Woutersen, Tiemen

Working Paper

Robustness against incidental parameters

Research Report, No. 2002-8

Provided in Cooperation with:
Department of Economics, University of Western Ontario

Suggested Citation: Woutersen, Tiemen (2002) : Robustness against incidental parameters, Research Report, No. 2002-8, The University of Western Ontario, Department of Economics, London (Ontario)

This Version is available at:
https://hdl.handle.net/10419/70388

ROBUSTNESS AGAINST INCIDENTAL PARAMETERS*

By Tiemen Woutersen†

October 2002

Abstract

Neyman and Scott (1948) define the incidental parameter problem. In panel data with T observations per individual and unobservable individual-specific effects, the maximum likelihood estimator of the common parameters is in general inconsistent. This paper develops the integrated moment estimator. It shows that the inconsistency of the integrated moment estimator is of a low order, $O(T^{-2})$, and thereby offers an approximate solution for the incidental parameter problem. The integrated moment estimator allows for exogenous regressors, time dummies and lagged dependent variables and is efficient for an asymptotics in which $T \propto N^{\alpha}$ where $\alpha > \frac{1}{3}$. We adjust the integrated likelihood estimator to allow for general predetermined regressors. The paper also shows that methods that rely on differencing away the individual-specific effects can be viewed as special cases of the integrated moment estimator.

Keywords: Incidental parameters, predetermined variables, panel data

JEL classification: C33, C31, C14

*An earlier version of this paper circulated under the title "Robustness againt Incidental Parameters and Mixing Distributions". I gratefully acknowlegde stimulating suggestions from Manuel Arellano, Richard Blundell, Bo Honoré, Tony Lancaster, Nancy Reid, Geert Ridder, Peter Schmidt, Jeffrey Wooldridge and seminar participants at the Departments of Economics at Princeton, Queen's, Stanford, Tinbergen Institute, University College London, University of Toronto, CEMFI, and University of Western Ontario. Comments by Joel Horowitz improved the presentation and focus of the paper. The 'CIBC Human Capital and Productivity Program' and the 'Social Science and Humanities Council' in Canada provided financial support. All errors are mine.

†Mailing Address: University of Western Ontario, Department of Economics, Social Science Center, London, Ontario, N6A 5C2, Canada. Email: twouters@uwo.ca.

1 Introduction

One way to control for the heterogeneity in panel data is to allow for time-invariant, individual specific parameters. This fixed effect approach introduces many parameters into the model which causes the 'incidental parameter problem' of Neyman and Scott (1948): the maximum likelihood estimator is in general inconsistent. In particular, Neyman and Scott (1948), Mundlak (1961), Nickell (1981), and Chamberlain (1984) give examples in which the inconsistency of the maximum likelihood estimator is $O(T^{-1})$ where T is the number of periods for which we observe an individual.

Cox and Reid (1987) are not concerned with panel data but they propose a general method to deal with nuisance parameters. In particular, Cox and Reid (1987) propose to reparameterize the log likelihood so that its cross derivative with respect to the nuisance parameter and parameter of interest is zero in expectation. After this reparametrization, Cox and Reid (1987) apply their conditional profile likelihood method. A limit of this approach is that the reparametrization requires that a particular differential equation has an explicit solution. It is well known and discussed by Critchley (1987) and Hills (1987), that differential equations rarely have explicit solutions. Despite this drawback, Lancaster (2000 and 2002) applies the idea of reparametrization to panel data where the fixed effect plays the role of nuisance parameter. Lancaster (2000 and 2002) then integrates out the fixed effects but does not present general results. This paper shows that the inconsistency of this integrated likelihood estimator is $O(T^{-2})$. Moreover, this paper removes the need for reparametrization by using a moment function, $g(\beta, f)$, that approximately separates the individual nuisance parameter f from the parameter of interest β. Specifically, the partial derivative of $g(\beta, f)$ with respect to f is zero at the true values of the parameters,

$$Eg_f(\beta_0, f_0) = 0.$$

We show how to construct $g(\beta, f)$ for any likelihood and some quasi-likelihood models, subject to regularity conditions. We thus avoid reparametrizations and differential equations and the restrictions those imply. Moreover, we allow for general predetermined regressors,

time dummies, and the individual parameter does not necessarily have to be an individual mean but can also be an individual regressor coefficient or an individual elasticity. We then integrate the moment function $g(\beta, f)$ with respect to the likelihood of f, using a Laplace approximation. We use this approximation as an estimating function and define the integrated moment estimator accordingly. We then show that the inconsistency of the integrated moment estimator is $O(T^{-2})$.

Alvarez and Arellano (1998) develop an alternative asymptotic where T increases at the same rate as the number of individuals, N. We show that the integrated likelihood estimator is asymptotically unbiased under this alternative asymptotic. Given that T is smaller than N in most panel data, we argue that it might be more interesting to let T increase at a slower rate than N; we also show that the integrated moment estimator is asymptotically unbiased as long as T grows faster than $N^{1/3}$. The Cramér-Rao bound is well-defined for likelihood models with lagged dependent variables and the integrated moment estimator reaches the Cramér-Rao bound in this asymptotic. Simulations show the relevance of the asymptotic approximation.

Mundlak (1961), Chamberlain (1985), and Arellano and Honoré (2001) discuss how optimizing behavior of individuals causes dependence between the regressors and the heterogeneity. This dependence, as well as the unobservability of the heterogeneity distribution, motivates the fixed effect approach in which the heterogeneity distribution is left unspecified. If an applied researcher wants to specify a mixing distribution then the same line of reasoning implies that some robustness against the wrong choice of mixing distribution is desirable. Suppose a mixing distribution is specified for the individual parameter f. Using $g(\beta, f)$ where $Eg_f(\beta_0, f_0) = 0$ yields an estimator that is consistent for β if the model is correctly specified and whose inconsistency is $O(T^{-2})$ if the mixing distribution is misspecified. We thus generalize Mundlak's (1978) linear random effects model to nonlinear models. The argument against fixed effect models is usually that the set of models that can be estimated is so small. The reason for aversion against random effects models is usually based on the sensitivity to the choice of mixing distribution, see Nerlove (2000) and Trognon (2000)

for a recent exposition of these arguments. This paper gives an approximate solution for the incidental parameter problem and some algebra can be used to increase the robustness against the wrong choice of mixing distributions. For the linear and Weibull model with exogenous regressors, one can difference the (transformed) data to remove the fixed effects, see the overviews of Chamberlain (1984, 1985) and Arellano and Honoré (2001). The integrated moment estimator reproduces the difference estimator for these simple models.

The work of Hahn and Newey (2002) is related to this paper. They use a jackknife technique to reduce the bias in panel likelihood models with strictly exogenous regressors and without time dummies. Hahn and Newey (2002) assume that T increases as fast as N and show that the jackknife technique reduces the bias to $o(T^{-1})$, which is larger than $O(T^{-2})$.

Arellano and Honoré (2001) note "almost nothing is known about nonlinear models with general predetermined variables". Recent progress has been made by Honoré and Lewbel (2002) who consider the binary choice model with fixed effects and general predetermined variables. Honoré and Lewbel (2002) linearize the binary choice model by dividing by the density of an regressor which is assumed to be absolutely continuously distributed and conditionally independent of the fixed effect and error terms. Honoré and Lewbel (2002) then use instrumental variables to estimate the model. The integrated moment estimator allows the regressors to have discrete support and to depend on the fixed effects but the estimator needs T to increase slowly and is parametric.

This paper is organized as follows. Section 2 provides an informal description of the estimator. Section 3 presents formal results. Section 4 discusses the panel probit model with general predetermined regressors and fixed effects as an example. Section 5 concludes and all the proofs are in the appendix.

2 Informal description of the estimator

Suppose we observe N individuals for T periods. Let the log likelihood contribution of the t^{th} period of individual i be denoted by $L^{it}(\beta, f_i)$. Summing over i and t yields the log likelihood,

$$L(\beta, f) = \sum_i \sum_t L^{it}(\beta, f_i),$$

where β is the common parameter, f_i is the incidental parameter and $f = \{f_1, ..., f_N\}$ and we condition on exogenous regressors. Suppose that the parameter β is of interest and that the incidental parameter f_i is a nuisance parameter that controls for heterogeneity. One could estimate β by maximum likelihood. However, Neyman and Scott (1948), Mundlak (1961), Nickell (1981), and Chamberlain (1984) give examples in which the inconsistency of the maximum likelihood estimator is $O(T^{-1})$. The intuition for these examples is that the score $L_\beta(\beta, f)$ depends on f and that replacing f by its maximum likelihood estimate yields an estimating function with nonzero expectation since $L_\beta(\beta, f)$ and the maximum likelihood estimate for f are dependent. Specifically, $\frac{1}{NT} E L_\beta(\beta_0, \hat{f})$ is $O(T^{-1})$ where β_0 denotes the true value and $\hat{f}$ is the maximum likelihood of f given β_0. This paper gives an approximate solution to this incidental parameter problem by reducing the dependence between the moment function and $\hat{f}$. In particular, consider the moment function

$$(1) \qquad g(\beta, f) = \frac{\sum_i}{N}\{g^i(\beta, f)\} = \frac{\sum_i}{N}\{L^i_\beta(\beta, f) - L^i_f(\beta, f)\frac{\text{`}E\text{'}L^i_{\beta f_i}(\beta, f)}{\text{`}E\text{'}L^i_{f_i f_i}(\beta, f)}\}$$

$$\text{where } \frac{\text{`}E\text{'}L^i_{\beta f}(\beta, f)}{\text{`}E\text{'}L^i_{ff}(\beta, f)} = \frac{\int L^i_{\beta f}(\beta, f) e^{L^i(\beta, f)} dt}{\int L^i_{ff}(\beta, f) e^{L^i(\beta, f)} dt}$$

and partial derivatives with respect to β and f are denoted by the subscripts β and f while the subscript i is suppressed. Thus, 'E' denotes the expectation that is yielded by treating $e^{L^i(\beta, f)}$ as a density so that $\frac{\text{`}E\text{'}L^i_{\beta f}(\beta, f)}{\text{`}E\text{'}L^i_{ff}(\beta, f)}$ depends on β and f but is nonstochastic. Differentiating with respect to f gives

$$g_f(\beta, f) = \frac{\sum_i}{N} g^i_f(\beta, f) = \frac{\sum_i}{N}[L^i_{\beta f}(\beta, f) + L^i_{ff}(\beta, f)\frac{\text{`}E\text{'}L^i_{\beta f}(\beta, f)}{\text{`}E\text{'}L^i_{ff}(\beta, f)} + L^i_f(\beta, f)\frac{\partial\{\frac{\text{`}E\text{'}L^i_{\beta f}(\beta, f)}{\text{`}E\text{'}L^i_{ff}(\beta, f)}\}}{\partial\beta}].$$

Taking expectations and assuming that $\frac{`E'L^i_{\beta f}(\beta_0,f_0)}{`E'L^i_{ff}(\beta_0,f_0)} = \frac{EL^i_{\beta f}(\beta_0,f_0)}{EL^i_{ff}(\beta_0,f_0)}$ gives

$$Eg_f(\beta_0, f_0) = EL^i_{\beta f} - EL^i_{ff}\frac{`E'L^i_{\beta f}}{`E'L^i_{ff}} = 0.$$

Thus, the parameter of interest β is approximately separated from the nuisance parameter f. For two simple models, the Poisson model with exogenous regressors and the linear model with exogenous regressors[1], the moment $g(\beta, f)$ does not depend on f. However, for other models, the unknown parameter f does appear in the moment function of equation (1) and the next step is to integrate[2] out f with respect to the likelihood for given β. Let

$$g^{i,I}(\beta) = \frac{\int g^i(f)e^{L^i}df}{\int e^{L^i}df}.$$

Kass et al. (1990, theorem 7) give a Laplace approximation for this ratios of integrals,

$$g^{i,I}(\beta) = \frac{\int g^i(f)e^{L^i}df}{\int e^{L^i}df} = g^i(\hat{f}) - \frac{1}{2}\frac{g^i_{ff}(\hat{f})}{L^i_{ff}(\hat{f})} + \frac{1}{2}\frac{L^i_{fff}(\hat{f})g^i_f(\hat{f})}{\{L^i_{ff}(\hat{f})\}^2} + O_p(T^{-1}) \tag{2}$$

where $\hat{f}$ denotes the maximum likelihood estimate of f for given β. The expansion is valid for well-behaved likelihoods. We use this approximation[3] to estimate β and thus define the *integrated moment function.*

$$g^*(\beta, \hat{f}) = \frac{1}{NT}\sum_i \{g^i(\beta, \hat{f}) - \frac{1}{2}\frac{g^i_{ff}(\beta, \hat{f})}{L^i_{ff}(\beta, \hat{f})} + \frac{1}{2}\frac{L^i_{fff}(\beta, \hat{f})g^i_f(\beta, \hat{f})}{\{L^i_{ff}(\beta, \hat{f})\}^2}\} \tag{3}$$

We define the *integrated moment estimator* as follows

$$\hat{\beta} = \arg\min_{\beta,f}\{g^*(\beta, \hat{f})'g^*(\beta, \hat{f})\}.$$

Under conditions that are slightly weaker than correct specification of the model[4], the expectation of $g^*(\beta, \hat{f})$ is of a low order at the true value of the parameter, that is

$$Eg^*(\beta_0, \hat{f}) \text{ is } O(T^{-2}).$$

There are two causes of bias in panel data models. First, bias can be caused by a nonzero expectation of the estimating function. We dealt with this cause by applying approximate parameter separation so that $Eg^*(\beta_0, \hat{f})$ is of a low order. Bias can also be caused by

correlation between the estimating function and its derivative, i.e. the correlation between $g^*(\beta_0, \hat{f})$ and its derivative, $g^*_\beta(\beta_0)$. The latter source of bias is of small order in panel data models, $O(\frac{1}{NT})$, if N or T is large, see Newey and Smith (2001) for a discussion of this type of bias in cross section models. We approximate the finite sample properties of $\hat{\beta}$ by using an asymptotics in which N or T increases. In particular,

$$\sqrt{NT}(\hat{\beta} - \beta_0) \rightarrow_d N(\sqrt{NT}B, \Psi),$$

where Ψ denotes the variance-covariance matrix and B denotes the bias. We show that B is of a low order, $O(T^{-2})$. This implies that $\sqrt{NT}B$ is $O(\sqrt{\frac{N}{T^3}})$, which goes to zero if T grows faster then $N^{1/3}$. Moreover, we show in the next section that the integrated moment estimator reaches the Cramér-Rao efficiency bound in this asymptotic. An intuition for this efficiency result is that $g(\beta, f_0)$ is the efficient score and that the difference between the integrated moment and this efficient score is small in terms of mean squared error at β_0. That is,

$$\sqrt{NT}\{g^*(\beta_0, \hat{f}) - g(\beta_0, f_0)\} \text{ is } o_{ms}(1).$$

An important topic in panel data econometrics is robustness against common shocks. For example, labor force participation or expenditure decisions depend, among other things, on the business cycle. If the common shocks are caused by observables, such as the unemployment rate or the growth rate of Gross Domestic Product, then one may be able to control for the common shocks by including these observables as regressors. In other applications the common shock may not be observable to the econometrician and one prefers to include a vector of time dummies. A time dummy has value one for a fixed set of periods and is zero for all other periods. We can, therefore, only estimate time dummies at rate $N^{-1/2}$. We again assume that T grows faster than $N^{1/3}$. Applying the estimating function of equation (3) yields an efficient estimator with a bias of order $O(T^{-2})$ that converges at rate $(NT)^{-1/2}$ for the common parameter and at rate $N^{-1/2}$ for the time dummies.

