *Annals of Statistics 18*, 191–219.

Kleibergen, F. and R. Paap (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics 133*, 97–126.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica 65*, 1335–1364.

Kyriazidou, E. (2001). Estimation of dynamic panel data sample selection models. *Review of Economic Studies 68*, 543–572.

Lee, S. (2007). Endogeneity in quantile regression models: A control function approach. *Journal of Econometrics 141*, 1131–1158.

Li, Q. and J. S. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics 3*, 205–228.

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics 27*, 313–333.

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica 55*, 357–362.

Müller, H.-G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *Annals of Statistics 12*, 766–774.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *Econometrics Journal 12*, S217–S229.

Newey, W. K. and D. L. McFadden (1994). Large sample estimation and hypothesis testing. In R. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, Chapter 36, pp. 2111–2245. Elsevier.

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*, 1027–1057.

Powell, J. L. (1987). Semiparametric estimation of bivariate latent variable models. Working paper No. 8704, Social Systems Research Institute, University of Wisconsin-Madison.

Powell, J. L. (1994). Estimation of semiparametric models. In R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics*, Volume IV, Chapter 41, pp. 2443–2521. Elsevier.

Powell, J. L. (2001). Semiparametric estimation of censored selection models. In C. Hsiao, K. Morimune, and J. L. Powell (Eds.), *Nonlinear Statistical Modeling*. Cambridge University Press.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica 57*, 1403–1430.

Santos Silva, J. M. C. and S. Tenreyro (2006). The log of gravity. *Review of Economics and Statistics 88*, 641–658.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica 26*, 393–415.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica 61*, 123–137.

Terza, J. V. (1998). Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *Journal of Econometrics 84*, 129–154.

Vella, F. (1998). Estimating models with sample selection bias: A survey. *Journal of Human Resources 33*, 127–169.

Verbeek, M. and T. Nijman (1992). Testing for selectivity bias in panel data models. *International Economic Review 33*, 681–703.

Winkelmann, R. (1998). Count data models with selectivity. *Econometric Reviews 17*, 339–360.

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics 68*, 115–132.

Wooldridge, J. M. (1997). Multiplicative panel data models without the strict exogeneity assumption. *Econometric Theory 13*, 667–678.