Optimizing behavior of individuals often implies that the dependent variable influences the regressors in subsequent periods, as discussed in Arellano and Honoré (2001). We show

how to adjust the estimating function of equation (3) to allow for general predetermined regressors.

3 Assumptions and theorems

This section presents conditions under which the estimator for the common parameter, $\hat{\beta}$, can be efficiently estimated. Let

$$(4) \qquad g(\beta,f)=\frac{\sum_i}{NT}\{g^i(\beta,f)\}=\frac{\sum_i}{NT}\{L^i_\beta(\beta,f)-L^i_f(\beta,f)\frac{\int L^i_{\beta f}(\beta,f)e^{L^i(\beta,f)}dt}{\int L^i_{ff}(\beta,f)e^{L^i(\beta,f)}dt}\},$$

where we suppress the subscript of f. Consider the following objective function

$$Q(\beta,f)=-\{g^*(\beta,f)\}'\{g^*(\beta,f)\}-\{h(\beta,f)\}'\{h(\beta,f)\},$$

where

$$\begin{aligned} g^*(\beta,f) &= \frac{\sum_i}{NT}\{g^i(\beta,f)-\frac{1}{2}\frac{g^i_{ff}(\beta,f)}{L^i_{ff}(\beta,f)}+\frac{1}{2}\frac{L^i_{fff}(\beta,f)g^i_f(\beta,f)}{\{L^i_{ff}(\beta,f)\}^2}\} \text{ and} \\ h(\beta,f) &= \frac{L_f(\beta,f)}{T}. \end{aligned}$$

We assume that the moments $g(\beta,f)$ and $h(\beta,f)$ identify the parameter of interest β and the nuisance parameter f. Let $z=\{x,y\}$.

Assumption 1 (Moments): *Let* $\{Eg(z,\beta,f)=0,\ Eh(z,\beta,f)=0\}$ *be uniquely solved for* $\{\beta,f\}=\{\beta_0,f_0\}$ *where* $\{\beta_0,f_0\}\in\Theta$ *and* Θ *is compact.*

The compactness assumption is standard in econometrics and can be replaced by the requirement that $g(\beta,f)'g(\beta,f)+h(\beta,f)'h(\beta,f)$ is concave in its parameters. The following assumptions ensures that $g(\beta,f)$, $g^*(\beta,f)$ and $h(\beta,f)$ converge uniformly to their expectations; let $||\cdot||$ denote the Euclidean norm.

Assumption 2 (Uniform Convergence): *Let* $g(z,\beta,f)$, $\frac{g^i_{ff}(z,\beta,f)}{L^i_{ff}(z,\beta,f)}$, $\frac{L^i_{fff}(z,\beta,f)g^i_f(z,\beta,f)}{\{L^i_{ff}(z,\beta,f)\}^2}$, *and* $h(z,\beta,f)$ *be continuous for all* $\{\beta,f\}\in\Theta$ *with probability one for all* i; *let* $||g(z,\beta,f)||<a(z)$, $||\frac{g^i_{ff}(z,\beta,f)}{L^i_{ff}(z,\beta,f)}||<b(z)$, $||\frac{L^i_{fff}(z,\beta,f)g^i_f(z,\beta,f)}{\{L^i_{ff}(z,\beta,f)\}^2}||<c(z)$, *and* $||h(z,\beta,f)||<d(z)$ *for all* i *and for all* $\{\beta,f\}\in\Theta$ *where* $E[a(z)]<\infty$. $E[b(z)]<\infty$, $E[c(z)]<\infty$ *and* $E[d(z)]<\infty$.

Assumption 3 (Stationary and Exogeneity): *Let* $z_i = \{x_i, y_i\}$ *be strictly stationary and ergodic for all* i; *let* x_i *be exogenous.*

Assumption 4 (Asymptotics): Let $T \to \infty$.

Proposition 1

Suppose $\{\hat{\beta}, \hat{f}\} = \arg\min_{\beta, f \in \Theta}\{Q(\beta, f\}$. *Let assumption 1-4 hold. Then* $\hat{\beta} \to_p \beta_0$ and $\hat{f} \to_p f_0$.

Proof: See appendix.

Examples by Neyman and Scott (1948), Nickell (1981) and Chamberlain (1984) show that the asymptotic bias of the maximum likelihood is $O(T^{-1})$. Neyman and Scott called this the incidental parameter problem. We now show that approximate separation reduces the asymptotic bias of the maximum likelihood estimator to $O(T^{-2})$ for large N.

Assumption 5 (Approximate separation):

Let (i) $Eg_f(\beta_0, f_0) = 0$ *and (ii)* $EL_f(\beta_0, f_0)g_f(\beta_0, f_0) + Eg_{ff}(\beta_0, f_0) = 0$.

Condition (i) is implied by 'E'$L_{ff} = EL_{ff}$ and 'E'$L_{\beta f} = EL_{\beta f}$; condition (ii) holds if, in addition, $E(L_f)^2 = EL_{ff}$, 'E'$L_{\beta ff} = EL_{\beta ff}$, '$E$'$L_{\beta f}L_f = EL_{\beta f}L_f$ and 'E'$L_{fff} = EL_{fff}$, where 'E' denotes the expectation that is yielded by treating $e^{L^i(\beta,f)}$ as a density. Moreover, condition 5 holds under correct specification of the model, as shown in the appendix. In an dynamic linear model[5] with an individual parameter, we can assume normality of the error terms in order to derive moment functions but it follows from Assumption 5 that approximate separation only depends on correct specification of the first two moments of the error term.

We need T to increase with N and assume the following.

Assumption 6 (Asymptotics):

Let $T \propto N^{\alpha}$ *where* $\alpha > \frac{1}{3}$.

We also impose an assumption about the conditionally independent across individuals. For a misspecified model, assuming that the data generating process is conditional independent across individuals is not sufficient. The reason for this is that a common regressor, in

combination with some misspecification, could induce dependence of the contribution to the moment functions of different individuals, i.e. that $g^i(\beta_0, f_0)$ and $g^j(\beta_0, f_0)$ are dependent for $i \neq j$. The following assumption assumes that $g^i(\beta_0, f_0)$ and $g^j(\beta_0, f_0)$ are uncorrelated and also assumes their approximations, $g^{*,i}(\beta_0, \hat{f})$ and $g^{*,j}(\beta_0, \hat{f})$ are uncorrelated as well, where $\hat{f}$ solves $L_f(\beta_0, \hat{f}) = 0$.

Assumption 7 (Conditional independence across i):
Let (i) $E\frac{1}{NT}\{\sum_i g^{i,*}(\beta_0, \hat{f}) \sum_i g^{i,*}(\beta_0, \hat{f})'\} = \frac{1}{NT}\sum_i E\{g^{i,*}(\beta_0, \hat{f}) g^{i,*}(\beta_0, \hat{f})'\}$ *and*
(ii) $E\frac{1}{NT}\{\sum_i g^{i}(\beta_0, f_0) \sum_i g^{i}(\beta_0, f_0)'\} = \frac{1}{NT}\sum_i E\{g^{i}(\beta_0, f_0) g^{i}(\beta_0, f_0)'\}$.

For correctly specified models, assumption 7 can be replaced by the requirement that the expectations exist and that $p(y_{it}|x_{it}, f_i, y_1, ...y_{i-1}, y_{i+1}, y_N) = p(y_{it}|x_{it}, f_i)$. That is, conditional on the regressors and incidental parameters, y_i is independent of the values of y of other individuals.

Assumption 8
Let (i) $\{\beta_0, f_0\} \in$ *interior of* Θ; *(ii)* $\frac{1}{T}L_f(z, \beta, f)$ *be three times continuously differentiable with respect to* f *in a neighborhood* $\mathfrak{N}$ *of* $\{\beta_0, f_0\}$ *with probability approaching one*; *(iii)* $\exists\, T_0$ *such that, for all* i, $-\infty < \frac{1}{T}L^i_{ff}(z, \beta, f) < 0$ *in a neighborhood* $\mathfrak{N}$ *of* $\{\beta_0, f_0\}$ and $T > T_0$; *(iv)* $E[||\frac{1}{T}g^i_f(z, \beta_0, f_0)||^2]$, $E[||\frac{1}{T}L^i_f(z, \beta_0, f_0)||^2]$, $E[||\frac{1}{T}L^i_{\beta f}(z, \beta_0, f_0)||^2]$, $E[||\frac{1}{T}L^i_{fff}(z, \beta_0, f_0)||^2]$, *and* $E[||\frac{1}{T}L^i_{ffff}(z, \beta_0, f_0)||^2]$ *are finite for all* i *(v)* $\frac{dg^*(\beta, \hat{f}(\beta))}{d\beta}$ *is continuous in a neighborhood* $\mathfrak{N}$ *of* $\{\beta_0\}$ *with probability approaching one and* $\sup_{\beta \in \mathfrak{N}} ||\frac{dg^*(\beta, \hat{f}(\beta))}{d\beta}|| < \infty$.

At the cost of a more complicated proof, assumption 8 (iii) only has to hold with a probability larger than $1 - e^{-cT}$ for some $c > 0$, see appendix. Assumptions 1-8 ensure that $\sqrt{NT} * g^*(\beta_0, \hat{f}) = \sqrt{NT} * g(\beta_0, f_0) + o_{ms}(1)$. After ensuring that the asymptotic distribution of $g^*(\beta_0, \hat{f})$ is determined by the asymptotic distribution of $g(\beta_0, f_0)$, we impose the following standard method of moment condition to ensure asymptotic normality of the estimator if $g(\beta, f_0)$ would be used as a moment. Let $G = E\{g_\beta(\beta_0, f_0)\} = E\{\frac{\partial g(\beta, f_0)}{\partial \beta}|_{\beta=\beta_0}\}$.

Assumption 9

Let (i) $g(z,\beta,f)$ be continuously differentiable in a neighborhood $\mathfrak{N}$ of $\{\beta_0, f_0\}$ with probability approaching one; (ii) $E[||g(z,\beta_0,f_0)||^2]$ be finite; (iii) $sup_{\beta,f\in\mathfrak{N}}||g_\beta(z,\beta,f)|| < \infty$; (iv) G be nonsingular.

Theorem 1

Let assumption 1-3 and 5-9 hold and $\{\hat{\beta},\hat{f}\} = \arg\min_{\beta,f\in\Theta}\{Q(\beta,f\}$ where β is a common parameter and $f = \{f_1, ..., f_N\}$ is a vector of individual parameters. Then

$$\sqrt{NT}(\hat{\beta} - \beta_0) \rightarrow_d N(0,\Psi)$$

where

$$\Psi = G^{-1}E\{NT * g(\beta_0,f_0)g(\beta_0,f_0)'\}G^{-1}.$$

Under the additional assumption that $EL^i_{f_i}L^i_\beta = -$'E'$L^i_{\beta f_i}$, $EL^i_\beta L^i_\beta = -$'E'$L^i_{\beta\beta}$, and $EL^2_{f_i} = -EL_{f_if_i}$ for all i, the integrated moment estimator reaches the Cramér-Rao efficiency bound and

$$\Psi = -G^{-1} = -[\frac{1}{NT}E(L_{\beta\beta}) - \frac{1}{NT}E(L_{\beta f})\{E(L_{ff})\}^{-1}E(L_{f\beta})]^{-1}.$$

Proof: See appendix.

Under the conditions of theorem 1, G can be consistently estimated by $\frac{1}{NT}\sum_i g^i_\beta(\hat{\beta},\hat{f})$ and $E\{NT * g(\beta_0,f_0)g(\beta_0,f_0)'\}$ can be consistently estimated by $\frac{1}{NT}\sum_i g^i(\hat{\beta},\hat{f})g^i(\hat{\beta},\hat{f})'$. Following LeCam (1953), Van der Vaart (1998, chapter 8) discusses how the Cramér-Rao bound gives a lower bound on the variance for regular estimators.

In section 2, we introduce the integrated moment function as an approximation to a ratio of integrals. Under regularity and smoothness conditions, this approximation is still valid if we use a prior for f. That is

$$\frac{\int g(\beta,f)e^{L^i}\pi(\beta,f)df}{\int e^{L^i}\pi(\beta,f)df} = g^{i,*}(\beta,\hat{f}) + O_p(T^{-1}). \tag{5}$$

The regularity conditions are implied by Kass et al. (1990, lemma 2, theorem 4 and 7) and are stated in the appendix. Using a prior is formally equivalent to using a mixing distribution. Let the mixing distribution be denoted by $\pi(\delta,f)$ where δ is a finite parameter

vector. Replacing $\pi(\beta, f)$ by $\pi(\delta, f)$ does not change the approximation result in equation (5). Theorem 1 is based on properties of $g^{i,*}(\beta, \hat{f})$ and still holds if we use a prior or mixing distribution for f. Thus, approximately separating β from f by using $g(\beta, f)$ results in a certain robustness against the wrong choice of prior or mixing distribution. In particular, misspecifying the initial conditions or misspecifying the dependence between the heterogeneity and the regressors only results in an asymptotic bias of low order, $O(T^{-2})$. See Wooldridge (2001) for an overview of models with mixing distributions and Gelman et al. (1995) for a motivation of Bayesian techniques.

Cox and Reid (1987) propose to reparameterize the log likelihood so that its cross derivative with respect to the nuisance parameter and parameter of interest is zero in expectation. After this reparametrization, Cox and Reid (1987) apply their conditional profile likelihood method. A limit of this approach is that the reparametrization requires that a particular differential equation has an explicit solution. It is well known and discussed by Critchley (1987) and Hills (1987), that differential equations rarely have explicit solutions. Despite this drawback, Lancaster (2000 and 2002) applies the idea of reparametrization to panel data where the fixed effect plays the role of nuisance parameter. Lancaster (2000 and 2002) then integrates out the fixed effects and derives a new estimator for the dynamic linear model with fixed effects. However, Lancaster (2000 and 2002) does not present general results. It is easy to show that the orthogonality of Cox and Reid (1987), $EL_{\beta f}(\beta_0, f_0) = 0$, for all $\{\beta_0, f_0\}$ is equivalent to 'E'$L_{\beta f}(\beta, f) = 0$ for all $\{\beta, f\}$, so that a simplified version of the arguments in this section yields that the inconsistency of this integrated likelihood estimator is $O(T^{-2})$. Applying the integrated moment estimator to the dynamic linear model yields Lancaster's (2002) estimator without a need for reparametrization. The beauty of the integrating moment approach is its combination of simplicity with generality: It is easy to compute and can be applied to a large range of models.

3.1 Time Dummies[6]

The individuals in a panel dataset may be exposed to common shocks. If these common shocks are caused by observables then one may be able to control for the common shocks by

including these observables as regressors. In other applications the common shock may not be observable to the econometrician and one prefers to include a vector of time dummies. This section shows that the conclusion of the last section, that the moment function $g(\beta, f)$ delivers asymptotically unbiased and efficient estimators, remains true in an asymptotic in which T grows slower than N but faster than $N^{1/3}$.

A time dummy has the value one for a fixed set of periods and is zero for all other periods. As a consequence, these time dummies are only estimated at rate $N^{-1/2}$ in an asymptotic with increasing T and N. We show that the common parameter still converges at a rate $(NT)^{-1/2}$ and, therefore, partition the parameter vector β. Let $\beta = \{\beta_c, \beta_d\}$ where β_c is the common parameter that appears in the quasi log likelihood contributions of all individuals for all time periods and β_d is the vector of dummy variables. In order that the derivative of $g(\beta, f)$ converges to a constant matrix G we also change the normalization. Let $g(\beta, f) = \{g^c(\beta, f), g^d(\beta, f)\}$ where $g^c(\beta, f)$ is defined as in the last section, equation (4), and $g^d(\beta, f)$ is defined similarly but is normalized by N instead of NT. That is,

$$g^d(\beta, f) = \frac{\sum_i}{N} g^{i,d}(\beta, f) = \frac{\sum_i}{N} \{L^i_{\beta_d}(\beta, f) - L^i_f(\beta, f) \frac{\int L^i_{\beta_d f}(\beta, f) e^{L^i(\beta, f)} dt}{\int L^i_{ff}(\beta, f) e^{L^i(\beta, f)} dt}\}.$$

As is shown in the appendix, $\frac{\partial g(\beta, f)}{\partial \beta}$ converges to a matrix G if N and $T \to \infty$, $\beta \to_p \beta_0$, and $f \to_p f_0$. Under assumptions given below, the integrated moment estimator is consistent and has the following variance-covariance matrix,

$$\Pi = G^{-1} E\{g(\beta_0, f_0) g(\beta_0, f_0)'\} G^{-1}.$$

Note that the upper left block of the matrix $E\{g(\beta_0, f_0) g(\beta_0, f_0)'\}$ is normalized by $\frac{1}{NT}$ and is $O((NT)^{-1})$ while the lower right block is $O(N^{-1})$. In particular, let the dimension of the common parameter, β_c, be equal to K and the dimension of the vector of time dummies, β_d, be D. Let

$$\Pi = \begin{pmatrix} \Pi_{cc} & \Pi_{cd} \\ \Pi_{dc} & \Pi_{dd} \end{pmatrix}$$

where Π_{cc} is K by K and Π_{dd} is D by D. Define $\Psi_c = NT * \Pi_{cc}$ and $\Psi_d = N * \Pi_{dd}$ and note that Ψ_c and Ψ_d converge to constant matrices for $N, T \to \infty$.

We assume[7] that T grows slower than N but faster than $N^{1/3}$.

Assumption 10 (Asymptotics):

Let $T \propto N^{\alpha}$ *where* $\frac{1}{3} < \alpha < 1$.

The following assumption formalizes the notion that β_d is the vector of time dummies. It is satisfied if, for example, each time dummy only appears in the likelihood contribution of a fixed set of periods.[8]. Let $g^c_{\beta_c} = \frac{\partial g^c(\beta,f_0)}{\partial \beta_c}|_{\beta=\beta_0}$, $g^c_{\beta_d} = \frac{\partial g^d(\beta,f_0)}{\partial \beta_d}|_{\beta=\beta_0}$, $g^d_{\beta_c} = \frac{\partial g^d(\beta,f_0)}{\partial \beta_c}|_{\beta=\beta_0}$ and $g^d_{\beta_d} = \frac{\partial g^d(\beta,f_0)}{\partial \beta_d}|_{\beta=\beta_0}$.

Assumption 11 (Time Dummies):

Let

$$g^c_{\beta_d} \text{ be } O_p(T^{-1}); \; E\{g^c(\beta_0,f_0)g^d(\beta_0,f_0)\} \text{ be } O((NT)^{-1});$$
$$\{Eg^c_{\beta_c} - Eg^c_{\beta_d}[Eg^d_{\beta_d}]^{-1}Eg^d_{\beta_c}\}^{-1} \text{ be } O(1).$$

Theorem 2

Let assumption 1-3,5 and 7-11 hold and $\{\hat{\beta},\hat{f}\} = \arg\min_{\beta,f\in\Theta}\{g(\beta,f)'g(\beta,f)+h_f(\beta,f)'h_f(\beta,f)\}$. *Then*

$$\begin{aligned}\sqrt{NT}(\hat{\beta}_c - \beta_0) &\rightarrow {}_dN(0,\Psi_c) \text{ and}\\ \sqrt{N}(\hat{\beta}_d - \beta_0) &\rightarrow {}_dN(0,\Psi_d),\end{aligned}$$

where $\Psi_c = NT * \Pi_{cc}$ *and* $\Psi_d = N * \Pi_{dd}$. *Under the additional assumption that* $EL^i_{f_i}L^i_{\beta} = -$'$E$'$L^i_{\beta f_i}$, $EL^i_{\beta}L^i_{\beta} = -$'$E$'$L^i_{\beta\beta}$, *and* $EL^2_{f_i} = -EL_{f_if_i}$ *for all* i *the integrated moment estimator reaches the Cramér-Rao efficiency bound and*

$$\Pi = \begin{pmatrix} \Pi_{cc} & \Pi_{cd} \\ \Pi_{dc} & \Pi_{dd} \end{pmatrix} = -[E(L_{\beta\beta}) - E(L_{\beta f})\{E(L_{ff})\}^{-1}E(L_{f\beta})]^{-1},$$

where Π_{cc} is K by K and Π_{dd} is D by D.

$$\begin{aligned}\Psi_c &= NT * \Pi_{cc},\\ \Psi_d &= N * \Pi_{dd} = -[\frac{1}{N}E(L_{\beta_d\beta_d}) - \frac{1}{N}E(L_{\beta_d f})\{E(L_{ff})\}^{-1}E(L_{f\beta_d})].\end{aligned}$$

Proof: See appendix.

3.2 General Predetermined Regressors

Neyman and Scott (1948) describe the incidental parameter problem by showing that the maximum likelihood estimator fails to be consistent in a couple of examples. The regressors of these examples are all exogenous but the incidental parameter problem obviously remains when the assumption of exogeneity is relaxed. In the previous sections, we derived an approximate solution to the incidental parameter problem by using a moment function that approximately separates the nuisance parameters from the parameters of interest. As discussed above, this framework allows for predetermined regressors that are lagged dependent variables or whose density is known, up to a finite parameter vector. In some applications, however, one is not willing to specify the stochastic process of such explanatory variables. In their handbook chapter, Arellano and Honoré (2001) note "almost nothing is known about nonlinear models with general predetermined variables". This section derives new estimators for (quasi) likelihood models with general predetermined regressors and incidental parameters[9]. For models with weakly exogenous regressors, the technique to condition on a sufficient statistic is unlikely to work. A sufficient statistic for the incidental parameter f_i would be a function of $y_{i1}, ..., y_{iT}$. Conditioning requires the distribution of the sufficient statistic conditional on the predetermined regressors of all periods. This distribution is not specified since the regressors are only required to be predetermined.

Analogously to the last section, we develop a moment function $g(\beta, f)$ which approximately separates β and f, i.e. $Eg(\beta_0, f_0) = 0$ and $Eg_f(\beta_0, f_0) = 0$. However, we can no longer calculate a probability distribution for predetermined variables. We, therefore, just use the fact that the regressors are predetermined and predict $L^{it}_{\beta f}$ using the predetermined[10] $x^t_i = x\{x_{i1}, ..., x_{it}\}$ and the parameters β and f. In particular, consider

$$g^i(\beta, f) = \sum_{t=1}^{T} \{L^{it}_{\beta}(\beta, f) - \frac{\text{`}E\text{'}(L^{it}_{\beta f}|x^t_{i,}, \beta, f)}{\text{`}E\text{'}(L^{it}_{ff}|x^t_i, \beta, f)} L^{it}_{f}(\beta, f)\} \tag{6}$$

$$\text{where } \frac{\text{`}E\text{'}(L^{it}_{\beta f}|x^t_{i,}, \beta, f)}{\text{`}E\text{'}(L^{it}_{ff}|x^t_i, \beta, f)} = \frac{\int L^{it}_{\beta f}(\beta, f) e^{L^{it}(\beta, f)} dt}{\int L^{it}_{ff}(\beta, f) e^{L^{it}(\beta, f)} dt}.$$

Analogously to the last sections, let $g(\beta,f)=\frac{\sum_i}{NT}\{g^i(\beta,f)\}$. This yields

$$g(\beta,f)=\frac{L_\beta(\beta,f)}{NT}-\frac{\sum_{i=1}^N}{N}\frac{\sum_{t=1}^T}{T}\{\frac{\text{`}E\text{'}(L^{it}_{\beta f}|x^t_{i,},\beta,f)}{\text{`}E\text{'}(L^{it}_{ff}|x^t_i,\beta,f)}L^{it}_f(\beta,f)\} \tag{7}$$

and note that $Eg_f(\beta_0,f_0)=0$. Suppose the incidental parameter is a fixed effect and appears through an index in the likelihood function. An example is the probit model where $Pr(Y_{it}=1)=\Phi(\mu_{it})$ and $\mu_{it}=x_{it}\beta+f_i$. Using the moment function of equation (7) does not identify the parameter of the regressors appearing in the index. Therefore, we use the following moment if the incidental parameter appears through an index in the likelihood function. Let $L_{\mu_{it}}$ denote the first derivative of the log likelihood with respect to μ_{it} and let $L_{\mu_{it}\mu_{it}}$ denote the second derivative. Consider

$$g^i(\beta,f)=\sum_{t=1}^{T-1}\frac{\partial\mu_{it}}{\partial\beta}\{L^{it}_{\mu_{it}}-\frac{\text{`}E\text{'}(\frac{\partial\mu_{it}}{\partial f}L^{it}_{\mu_{it}\mu_{it}}|x^t_{i,},\beta,f)}{\text{`}E\text{'}(\frac{\partial\mu_{i,t+1}}{\partial f}L^{i,t+1}_{\mu_{i,t+1}\mu_{i,t+1}}|x^{t+1}_i,\beta,f)}L^{i,t+1}_{\mu_{i,t+1}}\}, \tag{8}$$

where x_{it} denotes a vector of predetermined variables. If the incidental parameter is a fixed effect, then we have

$$g^i(\beta,f)=\sum_{t=1}^{T-1}x_{it}\{L^{it}_{\mu_{it}}-\frac{\text{`}E\text{'}(L^{it}_{\mu_{it}\mu_{it}}|x^t_{i,},\beta,f)}{\text{`}E\text{'}(L^{i,t+1}_{\mu_{i,t+1}\mu_{i,t+1}}|x^{t+1}_i,\beta,f)}L^{i,t+1}_{\mu_{i,t+1}}\}. \tag{9}$$

The ratio $\frac{\text{`}E\text{'}(L^{it}_{\mu_{it}\mu_{it}}|x^t_{i,},\beta,f)}{\text{`}E\text{'}(L^{i,t+1}_{\mu_{i,t+1}\mu_{i,t+1}}|x^{t+1}_i,\beta,f)}$ has a similar role as the ratio $\frac{\text{`}E\text{'}(L^{it}_{\beta f}|x^t_{i,},\beta,f)}{\text{`}E\text{'}(L^{it}_{ff}|x^t_i,\beta,f)}$ above. However, notice that the expectations are now conditional on x^t_i and x^{t+1}_i, respectively. The idea behind the moment functions is the same and in both cases, β is approximately separated[11] from f if

$E(L^{it}_{\mu_{it}\mu_{it}}|x^t_{i,},\beta_0,f_0)=\text{`}E\text{'}(L^{it}_{\mu_{it}\mu_{it}}|x^t_{i,},\beta_0,f_0)\}$,

$$Eg^i_f(\beta_0,f_0)=E\sum_{t=1}^{T}x_{it}\{L^{it}_{\mu_{it}\mu_{it}}-E(L^{it}_{\mu_{it}\mu_{it}}|x^t_{i,},\beta_0,f_0)\}=0.$$

The same vector moment can be used for the regressor coefficients of the transformation model with a parametric transformation or the Weibull model[12]. Suppose $y_{it}=H(\mu_{it})+\varepsilon_{it}$ where $\mu_{it}=X_{it}\beta+f_i$, $H()$ is a parametric function, $E(\varepsilon_{it}|x^t_i)=0$ and $E(\varepsilon^2_{it}|x^t_i)<\infty$. Assuming normality and homoscedasticity of the error term yields a quasi likelihood. Applying

the moment function of equation (9) to the linear model with predetermined variables yields a familiar[13] moment function,

$$g^i(\beta) = \sum_{t=1}^{T-1} x_{it}(\varepsilon_{it} - \varepsilon_{i,t+1}).$$

We thus use the moment vector function of equation (6). If the incidental parameter is a fixed effect and appears through an index in the likelihood function then we replace those moments by (8). This yields $g(\beta, f)$. We then use the same objective function as in the last section,

$$Q(\beta, f) = -\{g^*(\beta, f)\}'\{g^*(\beta, f)\} - \{h(\beta, f)\}'\{h(\beta, f)\}$$

where

$$\begin{aligned} g^*(\beta, f) &= \frac{\sum_i}{NT}\{g^i(f) - \frac{1}{2}\frac{g^i_{ff}(f)}{L^i_{ff}(f)} + \frac{1}{2}\frac{L^i_{fff}(f)g^i_f(f)}{\{L^i_{ff}(f)\}^2}\} \\ h(\beta, f) &= \frac{L_f(\beta, f)}{T}. \end{aligned}$$

Imposing the same assumptions as in Theorem 1, except that x is exogenous, gives the following theorem.

Theorem 3

Let $\{x_i, y_i\}$ be strictly stationary and ergodic for all i and let x_i be predetermined for all i. Let assumption 1-2 and 5-9 hold and $\hat{\beta} = \arg\min_{\beta, f \in \Theta}\{Q(\beta, f)\}$. Then

$$\sqrt{NT}(\hat{\beta} - \beta_0) \rightarrow_d N(0, \Psi)$$

where

$$\Psi = G^{-1}E\{NT * g(\beta_0, f_0)g(\beta_0, f_0)'\}G^{-1}.$$

Proof: See appendix.

4 Simulations

Heckman (1981a,b) introduces the distinction between "state dependence" and "spurious correlation" as two sources of correlation over time between outcomes of labor market participation or other decisions by an individual. "State dependence" is caused by a lag dependent variable whereas "spurious correlation" is the result of heterogeneity. Heckman (1981b) presents a simulation study and notes that the fixed effect estimator is biased, in particular if lagged dependent variables are present. The simulation study of this section is intended to illustrate the integrated likelihood approach and shows that the integrated moment estimator is unbiased in the simulation designs considered by Heckman (1981b).

Exogenous Regressors

Heckman (1981b) assumes the following. Let

$$Y_{it}^{*} = X_{it}\beta + \sigma_t \tau_i + \varepsilon_{it},$$

where X_{it} is generated by a Nerlove (1971) process

$$\begin{aligned} X_{it} &= 0.1t + 0.5X_{i,t-1} + U_{it,} \quad U_{it} \sim U\left(-0.5, 0.5\right) \\ X_{it} &= 5 + 10U_{i0,} \\ \tau_i &\sim N\left(0,1\right), \quad \varepsilon_{it} \sim N\left(0,1\right) \\ Y_{it} &= 1\left[Y_{it}^{*} > 0\right]. \end{aligned}$$

Using the fixed effect probit estimator yields the following average estimates for β.

Table 1

	$\beta = 1$	$\beta = -0.1$	$\beta = -1$
$\sigma_t^2 = 3$	1.3052	−0.1097	−1.2422
$\sigma_t^2 = 1$	1.1598	−0.1042	−1.2026
$\sigma_t^2 = 0.5$	1.1028	−0.1056	−1.1825

Based on 1000 replications

These results differ somewhat from Heckman's (1981b) findings, probably because the number of replications is larger[14] but the conclusion, namely that the fixed effect probit estimator is biased, remains the same.

As is apparent from the following table, the integrated moment estimator is unbiased.

Table 2

	$\beta = 1$	$\beta = -0.1$	$\beta = -1$
$\sigma_t^2 = 3$ (MSE)	0.9969 (0.0216)	−0.0879 (0.0208)	−0.9998 (0.0219)
$\sigma_t^2 = 1$ (MSE)	1.0070 (0.0209)	−0.0978 (0.0211)	−0.9964 (0.0212)
$\sigma_t^2 = 0.5$ (MSE)	0.9973 (0.0214)	−0.1045 (0.0200)	−1.0010 (0.0212)

Based on 1000 replications

See the appendix for details.

Lagged Dependent Variables

The data generating process is the same as in the previous case but now a lagged dependent variable is included as an explanatory variable. In particular,

$$Y_{it}^* = X_{it}\beta + Y_{i,t-1}\gamma + \sigma_t\tau_i + \varepsilon_{it}.$$

Using Heckman's (1981b) design, the conditional fixed effect probit estimator yields the following average estimates for β and γ for various values of the parameters.

Table 3

		$\sigma_t^2 = 3$ $\beta = -0.1$	$\sigma_t^2 = 3$ $\beta = 1$	$\sigma_t^2 = 3$ $\beta = 0$
$\gamma = 0.5$	$\widehat{\gamma}$ (MSE)	−0.4401 (0.9307)	1.7461 (194.22)	−0.5424 (0.9604)
$\gamma = 0.5$	$\widehat{\beta}$ (MSE)	−0.2008 (0.0442)	3.7034 (55.4247)	0.0376 (0.0227)
$\gamma = 0.1$	$\widehat{\gamma}$ (MSE)	−0.3898 (0.2568)	−4.5833 (395.79)	−0.4785 (0.3843)
$\gamma = 0.1$	$\widehat{\beta}$ (MSE)	−0.1353 (0.0199)	13.0668 (1011.52)	0.0312 (0.0305)
		$\sigma_t^2 = 1$ $\beta = -0.1$	$\sigma_t^2 = 1$ $\beta = 1$	$\sigma_t^2 = 1$ $\beta = 0$
$\gamma = 0.5$	$\widehat{\gamma}$ (MSE)	−0.3503 (0.7520)	0.5079 (0.0003)	−0.2907 (0.6446)
$\gamma = 0.5$	$\widehat{\beta}$ (MSE)	−0.1167 (0.0102)	1.0444 (0.0051)	0.0513 (0.0105)
$\gamma = 0.1$	$\widehat{\gamma}$ (MSE)	−0.4066 (0.2807)	0.1174 (0.0006)	−0.4312 (0.2999)
$\gamma = 0.1$	$\widehat{\beta}$ (MSE)	−0.1692 (0.0143)	1.0671 (0.0106)	0.0457 (0.0024)

Based on 1000 replications

The following table contains the mean of the estimates for β and γ as well as the mean squared error for the integrated moment estimator.

Table 4

		$\sigma_t^2 = 3$ $\beta = -0.1$	$\sigma_t^2 = 3$ $\beta = 1$	$\sigma_t^2 = 3$ $\beta = 0$
$\gamma = 0.5$	$\widehat{\gamma}$ (MSE)	0.4973 (0.0214)	0.4922 (0.0202)	0.5058 (0.0202)
$\gamma = 0.5$	$\widehat{\beta}$ (MSE)	−0.1041 (0.0211)	1.0058 (0.0201)	−0.0025 (0.0213)
$\gamma = 0.1$	$\widehat{\gamma}$ (MSE)	0.1008 (0.0207)	0.0946 (0.0224)	0.1005 (0.0205)
$\gamma = 0.1$	$\widehat{\beta}$ (MSE)	−0.0973 (0.0196)	0.9994 (0.0215)	−0.0043 (0.0213)
		$\sigma_t^2 = 1$ $\beta = -0.1$	$\sigma_t^2 = 1$ $\beta = 1$	$\sigma_t^2 = 1$ $\beta = 0$
$\gamma = 0.5$	$\widehat{\gamma}$ (MSE)	0.5019 (0.0210)	0.4952 (0.0204)	0.4908 (0.0197)
$\gamma = 0.5$	$\widehat{\beta}$ (MSE)	−0.1081 (0.0212)	0.9933 (0.0215)	0.0016 (0.0209)
$\gamma = 0.1$	$\widehat{\gamma}$ (MSE)	0.0942 (0.0213)	0.1073 (0.0206)	0.0953 (0.0214)
$\gamma = 0.1$	$\widehat{\beta}$ (MSE)	−0.0959 (0.0208)	1.0034 (0.0205)	0.0052 (0.0212)

Based on 1000 replications

5 Conclusion

This paper develops the integrated moment estimator. It shows that the integrated moment estimator yields an approximate solution to the incidental parameter problem of Neyman and Scott (1948). A nice feature of this approximate solution is that estimators that rely on 'differencing' are shown to be special cases and that the estimator is efficient in an asymptotic where T increases slowly, in particular, T is only required to increase faster than $N^{1/3}$.

In their conclusion, Arellano and Honoré (2001) note that almost nothing is known about nonlinear models with general predetermined variables. Using approximate separation and the integrated moment function, we derive new estimators for (quasi) likelihood models with incidental parameters and predetermined regressors. It thus seems that approximate parameter separation and the integrated moment function are very promising approaches to study models with general predetermined variables and many nuisance parameters. In particular, the idea of approximate separation is not limited to quasi-likelihood functions and it seems to be possible to apply it to other smooth objective functions such as the smoothed maximum score estimator of Horowitz (1992). This is a topic of further study.

6 Appendices

Proposition 1

Suppose $\{\hat{\beta}, \hat{f}\} = \arg\min_{\beta, f \in \Theta}\{Q(\beta, f\}$. *Let assumption 1-4 hold. Then* $\hat{\beta} \rightarrow_p \beta_0$ and $\hat{f} \rightarrow_p f_0$.

Proof: Assumption 1 implies that $Q_0(\beta, f) = -\{Eg(\beta, f)\}'\{Eg(\beta, f)\} - \{Eh(\beta, f)\}'\{Eh(\beta, f)\}$ is uniquely solved for $\{\beta, f\} = \{\beta_0, f_0\}$. Assumption 2-4 and the compactness assumption of assumption 1 imply that (1) $g(\beta, f)$, $\frac{g^i_{ff}(\beta,f)}{L^i_{ff}(\beta,f)}$, $\frac{L^i_{fff}(\beta,f)g^i_f(\beta,f)}{\{L^i_{ff}(\beta,f)\}^2}$ and $h(\beta, f)$ converges uniformly to their expectations and (2) $Eg(\beta, f)$, $E\frac{g^i_{ff}(\beta,f)}{L^i_{ff}(\beta,f)}$, $E\frac{L^i_{fff}(\beta,f)g^i_f(\beta,f)}{\{L^i_{ff}(\beta,f)\}^2}$ and $Eh(\beta, f)$ are continuous, see Newey and McFadden (1994, Lemma 2.4, where Newey and McFadden note that "The conclusion [of Lemma 2.4] remains true if the i.i.d. hypothesis is changed to strict stationarity and ergodicity of z_i", see Newey and McFadden (1994, page 2129)). This implies that $g^*(\beta, f) - g(\beta, f) = \frac{1}{T}\frac{\sum_i}{N}\{-\frac{1}{2}\frac{g^i_{ff}(\beta,f)}{L^i_{ff}(\beta,f)} + \frac{1}{2}\frac{L^i_{fff}(\beta,f)g^i_f(\beta,f)}{\{L^i_{ff}(\beta,f)\}^2}\}$ converges uniformly to zero so that $Q(\beta, f)$ converges uniformly to $Q_0(\beta, f)$. All assumptions of Newey and McFadden (1994, Theorem 2.1) are satisfied and consistency follows. Q.E.D.

We frequently use the fact that the product of three normally distributed random variables has expectation zero. Seminar participants were often surprised by this so we state the result as a lemma.

Lemma 1: Let η_1, η_2 and η_3 be scalars and let

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} \sim N(0, \Sigma) \text{ where } \Sigma \text{ is a } 3 \times 3 \text{ matrix and } \det(\Sigma) < \infty.$$

Then $E(\eta_1\eta_2\eta_3) = 0$.

Proof: We first consider $\det(\Sigma) > 0$. Note that $E(\eta_1|\eta_2, \eta_3) = \gamma_2\eta_2 + \gamma_3\eta_3$ where $\gamma_2 =$

$\frac{\Sigma_{12}\Sigma_{33}-\Sigma_{13}\Sigma_{23}}{\Sigma_{22}\Sigma_{33}-(\Sigma_{23})^2}$ and $\gamma_3 = \frac{\Sigma_{13}\Sigma_{22}-\Sigma_{12}\Sigma_{23}}{\Sigma_{22}\Sigma_{33}-(\Sigma_{23})^2}$. This gives

$$\begin{aligned} E(\eta_1\eta_2\eta_3) &= E\{E(\eta_1\eta_2\eta_3|\eta_2,\eta_3) \\ &= E\{(\gamma_2\eta_2+\gamma_3\eta_3)\eta_2\eta_3\} \\ &= E\{E(\gamma_2\eta_2^2\eta_3|\eta_3)\}+E\{E(\gamma_3\eta_2\eta_3^2|\eta_2)\} \\ &= \Sigma_{22}\gamma_2E(\eta_3)+\Sigma_{33}\gamma_3E(\eta_2)=0, \end{aligned}$$

using the fact that $E(\eta_l^2|\eta_k) = E(\eta_l^2) = \Sigma_{ll}$ for $l \neq k$.

For $\det(\Sigma) = 0$, we have (a) η_1 is a linear combination of η_2 and η_3 or (b) $\eta_2 = c\eta_3$ for some finite c. If (a) applies then $\eta_1 = \gamma_2\eta_2 + \gamma_3\eta_3$ and the above proof holds. If $\eta_2 = c\eta_3$ then $E(\eta_1|\eta_2,\eta_3) = E(\eta_1|\eta_2) = \frac{\Sigma_{12}}{\Sigma_{22}}\eta_2$ and $E(\eta_1\eta_2\eta_3) = E(\frac{\Sigma_{12}}{\Sigma_{22}}\eta_2^2\eta_3) = E(\Sigma_{12}\eta_3) = 0$.

Q.E.D.

Differentiating $g^i(\beta, f)$

$$\begin{aligned} g^i(\beta,f) &= L^i_\beta(\beta,f) - L^i_f(\beta,f)\delta(\beta,f) \\ \text{where } \delta^i(\beta,f) &= \frac{\int L^i_{\beta f}(\beta,f)e^{L^i(\beta,f)}dt}{\int L^i_{ff}(\beta,f)e^{L^i(\beta,f)}dt}. \end{aligned}$$

Differentiating $g^i(\beta, f)$ with respect to f gives

$$g^i_f(\beta,f) = L^i_{\beta f}(\beta,f) - L^i_{ff}(\beta,f)\frac{\int L^i_{\beta f}(\beta,f)e^{L^i(\beta,f)}dt}{\int L^i_{ff}(\beta,f)e^{L^i(\beta,f)}dt} + L^i_f(\beta,f)\delta^i_f(\beta,f)$$

where

$$\begin{aligned} \delta^i_f(\beta,f) &= -\frac{\int\{L^i_{\beta ff}(\beta,f)+L^i_{\beta f}L^i_f\}e^{L^i(\beta,f)}dt}{\int L^i_{ff}(\beta,f)e^{L^i(\beta,f)}dt} \\ &\quad +\frac{\int\{L^i_{fff}(\beta,f)+L^i_{ff}(\beta,f)L^i_f(\beta,f)\}e^{L^i(\beta,f)}dt\int L^i_{\beta f}(\beta,f)e^{L^i(\beta,f)}dt}{\{\int L^i_{ff}(\beta,f)e^{L^i(\beta,f)}dt\}^2}. \end{aligned}$$

Note that $\delta^i_f(\beta, f)$ is nonstochastic and that,

$$Eg^i_f(\beta_0,f_0) = EL^i_{\beta f} - \frac{\int L^i_{\beta f}(\beta_0,f_0)e^{L^i(\beta_0,f_0)}dt}{\int L^i_{ff}(\beta_0,f_0)e^{L^i(\beta_0,f_0)}dt}EL^i_{ff},$$

since $EL_f^i(\beta_0, f_0) = 0$ by assumption 1. Differentiating $g^i(\beta, f)$ with respect to β gives

$$\begin{aligned} g_\beta^i(\beta, f) &= \frac{\sum_i}{N}[L_{\beta\beta}^i(\beta, f) + L_{\beta f}^i(\beta, f)\frac{\int L_{\beta f}^i(\beta, f)e^{L^i(\beta,f)}dt}{\int L_{ff}^i(\beta, f)e^{L^i(\beta,f)}dt} \\ &+L_f^i(\beta, f)\frac{\int\{L_{\beta\beta f}^i(\beta, f) + L_{\beta f}^i(\beta, f)L_\beta^i(\beta, f)\}e^{L^i(\beta,f)}dt}{\int L_{ff}^i(\beta, f)e^{L^i(\beta,f)}dt} \\ &-L_f^i(\beta, f)\frac{\int L_{\beta f}^i(\beta, f)e^{L^i(\beta,f)}dt\int\{L_{\beta ff}^i(\beta, f) + L_{ff}^i(\beta, f)L_\beta^i(\beta, f)\}e^{L^i(\beta,f)}dt}{(\int L_{ff}^i(\beta, f)e^{L^i(\beta,f)}dt)^2})]. \end{aligned}$$

$$\begin{aligned} g_\beta^i(\beta_0, f_0) &= \frac{\sum_i}{N}[L_{\beta\beta}^i + L_{\beta f}^i\frac{\text{`}E\text{'}L_{\beta f}^i}{\text{`}E\text{'}L_{ff}^i} \\ &+L_f^i\frac{\text{`}E\text{'}L_{\beta\beta f}^i + \text{`}E\text{'}L_{\beta f}^iL_\beta^i}{\text{`}E\text{'}L_{ff}^i} \\ &-L_f^i\frac{\text{`}E\text{'}L_{\beta f}^i\{\text{`}E\text{'}(L_{\beta ff}^i + L_{ff}^iL_\beta^i)\}}{(\text{`}E\text{'}L_{ff}^i)^2}], \end{aligned}$$

where we omitted the argument if a function is evaluated at (β_0, f_0).

Stronger alternatives to Assumption 5

Assumption 5 (Approximate separation): Let (i) $Eg_f(\beta_0, f_0) = 0$ and (ii) $EL_f(\beta_0, f_0)g_f(\beta_0, f_0) + Eg_{ff}(\beta_0, f_0) = 0$.

The following two assumptions imply assumption 5.

Assumption 5*: $\text{`}E\text{'}L_{ff} = EL_{ff}$, $\text{`}E\text{'}L_{\beta f} = EL_{\beta f}$; $E(L_f)^2 = EL_{ff}$, $\text{`}E\text{'}L_{\beta ff} = EL_{\beta ff}$, $\text{`}E\text{'}L_{\beta f}L_f = EL_{\beta f}L_f$ and $\text{`}E\text{'}L_{fff} = EL_{fff}$.

Assumption 5**: (Correct specification): $p(y_i|x_i, \beta, f) = e^{L^i(\beta,f|x_i,y_i)}$ for all i.

In particular, assumption 5* implies assumption 5 since

$$\begin{aligned} & Eg^i_{ff}(\beta_0, f_0) + EL^i_f(\beta_0, f_0) g^i_f(\beta_0, f_0) \\ = \; & EL^i_{\beta ff} - EL^i_{fff} \frac{\text{`}E\text{'}L^i_{\beta f}}{\text{`}E\text{'}L^i_{ff}} - 2\text{`}E\text{'}(L^i_{\beta ff} + L^i_{\beta f} L^i_f) \\ & + 2 \frac{\text{`}E\text{'}L^i_{\beta f}\{\text{`}E\text{'}(L^i_{fff} + L^i_{ff} L^i_f)\}}{\text{`}E\text{'}L^i_{ff}} \\ & E[L^i_{\beta f} L^i_f - L^i_{ff} L^i_f \frac{\text{`}E\text{'}L^i_{\beta f}}{\text{`}E\text{'}L^i_{ff}} - L^i_f L^i_f \frac{\text{`}E\text{'}(L^i_{\beta ff} + L^i_{\beta f} L^i_f)}{\text{`}E\text{'}L^i_{ff}} \\ & + L^i_f L^i_f \frac{\text{`}E\text{'}L^i_{\beta f}\{\text{`}E\text{'}(L^i_{fff} + L^i_{ff} L^i_f)\}}{(\text{`}E\text{'}L^i_{ff})^2}] = 0. \end{aligned}$$

Assumption 5** implies assumption 5 since $\text{`}E\text{'}L^i_{\beta f} = EL^i_{\beta f}$ so that $Eg^i_f = 0$ and differentiating both sides of the equation $Eg^i_f = 0$ with respect to f gives $E\{g^i_f L^i_f + g^i_{ff}\} = 0$.

Lemma 2: *Let the assumptions of theorem 1 hold, including the assumption that* $E\{g^i_f L^i_f + g^i_{ff}\} = 0$. *Then* $E\{\frac{g^i_f L^i_f + g^i_{ff}}{L^i_{ff}}\}$ is $O(T^{-1})$.

Proof: $E\{g^i_f L^i_f + g^i_{ff}\} = 0$ implies that $E\{\frac{g^i_f L^i_f + g^i_{ff}}{EL^i_{ff}}\} = 0$. Thus,

$$\begin{aligned} E\{\frac{g^i_f L^i_f + g^i_{ff}}{L^i_{ff}}\} &= E\{\frac{g^i_f L^i_f + g^i_{ff}}{L^i_{ff}}\} - E\{\frac{g^i_f L^i_f + g^i_{ff}}{EL^i_{ff}}\} \\ &= E\{\frac{(g^i_f L^i_f + g^i_{ff})(EL^i_{ff} - L^i_{ff})}{L^i_{ff} EL^i_{ff}}\} \\ &= E\{\frac{(g^i_f L^i_f + g^i_{ff})(EL^i_{ff} - L^i_{ff})}{(EL^i_{ff})^2}\} + O(T^{-1}) \end{aligned}$$

since $\frac{(EL^i_{ff} - L^i_{ff})^2}{(EL^i_{ff})^2}$ is $O(T^{-1})$. Thus,

$$E\{\frac{g^i_f L^i_f + g^i_{ff}}{L^i_{ff}}\} = \frac{1}{\sqrt{T}} E\{\frac{T}{EL^i_{ff}} \frac{T}{EL^i_{ff}} \frac{g^i_f}{\sqrt{T}} \frac{L^i_f}{\sqrt{T}} \frac{EL^i_{ff} - L^i_{ff}}{\sqrt{T}}\} + O(T^{-1}).$$

The terms $\frac{g^i_f}{\sqrt{T}}$, $\frac{L^i_f}{\sqrt{T}}$, and $\frac{(EL^i_{ff} - L^i_{ff})}{\sqrt{T}}$ each have expectation zero and their asymptotic distribution is normal with mean zero (with remainder term $O_{ms}(T^{-1/2})$). Lemma 1 states that the product of three normally distributed stochasts, with mean zero, has expectation zero and the result follows. *Q.E.D.*

Lemma 3: *Let the assumptions of theorem 1 hold, including the assumption that* $Eg_f^i = 0$. *Then* $E[\{\frac{L_{ff}^i - (L_f^i)^2}{L_{ff}^i}\}\{\frac{L_{fff}^i g_f^i}{(L_{ff}^i)^2}\}]$ *is* $O(T^{-1})$.

Proof: $Eg_f^i = 0$ and $E[||\frac{1}{T} g_f||^2] < \infty$ imply that g_f^i is $O(\sqrt{T})$. In particular,

$$\begin{aligned} E[\{\frac{L_{ff}^i - (L_f^i)^2}{L_{ff}^i}\}\{\frac{L_{fff}^i g_f^i}{(L_{ff}^i)^2}\}] &= E[\{\frac{L_{fff}^i g_f^i}{(EL_{ff}^i)^2}\}] - E[\{\frac{(L_f^i)^2}{EL_{ff}^i}\}\{\frac{L_{fff}^i g_f^i}{(EL_{ff}^i)^2}\}] + O(T^{-1}) \\ &= E[\{\frac{g_f^i (L_{fff}^i - EL_{fff}^i)}{(EL_{ff}^i)^2}\}] - E[\{\frac{(L_f^i)^2}{EL_{ff}^i}\}\{\frac{L_{fff}^i g_f^i}{(EL_{ff}^i)^2}\}] + O(T^{-1}) \\ &= -\frac{1}{\sqrt{T}} E\{(\frac{T}{EL_{ff}^i})^3 \frac{L_f^i}{\sqrt{T}} \frac{L_f^i}{\sqrt{T}} \frac{g_f^i}{\sqrt{T}} \frac{EL_{fff}^i}{T}\} + O(T^{-1}) \end{aligned}$$

since $\frac{L_{fff}^i - EL_{fff}^i}{\sqrt{T}}$, $\frac{L_f^i}{\sqrt{T}}$ and $\frac{g_f^i}{\sqrt{T}}$ have an asymptotic distribution that is normal with mean zero. *Q.E.D.*

Taylor expansion

The advantage of panel data over a single time series is that one can average over individuals. We often average over random variables with zero mean and bounded variance. We therefore introduce the following notation concerning a sequence of error terms that is bounded in mean square. The random variable W_T is $O_{ms}^{\text{zero mean}}(T^{-1/2})$ if and only if $\text{var}(W_T)$ is $O(T^{-1})$ *and* $EW_T = 0$. The definition of $\hat{f}$ implies that

$$L_f^i(\hat{f}) = L_f^i + (\hat{f} - f_0) L_{ff}^i(\bar{f}) = 0,$$

where $\bar{f}$ denotes an intermediate value, $\bar{f} \in [f_0, \hat{f}]$ and we omitted the argument when evaluating a function at the true value. Similarly,

$$L_f^i(\hat{f}) = L_f^i + (\hat{f} - f_0) L_{ff}^i + \frac{1}{2}(\hat{f} - f_0)^2 L_{fff}^i(\overline{\bar{f}}) = 0,$$

where $\overline{\bar{f}}$ denotes another intermediate value, $\overline{\bar{f}} \in [f_0, \hat{f}]$. Combining the last two equations gives

$$(\hat{f} - f_0) = -\frac{L_f^i}{L_{ff}^i} - \frac{1}{2}\{\frac{L_f^i}{L_{ff}^i(\bar{f})}\}^2 \frac{L_{fff}^i(\overline{\bar{f}})}{L_{ff}^i}.$$

Under the assumptions of theorem 1, we have

$$(\hat{f} - f_0) = -\frac{L_f^i}{L_{ff}^i} + O_{ms}(T^{-1}).$$

More algebra gives

$$\begin{aligned}(\hat{f}-f_0)^2 &= (\frac{L_f^i}{L_{ff}^i})^2 + O_{ms}(T^{-3/2}) \\ \frac{\partial \hat{f}_i}{\partial \beta} &= -\frac{L_{\beta f}^i}{L_{ff}^i} + \frac{L_f^i L_{\beta ff}^i}{(L_{ff}^i)^2} + O_{ms}(T^{-1/2}) \\ &= -\frac{L_{\beta f}^i}{L_{ff}^i} + O_{ms}(T^{-1/2}).\end{aligned}$$

Similarly,

$$L_f^i(\hat{f}) = L_f^i + (\hat{f}-f_0)L_{ff}^i + \frac{1}{2}(\hat{f}-f_0)^2 L_{fff}^i + \frac{1}{6}(\hat{f}-f_0)^3 L_{ffff}^i(\bar{f}) = 0$$

and

$$(\hat{f}-f_0) = -\frac{L_f^i}{L_{ff}^i} - \frac{1}{2}\frac{L_{fff}^i}{L_{ff}^i}(\frac{L_f^i}{L_{ff}^i})^2 + O_{ms}(T^{-3/2}).$$

If we replace the assumption $E[||\frac{1}{T}L_{ffff}(z,\beta_0,f_0)||^2]$ by $E[||\frac{1}{T}L_{ffff}(z,\beta_0,f_0)||^2]$ then we have

$$(\hat{f}-f_0) = -\frac{L_f^i}{L_{ff}^i} - \frac{1}{2}\frac{L_{fff}^i}{L_{ff}^i}(\frac{L_f^i}{L_{ff}^i})^2 + \frac{\eta_i}{T^{3/2}} + O_p(T^{-2})$$

where η_i is normally distributed. We maintain the assumption $E[||\frac{1}{T}L_{ffff}(z,\beta_0,f_0)||]$ since it simplifies the proof of theorem 1.

Lemma 4: *Let the assumptions of theorem 1 hold. Then*

$$\begin{aligned}g^*(\beta,\hat{f}) &= g(\beta,\hat{f}) + O_{ms}(T^{-1}), \text{ for } \beta \text{ in a neighborhood } \mathfrak{N} \text{ of } \beta_0, \\ g^{i,*}(\beta_0) &= g^i(\beta_0, f_{i,0}) + O_{ms}^{\text{zero mean}}(1) + O(T^{-1}) \text{ and} \\ \sqrt{NT} * g^*(\beta_0,\hat{f}) &= \sqrt{NT} * g(\beta_0,f_0) + o_{ms}(1).\end{aligned}$$

Proof: Equation (3) in the text states that

$$g^*(\beta,\hat{f}) = \frac{1}{NT}\sum_i \{g^i(\beta,\hat{f}) - \frac{1}{2}\frac{g_{ff}^i(\beta,\hat{f})}{L_{ff}^i(\beta,\hat{f})} + \frac{1}{2}\frac{L_{fff}^i(\beta,\hat{f})g_f^i(\beta,\hat{f})}{\{L_{ff}^i(\beta,\hat{f})\}^2}\}.$$

The consistency result of Proposition 1 together with continuity of L_f implies that $\hat{f}$ is in a neighborhood $\mathfrak{N}$ of f_0. This implies, by assumption 8, that $L^i_{ff}(\beta, \hat{f}) < 0$ so that $g^*(\beta, \hat{f}) = g(\beta, \hat{f}) + O_{ms}(T^{-1})$. Evaluating $g^*(\beta, \hat{f})$ at β_0 gives,

$$\sqrt{NT} * g^*(\beta_0, \hat{f}) = \frac{1}{\sqrt{NT}} \sum_i [g^i(\beta_0, \hat{f}) - \frac{1}{2} \frac{g^i_{ff}(\beta_0, \hat{f})}{L^i_{ff}(\beta_0, \hat{f})} + \frac{1}{2} \frac{L^i_{fff}(\beta_0, \hat{f}) g^i_f(\beta_0, \hat{f})}{\{L^i_{ff}(\beta_0, \hat{f})\}^2}].$$

A Taylor approximation about f_0 and omitting the argument β_0 gives

$$\begin{aligned}
g^{i,*} &= g^i + g^i_f(\hat{f} - f_0) + \frac{1}{2} g^i_{ff}(\hat{f} - f_0)^2 - \frac{1}{2} \frac{g^i_{ff}}{L^i_{ff}} + \frac{1}{2} \frac{L^i_{fff} g^i_f}{\{L^i_{ff}\}^2} + O(T^{-1}) + O^{\text{zero mean}}_{ms}(T^{-1/2}) \\
&= g^i - \frac{L^i_f g^i_f}{L^i_{ff}} - \frac{1}{2} \frac{L^i_{fff} g^i_f}{L^i_{ff}} (\frac{L^i_f}{L^i_{ff}})^2 + \frac{1}{2} (\frac{L^i_f}{L^i_{ff}})^2 g^i_{ff} - \frac{1}{2} \frac{g^i_{ff}}{L^i_{ff}} \\
&\quad + \frac{1}{2} \frac{L^i_{fff} g^i_f}{\{L^i_{ff}\}^2} + O(T^{-1}) + O^{\text{zero mean}}_{ms}(T^{-1/2}) \\
&= g^i - \{\frac{g^i_f L^i_f + g^i_{ff}}{L^i_{ff}}\} + \frac{1}{2} \{\frac{L^i_{ff} + (L^i_f)^2}{L^i_{ff}}\} \frac{g^i_{ff}}{L^i_{ff}} \\
&\quad + \frac{1}{2} \{\frac{L^i_{ff} - (L^i_f)^2}{L^i_{ff}}\} \{\frac{L^i_{fff} g^i_f}{(L^i_{ff})^2}\} + O(T^{-1}) + O^{\text{zero mean}}_{ms}(T^{-1/2}).
\end{aligned}$$

Note that

$$\begin{aligned}
E[\{\frac{L^i_{ff} + (L^i_f)^2}{L^i_{ff}}\} \frac{E g^i_{ff}}{L^i_{ff}}] &= E[\{\frac{L^i_{ff} + (L^i_f)^2}{L^i_{ff}}\} \frac{E g^i_{ff}}{E L^i_{ff}}] - E[\{\frac{L^i_{ff} + (L^i_f)^2}{L^i_{ff}}\} \frac{E g^i_{ff}}{E L^i_{ff}} (\frac{L_{ff} - E L^i_{ff}}{E L^i_{ff}})] \\
&= E[\{\frac{L^i_{ff} + (L^i_f)^2}{E L^i_{ff}}\} \frac{E g^i_{ff}}{E L^i_{ff}}] + O(T^{-2}),
\end{aligned}$$

since $E[\{\frac{L^i_{ff} + (L^i_f)^2}{E L^i_{ff}}\} \frac{E g^i_{ff}}{E L^i_{ff}} (\frac{L_{ff} - E L^i_{ff}}{E L^i_{ff}})]$ is $O(T^{-2})$ by Lemma 1.

$$g^{i,*} = g^i - \{\frac{g^i_f L^i_f + g^i_{ff}}{L^i_{ff}}\} + \frac{1}{2} \{\frac{L^i_{ff} - (L^i_f)^2}{L^i_{ff}}\} \{\frac{L^i_{fff} g^i_f}{(L^i_{ff})^2}\} + O(T^{-1}) + O^{\text{zero mean}}_{ms}(T^{-1/2}).$$

The expectation of the second and third term are $O(T^{-1})$ by lemma 2 and 3. Moreover, $g^i_{ff} - E g^i_{ff}$ is $O_{ms}(\sqrt{T})$ and $\frac{L^i_f(\beta_0) g^i_f(\beta_0) + E g^i_{ff}}{E L^i_{ff}(\beta_0)}$ is $O^{\text{zero mean}}_{ms}(1)$. This yields

$$\begin{aligned}
E g^{i,*}(\beta_0) &= O(T^{-1}) \text{ and} \\
g^{i,*}(\beta_0) &= g^i(\beta_0) + O^{\text{zero mean}}_{ms}(1) + O(T^{-1}).
\end{aligned}$$

Using the conditional independence assumption implies

$$\sqrt{NT} * g^*(\beta_0, \hat{f}) = \frac{\sum_i g^{i,*}(\beta_0)}{\sqrt{NT}} = \sqrt{NT} * g(\beta_0) + O_{ms}(T^{-1/2}) + O(\sqrt{\frac{N}{T^3}}).$$

Assumption 6, $T \propto N^\alpha$ where $\alpha > \frac{1}{3}$, implies that $O(\sqrt{\frac{N}{T^3}})$ is $o(1)$. Q.E.D.

As noted in the text, theorem 1 remains true if we somewhat weaken assumption 8.

Lemma 4*: *Let the assumptions of theorem 1 hold but assumption 8 (iii) only has to hold with probability one and* $E[||\frac{1}{T}L^i_{ffff}(z, \beta_0, f_0)||^2] < \infty$ *is replaced by* $E[||\frac{1}{T}L^i_{ffff}(z, \beta_0, f_0)||] < \infty$ *in assumption 8 (iv). Then*

$$\begin{aligned} g^*(\beta, \hat{f}) &= g(\beta, \hat{f}) + O_p(T^{-1}), \text{ for } \beta \text{ in a neighborhood } \mathfrak{N} \text{ of } \beta_0, \\ g^{i,*}(\beta_0) &= g^i(\beta_0, f_{i,0}) + \eta_i + O_p(T^{-1}) \text{ and} \\ \sqrt{NT} * g^*(\beta_0, \hat{f}) &= \sqrt{NT} * g(\beta_0, f_0) + o_p(1) \end{aligned}$$

∞ where $\eta_i \sim N(0, \sigma_i^2)$ and $\sigma_i^2 < \infty$.

Proof: Evaluating $g^*(\beta, \hat{f})$ at β_0 gives

$$\sqrt{NT} * g^*(\beta_0, \hat{f}) = \frac{1}{\sqrt{NT}} \sum_i [g^i(\beta_0, \hat{f}) - \frac{1}{2}\frac{g^i_{ff}(\beta_0, \hat{f})}{L^i_{ff}(\beta_0, \hat{f})} + \frac{1}{2}\frac{L^i_{fff}(\beta_0, \hat{f}) g^i_f(\beta_0, \hat{f})}{\{L^i_{ff}(\beta_0, \hat{f})\}^2}].$$

A Taylor approximation about f_0 and omitting the argument β_0 gives

$$\begin{aligned} g^{i,*} &= g^i + g^i_f(\hat{f} - f_0) + \frac{1}{2}g^i_{ff}(\hat{f} - f_0)^2 - \frac{1}{2}\frac{g^i_{ff}}{L^i_{ff}} + \frac{1}{2}\frac{L^i_{fff} g^i_f}{\{L^i_{ff}\}^2} + \frac{\eta_i}{T^{1/2}} + O_p(T^{-1}) \\ &= g^i - \{\frac{g^i_f L^i_f + g^i_{ff}}{L^i_{ff}}\} + \frac{1}{2}\{\frac{L^i_{ff} + (L^i_f)^2}{L^i_{ff}}\}\frac{g^i_{ff}}{L^i_{ff}} \\ &\quad + \frac{1}{2}\{\frac{L^i_{ff} - (L^i_f)^2}{L^i_{ff}}\}\{\frac{L^i_{fff} g^i_f}{(L^i_{ff})^2}\} + \frac{\eta_i}{T^{1/2}} + O_p(T^{-1}), \end{aligned}$$

where η_i is normally distributed but differs from equation to equation. Note that $E[\{\frac{L^i_{ff} + (L^i_f)^2}{EL^i_{ff}}\}\frac{Eg^i_{ff}}{EL^i_{ff}}(\frac{L_{ff} - EL^i_{ff}}{EL^i_{ff}})]$ is $O(T^{-2})$ by Lemma 1.

$$g^{i,*} = g^i - \{\frac{g^i_f L^i_f + g^i_{ff}}{EL^i_{ff}}\} + \frac{1}{2}\{\frac{L^i_{ff} - (L^i_f)^2}{EL^i_{ff}}\}\{\frac{L^i_{fff} g^i_f}{(EL^i_{ff})^2}\} + \frac{\eta_i}{T^{1/2}} + O_p(T^{-1}).$$

The expectation of the second and third term are zero and $O(T^{-1})$ by assumption 5 and lemma 3. Moreover, $g^i_{ff} - Eg^i_{ff}$ is $O_{ms}(\sqrt{T})$ and $\frac{L^i_f(\beta_0)g^i_f(\beta_0)+Eg^i_{ff}}{EL^i_{ff}(\beta_0)}$ is $\eta_i + O_{ms}(T^{-1/2})$. This yields

$$\begin{aligned} Eg^{i,*}(\beta_0) &= O(T^{-1}) \text{ and} \\ g^{i,*}(\beta_0) &= g^i(\beta_0) + \eta_i + O_p(T^{-1}). \end{aligned}$$

Using the conditional independence assumption implies

$$\sqrt{NT} * g^*(\beta_0, \hat{f}) = \frac{\sum_i g^{i,*}(\beta_0)}{\sqrt{NT}} = \sqrt{NT} * g(\beta_0) + O_p(T^{-1/2}) + O_p(\sqrt{\frac{N}{T^3}}).$$

Assumption 6, $T \propto N^\alpha$ where $\alpha > \frac{1}{3}$, implies that $O_p(\sqrt{\frac{N}{T^3}})$ is $o_p(1)$.

Lemma 5: *Let the assumptions of theorem 1 hold. Then*

$$g^*_\beta(\beta_0) = g_\beta(\beta_0, f_0) + O_{ms}(T^{-1/2}).$$

Proof: As mentioned in the text, the estimate for the incidental parameter, $\hat{f}$, depends on β. Applying the chain rule when differentiating $g(\beta, \hat{f}(\beta))$ with respect to β gives

$$g^*_\beta(\beta_0) = \frac{\partial g(\beta, \hat{f}(\beta))}{\partial \beta}|_{\beta=\beta_0} = g_\beta(\beta_0, f_0) + \frac{1}{NT}\sum_i g^{i,*}_f(\beta_0)\frac{\partial \hat{f}(\beta)}{\partial \beta}|_{\beta=\beta_0}.$$

Above we show that $\frac{\partial \hat{f}(\beta)}{\partial \beta}|_{\beta=\beta_0} = -\frac{L^i_{\beta f}}{L^i_{ff}} + O_{ms}(T^{-1/2})$. Note that

$$g^{i,*}_f(\beta_0) = g^i_f(\beta_0, \hat{f}) = g^i_f(\beta_0, f_0) + (\hat{f} - f)g^i_{ff}(\beta_0, \bar{f})$$

where $\bar{f} \in [f_0, \hat{f}]$ is an intermediate value. Reasoning similar[15] to lemma 2 and 3 gives that $\frac{1}{NT}\sum_i g^{i,*}_f(\beta_0)\frac{\partial \hat{f}(\beta)}{\partial \beta}|_{\beta=\beta_0}$ is $O_{ms}(T^{-1/2})$ and the lemma follows.

Lemma 6: *Let the assumptions of theorem 1 hold. Then*

$$\begin{aligned} G(\beta_0, f_0) &= Eg_\beta(\beta_0, f_0) \\ &= \frac{\sum_i}{NT}[EL^i_{\beta\beta}(\beta_0, f_0) + \frac{EL^i_{\beta f}(\beta_0, f_0)\text{`}E\text{'}L^i_{\beta f}(\beta_0, f_0)'}{\text{`}E\text{'}L^i_{ff}(\beta_0, f_0)}. \end{aligned}$$

Proof: Differentiating $g(\beta, f)$ with respect to β gives

$$\begin{aligned} g_\beta(\beta,f) &= \frac{\sum_i}{NT} g^i_\beta(\beta,f) = \frac{\sum_i}{NT}[L^i_{\beta\beta}(\beta,f) + L^i_{\beta f}(\beta,f)\frac{\int L^i_{\beta f}(\beta,f)e^{L^i(\beta,f)}dt}{\int L^i_{ff}(\beta,f)e^{L^i(\beta,f)}dt} \\ &+L^i_f(\beta,f)\frac{\int\{L^i_{\beta\beta f}(\beta,f) + L^i_{\beta f}(\beta,f)L^i_\beta(\beta,f)\}e^{L^i(\beta,f)}dt}{\int L^i_{ff}(\beta,f)e^{L^i(\beta,f)}dt} \\ &-L^i_f(\beta,f)\frac{\int L^i_{\beta f}(\beta,f)e^{L^i(\beta,f)}dt\int\{L^i_{\beta ff}(\beta,f) + L^i_{ff}(\beta,f)L^i_\beta(\beta,f)\}e^{L^i(\beta,f)}dt}{(\int L^i_{ff}(\beta,f)e^{L^i(\beta,f)}dt)^2}]. \end{aligned}$$

The only stochastic term in the last two lines is $L^i_f(\beta,f)$ and $EL^i_f(\beta_0,f_0)=0$. The lemma follows. Note that continuity of $\frac{dg^*(\beta,\hat f(\beta))}{d\beta}$ (with probability approaching one) and $\sup_{\beta\in\Re}||\frac{dg^*(\beta,\hat f(\beta))}{d\beta}|| < \infty$ implies that $g^*_\beta(\bar\beta,\hat f(\bar\beta)) \to_p g_\beta(\beta_0,f_0)$ for $\bar\beta\to_p\beta_0$.

Full Rank: Differentiating the moment $\begin{pmatrix} g(\beta,f) \\ \frac{1}{T}L_f(\beta,f)\end{pmatrix}$ with respect to β and f and taking expectation gives

$$\begin{pmatrix} G & \frac{1}{T}EL_{\beta f} \\ 0 & \frac{1}{T}EL_{ff}\end{pmatrix}$$

since $Eg_f(\beta_0,f_0)=0$. Note that $\frac{1}{T}EL_{ff}$ is a diagonal matrix with $\frac{1}{T}EL^i_{ff}$, $i=1,...,N$ as its elements. Thus, the assumption 8 and 9 (that G has full rank) implies that

$$\begin{pmatrix} G & \frac{1}{T}EL_{\beta f} \\ 0 & \frac{1}{T}EL_{ff}\end{pmatrix}$$

has full rank.

The Cramér-Rao bound:

Stuart et al. (1991, section 17.13-17.17, 17.24-17.28, and 18.15-18.16) give a clear exposition of the Cramér-Rao bound. We briefly review the derivation of the Cramér-Rao bound and allow for an asymptotic in which N and/or T increases. Let the estimator $\hat\theta$ be a function of the data and have bias $b(\theta)$. That is[16],

$$\int ...\int \hat\theta e^{L_1+...+L_N}dy_1...dy_N = \theta + b(\theta).$$

Differentiating gives

$$\begin{aligned} \sum_i\int \hat\theta L^i_\theta e^{L^i}dy_i &= I + b'(\theta) \\ \sum_i\int(\hat\theta-\theta)L^i_\theta e^{L^i}dy_i &= I + b(\theta) + b'(\theta) \end{aligned}$$

where $b'(\theta)$ denotes the derivative of $b(\theta)$ with respect to θ. The Cauchy-Schwartz inequality gives

$$\sum_i \{\int (\hat{\theta}-\theta)^2 e^L dy + \int L_\theta^i L_\theta^{i\prime} e^L dy\} \leq I + b'(\theta).$$

Thus

$$E\{(\hat{\theta}-\theta)^2\} \leq \{\sum_i E(L_\theta^i L_\theta^{i\prime})\}^{-1} + \{\sum_i E(L_\theta^i L_\theta^{i\prime})\}^{-1}\{b(\theta)+b'(\theta)\}.$$

Note that $\{\sum_i E(L_\theta^i L_\theta^{i\prime})\}^{-1} \to 0$ if either N or T increases. Let $\{\sum_i E(L_\theta^i L_\theta^{i\prime})\}^{-1}\{b(\theta)+b'(\theta)\} \to 0$ if both N and T increase. Thus, the bound on the variance for regular, asymptotically unbiased estimators is $\{\sum_i E(L_\theta^i L_\theta^{i\prime})\}^{-1}$ in this asymptotic. We can allow for exogenous regressors x by repeating the arguments above while conditioning on x and then taking expectation over x. Thus, the information bound for the incidental parameter models is given by

$$\begin{pmatrix} EL_{\beta\beta} & EL_{\beta f} \\ EL_{f\beta} & EL_{ff} \end{pmatrix}^{-1}.$$

Using the fact that EL_{ff} is a diagonal matrix gives

$$\begin{pmatrix} EL_{\beta\beta} & EL_{\beta f} \\ EL_{f\beta} & EL_{ff} \end{pmatrix}^{-1} = \begin{pmatrix} \{EL_{\beta\beta} - EL_{\beta f}(EL_{ff})^{-1}EL_{f\beta}\}^{-1} & -(EL_{\beta\beta})^{-1}(EL_{\beta f})F \\ -F(EL_{f\beta})(EL_{\beta\beta})^{-1} & F \end{pmatrix}^{-1},$$

where $F = \{EL_{ff} - EL_{f\beta}(EL_{\beta\beta})^{-1}EL_{\beta f}\}^{-1}$.

Following LeCam (1953), Van der Vaart (1998, chapter 8) discusses how the Cramér-Rao bound gives a lower bound on the variance for regular estimators.

Theorem 1

Let assumption 1-3 and 5-8 hold and $\hat{\beta} = \arg\min_{\beta,f\in\Theta}\{Q(\beta,f)\}$. *Then*

$$\sqrt{NT}(\hat{\beta}-\beta_0) \to_d N(0,\Psi)$$

where

$$\Psi = G^{-1}E\{NT * g(\beta_0,f_0)g(\beta_0,f_0)'\}G^{-1}. \tag{10}$$

Under the additional assumption that $EL^i_{f_i}L^i_\beta = -\text{`}E\text{'}L^i_{\beta f_i}$, $EL^i_\beta L^i_\beta = -\text{`}E\text{'}L^i_{\beta\beta}$, and $EL^2_{f_i} = -EL_{f_i f_i}$ *for all* i, *the integrated moment estimator reaches the Cramér-Rao efficiency bound and*

$$\Psi = -[\frac{1}{NT}E(L_{\beta\beta}) - \frac{1}{NT}E(L_{\beta f})\{E(L_{ff})\}^{-1}E(L_{f\beta})]^{-1}. \tag{11}$$

Proof: The consistency result of proposition 1 implies that we only need to consider parameter values close to the true values. Assumption 8 (i) is identical to the assumption of theorem 3.4, (i) of Newey and McFadden (1994). Assumption 1, combined with assumption 9, (i)-(iv), is identical to the assumption of theorem 3.4, (ii)-(v), of Newey and McFadden (1994)[17]. This implies that the score $\frac{g^i}{\sqrt{T}}$ is normally distributed so that $\frac{g^{i,*}}{\sqrt{T}}$ is normally distributed as well, since $\frac{g^{i,*}}{\sqrt{T}} = \frac{g^i}{\sqrt{T}} + O_{ms}^{\text{zero mean}}(T^{-1/2}) + O(T^{-3/2})$ by Lemma 4. Assumption 6 (conditional independence) implies that $\sqrt{NT}g(\beta_0, f_0) = \sqrt{NT}\frac{\sum_i g^i(\beta_0,f_0)}{NT} = \frac{1}{\sqrt{N}}\sum_i \frac{g^i(\beta_0,f_0)}{\sqrt{T}}$ is normally distributed with variance $\frac{1}{NT}\sum_i E(g^i g^{i\prime}) = NT * Egg'$. The rest of the proof follows Newey and McFadden (theorem 3.4). In particular, the first order condition states that

$$g^*(\hat{\beta}, \hat{f}(\hat{\beta})) = 0.$$

Applying the delta method[18] gives

$$(\hat{\beta} - \beta_0) = \{g^*_\beta(\bar{\beta}, , \hat{f}(\bar{\beta}))\}^{-1} g^*(\beta_0, \hat{f})$$

where $\bar{\beta} \in [\hat{\beta}, \beta_0]$ is an intermediate value. By lemma 3 and 4 we have

$$(\hat{\beta} - \beta_0) = \{g_\beta(\bar{\beta})\}^{-1} g(\beta_0) + o_{ms}(\frac{1}{\sqrt{NT}}).$$

In particular, applying the Slutsky theorem gives

$$(\hat{\beta} - \beta_0) = \{G\}^{-1} g + o_{ms}(\frac{1}{\sqrt{NT}}),$$

where we omitted the argument when evaluating a function at $\{\beta_0, f_0\}$. Thus, the asymptotic variance of $\sqrt{NT}(\hat{\beta} - \beta_0)$ equals $\Psi = G^{-1}E\{NT * gg'\}G^{-1}$ and the result of equation (10) follows.

Under the additional assumption that $EL^i_{f_i}L^i_\beta = -\text{`}E\text{'}L^i_{\beta f_i}$, $EL^i_\beta L^i_\beta = -\text{`}E\text{'}L^i_{\beta\beta}$, and $EL^2_{f_i} = -EL_{f_if_i}$ for all i we have

$$\begin{aligned} E\{g^i g^{i\prime}\} &= E\{(L^i_\beta)(L^i_\beta)'\} - 2E\{(L^i_\beta)(L^i_f)\}\{\frac{\text{`}E\text{'}L^i_{\beta f}}{\text{`}E\text{'}L^i_{ff}}\}' \\ &\quad + E\{(L^i_f)(L^i_f)\}\{\frac{\text{`}E\text{'}L^i_{\beta f}}{\text{`}E\text{'}L^i_{ff}}\}\{\frac{\text{`}E\text{'}L^i_{\beta f}}{\text{`}E\text{'}L^i_{ff}}\}', \end{aligned} \tag{12}$$

where $\frac{\text{`}E\text{'}L^i_{\beta f}(\beta,f)}{\text{`}E\text{'}L^i_{ff}(\beta,f)} = \frac{\int L^i_{\beta f}(\beta,f)e^{L^i(\beta,f)}dt}{\int L^i_{ff}(\beta,f)e^{L^i(\beta,f)}dt}$. This gives

$$\Psi = -[\frac{1}{NT}E(L_{\beta\beta}) - \frac{1}{NT}E(L_{\beta f})\{E(L_{ff})\}^{-1}E(L_{f\beta})]^{-1} = -G^{-1},$$

using the fact that $E(L_{ff})$ is diagonal.

Q.E.D.

Relaxing Assumption 8:

Assumption 8 (iii) can be replaced by the following: "(iii) for all i, let $\frac{1}{T}L^i_{ff}(z,\beta,f) < 0$ in a neighborhood N of $\{\beta_0, f_0\}$ with probability larger than $1 - e^{-cT}$ for some $c > 0$." To see why assumption 8 (iii) can be relaxed let the distribution of $\sqrt{NT}g(\beta_0, f_0)$ be denoted by $p(\sqrt{NT}g(\beta_0, f_0))$. Note that

$$\begin{aligned} p(\sqrt{NT}g(\beta_0,f_0)) &= p(\sqrt{NT}g(\beta_0,f_0)|\frac{1}{T}L^i_{ff}(z,\beta,f) < 0 \text{ for all } i)p(\frac{1}{T}L^i_{ff}(z,\beta,f) < 0 \text{ for all } i) \\ &+ p(\sqrt{NT}g(\beta_0,f_0)|\frac{1}{T}L^i_{ff}(z,\beta,f) \geq 0 \text{ for some } i)p(\frac{1}{T}L^i_{ff}(z,\beta,f) \geq 0 \text{ for some } i), \end{aligned}$$

where $p(\frac{1}{T}L^i_{ff}(z,\beta,f) \geq 0 \text{ for some } i)$ is smaller than Ne^{-cT} so that the distribution of $p(\sqrt{NT}g(\beta_0,f_0)|\frac{1}{T}L^i_{ff}(z,\beta,f) \geq 0)$ converges to $p(\sqrt{NT}g(\beta_0,f_0))$ by the Slutsky theorem.

Priors and Mixing Distributions

Kass et al. (1990, theorem 4) states that

$$\frac{\int \frac{1}{T}g(\beta,f)e^{L^i(\beta,f)}\pi(\beta,f)df}{\int e^{L^i(\beta,f)}\pi(\beta,f)df} = \frac{1}{T}g^{i,*}(\beta,\hat{f}) + O_p(T^{-2}). \tag{13}$$

Kass et al. (1990) assume that

(1) $\pi(\beta,f)$ is four times differentiable with respect to f for $\{\beta,f\} \in \Theta$ with probability one and $\int e^{L^i}\pi(\beta,f)df < M_1$ for some finite M_1 (see Kass et al. (1990, lemma 2).

(2) $g(\beta, f)$ is four time differentiable with respect to f for $\{\beta, f\} \in \Theta$ with probability one and $|\int \frac{1}{T} g^i(\beta, f) e^{L^i} \pi(\beta, f) df| < M_2$ for some finite M_2 (see Kass et al. (1990, theorem 4); note that the function γ that is used by Kass et al. (1990, theorem 4) equals one in our application).

Kass et al. (1990, lemma 2 and theorem 4) assume that the Laplace approximation is "regular". A Laplace approximation of a function that depends on random variables is "regular" under the following conditions (these conditions are slightly stronger than Kass et al. (1990, theorem 7).

(3a) $L^i(\beta, f)$ is six times continuously differentiable with respect to f for $\{\beta, f\} \in \Theta$ with probability one.

(3b) Let $f_0(\beta)$ be the unique probability limit of the maximum likelihood estimator $\hat{f}$ for given β. Let $B_\varepsilon(\beta)$ be a ball with radius ε around $f_0(\beta)$ and let $f \in B_\varepsilon(\beta)$ where ε can be arbitrarily small. Assume that, $\frac{1}{T} \frac{\partial^j L^i(\beta, f)}{(\partial f)^j}$ is bounded in probability for $j = 1, ..., 6$.

(3c) $\lim_{T \to \infty} \frac{1}{T} L_{ff}(\beta, f) < 0$ for all $f \in B_\varepsilon(\beta)$.

Note that requirement (iv) of Kass et al. (1990, theorem 7) is implied by the assumed uniqueness of $\hat{f}$ in 3b. Finally, we need a condition on β to make the Laplace approximation useful. In particular, we need that

(4) (i) the Laplace approximation holds for all $\beta \in \Theta$ or (ii) that $\beta \to_p \beta_0$ in which case we only need the Laplace approximation to hold for β in the neighborhood of β_0.

Assume that the assumptions of theorem 1 hold and that (1)-(4) hold for all i. Thus, $\frac{\int \frac{1}{T} g^i(\beta, f) e^{L^i(\beta, f)} \pi(\beta, f) df}{\int e^{L^i(\beta, f)} \pi(\beta, f) df} = \frac{1}{T} g^{i,*}(\beta, \hat{f}) + O_p(T^{-2})$. By the assumptions of theorem 1 and the assumption 1 and 2 above, $\frac{\int \frac{1}{T} g^i(\beta, f) e^{L^i(\beta, f)} \pi(\beta, f) df}{\int e^{L^i(\beta, f)} \pi(\beta, f) df}$ and $\frac{1}{T} g^{i,*}(\beta, \hat{f})$ are finite so that the difference is finite as well. This implies that the difference has a finite variance so that $\frac{\int \frac{1}{T} g^i(\beta, f) e^{L^i(\beta, f)} \pi(\beta, f) df}{\int e^{L^i(\beta, f)} \pi(\beta, f) df} = \frac{1}{T} g^{i,*}(\beta, \hat{f}) + O_{ms}(T^{-2})$. Summing over individuals and normalizing gives

$$\sqrt{\frac{T}{N}} \sum_i \frac{\int \frac{1}{T} g^i(\beta, f) e^{L^i(\beta, f)} \pi(\beta, f) df}{\int e^{L^i(\beta, f)} \pi(\beta, f) df} = \sqrt{NT} g^*(\beta, \hat{f}) + O_{ms}(\sqrt{\frac{N}{T^3}}),$$

where $O_{ms}(\sqrt{\frac{N}{T^3}})$ is $o_{ms}(1)$ in the asymptotic of theorem 1 so that the asymptotic distribu-

tion of $\sqrt{\frac{T}{N}}\sum_i \frac{\int \frac{1}{T} g^i(\beta,f) e^{L^i(\beta,f)} \pi(\beta,f) df}{\int e^{L^i(\beta,f)} \pi(\beta,f) df}$ is determined by $\sqrt{NT} g^*(\beta, \hat{f})$ and theorem 1 applies. Replacing $\pi(\beta, f)$ by $\pi(\delta, f)$ does not change the approximation result in equation (13) so that the same arguments yield robustness against mixing distributions.

Integrated Likelihood

To be shown: $EL_{\beta f}(\beta_0, f_0) = 0$, for all $\{\beta_0, f_0\}$ is equivalent to 'E'$L_{\beta f}(\beta, f) = 0$ for all $\{\beta, f\}$.

Proof:

$$\begin{aligned} EL_{\beta f}(\beta_0, f_0) &= 0, \text{ for all } \{\beta_0, f_0\} \Leftrightarrow \\ \int L_{\beta f}(\beta_0, f_0) e^{L(\beta_0, f_0)} dt &= 0 \text{ for all } \{\beta_0, f_0\} \Leftrightarrow \\ \int L_{\beta f}(\beta, f) e^{L(\beta, f)} dt &= 0 \text{ for all } \{\beta, f\} \Leftrightarrow \\ \text{`}E\text{'}L_{\beta f}(\beta, f) &= 0, \text{ for all } \{\beta, f\}. \end{aligned}$$

Note that neither $EL_{\beta f}(\beta_0, f_0) = 0$ nor 'E'$L_{\beta f}(\beta, f) = 0$ depends on the realization of the data and that the proof only depends on relabelling.

Time Dummies

Theorem 2

Let assumption 1-3,5 and 7-11 hold and $\{\hat{\beta}, \hat{f}\} = \arg\min_{\beta, f \in \Theta} \{g(\beta, f)' g(\beta, f) + h_f' h_f\}$. *Let the dimension of the common parameter,* β_c, *be equal to* K *and the dimension of the vector of time dummies,* β_d, *be* D. *Then*

$$\sqrt{NT}(\hat{\beta}_c - \beta_0) \rightarrow {}_d N(0, \Psi_c) \textit{ and} \tag{14}$$

$$\sqrt{N}(\hat{\beta}_d - \beta_0) \rightarrow {}_d N(0, \Psi_d), \tag{15}$$

where $\Psi_c = NT * \Pi_{cc}$ *and* $\Psi_d = N * \Pi_{dd}$. *Under the additional assumption that* $EL^i_{f_i} L^i_\beta = -$'E'$L^i_{\beta f_i}$, $EL^i_\beta L^i_\beta = -$'E'$L^i_{\beta\beta}$, *and* $EL^2_{f_i} = -EL_{f_i f_i}$ *for all* i, *the integrated moment estimator reaches the Cramér-Rao efficiency bound and*

$$\Pi = \begin{pmatrix} \Pi_{cc} & \Pi_{cd} \\ \Pi_{dc} & \Pi_{dd} \end{pmatrix} = -[E(L_{\beta\beta}) - E(L_{\beta f})\{E(L_{ff})\}^{-1} E(L_{f\beta})]^{-1},$$

where , Π_{cc} is K by K, and Π_{dd} is D by D.

$$\begin{aligned}\Psi_c &= NT * \Pi_{cc}, \\ \Psi_d &= N * \Pi_{dd} = -[\frac{1}{N}E(L_{\beta_d\beta_d}) - \frac{1}{N}E(L_{\beta_d f})\{E(L_{ff})\}^{-1}E(L_{f\beta_d})].\end{aligned}$$

Proof: The proof of equation (14) and (15) is nearly identical to the proof of Theorem 1 and therefore omitted. To prove the remainder of the theorem, consider partitioning G. Let

$$\begin{aligned}G_{cc} &= \frac{1}{NT}\{EL_{\beta_c\beta_c} - \frac{EL_{\beta_c f}EL'_{\beta_c f}}{EL_{ff}}\} \\ G_{dd} &= \frac{1}{N}\{EL_{\beta_d\beta_d} - \frac{EL_{\beta_d f}EL'_{\beta_d f}}{EL_{ff}}\} \\ G_{dc} &= \frac{1}{N}\{EL_{\beta_d\beta_c} - \frac{EL_{\beta_c f}EL'_{\beta_d f}}{EL_{ff}}\} \\ G_{cd} &= \frac{1}{T}G'_{dc},\end{aligned}$$

and

$$Egg' = \begin{pmatrix} Eg^c g^{c\prime} & Eg^c g^{d\prime} \\ Eg^d g^{c\prime} & Eg^d g^{d\prime} \end{pmatrix}$$

where $g^c = L_{\beta_c} - L_f\frac{EL_{\beta_c f}}{EL_{ff}}$ and $g^d = L_{\beta_d} - L_f\frac{EL_{\beta_d f}}{EL_{ff}}$. This gives

$$\begin{aligned}Eg^c g^{c\prime} &= \frac{1}{(NT)^2}\{-EL_{\beta_c\beta_c} + \frac{EL_{\beta_c f}EL'_{\beta_c f}}{EL_{ff}}\} = -\frac{1}{NT}Eg^c_{\beta_c} = -\frac{1}{NT}G_{cc} \\ Eg^d g^{d\prime} &= \frac{1}{N^2}\{-EL_{\beta_d\beta_d} + \frac{EL_{\beta_d f}EL'_{\beta_d f}}{EL_{ff}}\} = -\frac{1}{N}Eg^d_{\beta_d} = -\frac{1}{N}G_{dd} \\ Eg^c g^{d\prime} &= \frac{1}{N^2T}E\{(L_{\beta_c} - L_f\frac{EL_{\beta_c f}}{EL_{ff}})(L_{\beta_d} - L_f\frac{EL_{\beta_d f}}{EL_{ff}})\} \\ &= \frac{1}{N^2T}\{-EL_{\beta_c\beta_d} + \frac{EL_{\beta_c f}EL'_{\beta_d f}}{EL_{ff}}\} = -\frac{1}{N}G_{cd} \\ &= (Eg^d g^{c\prime})' = -\frac{1}{NT}G'_{dc}.\end{aligned}$$

Thus

$$Egg' = -\frac{1}{NT}\begin{pmatrix} G_{cc} & T * G_{cd} \\ G_{dc} & T * G_{dd} \end{pmatrix}.$$

Inverting G gives (see, for example, Greene (2002b))

$$G^{-1} = \begin{pmatrix} \{G_{cc} - \frac{1}{T}G'_{dc}G_{dd}G_{dc}\}^{-1} & -G_{cc}^{-1}\frac{1}{T}G'_{dc}\{G_{dd} - \frac{1}{T}G_{dc}G_{cc}G'_{dc}\}^{-1} \\ -\{G_{dd} - \frac{1}{T}G_{dc}G_{cc}G'_{dc}\}^{-1}G_{dc}G_{cc}^{-1} & \{G_{dd} - \frac{1}{T}G_{dc}G_{cc}G'_{dc}\}^{-1} \end{pmatrix},$$

where $G_{cd} = \frac{1}{T} G'_{dc}$. This yields,

$$G^{-1} E g g' = - \begin{pmatrix} \frac{1}{NT} I_K & 0 \\ 0 & \frac{1}{N} I_D \end{pmatrix}$$

where I_K and I_D are identity matrices with dimensions K and D respectively. Thus,

$$G^{-1} E g g' G^{-1} = \begin{pmatrix} \frac{1}{NT}\{G_{cc} - \frac{1}{T} G'_{dc} G_{dd} G_{dc}\}^{-1} & -\frac{1}{NT} G_{cc}^{-1} \frac{1}{T} G'_{dc} \{G_{dd} - \frac{1}{T} G_{dc} G_{cc} G'_{dc}\}^{-1} \\ -\frac{1}{N}\{G_{dd} - \frac{1}{T} G_{dc} G_{cc} G'_{dc}\}^{-1} G_{dc} G_{cc}^{-1} & \frac{1}{N}\{G_{dd} - \frac{1}{T} G_{dc} G_{cc} G'_{dc}\}^{-1} \end{pmatrix}.$$

Above we discussed the Cramér-Rao bound for the integrated moment estimator for the parameter of interest,

$$\{EL_{\beta\beta} - EL_{\beta f}(EL_{ff})^{-1} EL_{f\beta}\}^{-1} = [\sum_i \{EL^i_{\beta\beta} - \frac{EL^i_{\beta f} EL^i_{f\beta}}{EL^i_{ff}}\}]^{-1}.$$

Partitioning β using $\beta = \{\beta_c, \beta_d\}$ gives

$$\begin{aligned} & \{EL_{\beta\beta} - EL_{\beta f}(EL_{ff})^{-1} EL_{f\beta}\}^{-1} \\ = & \begin{pmatrix} \sum_i \{EL^i_{\beta_c\beta_c} - \frac{EL^i_{\beta_c f} EL^i_{f\beta_c}}{EL^i_{ff}}\} & \sum_i \{EL^i_{\beta_c\beta_d} - \frac{EL^i_{\beta_c f} EL^i_{f\beta_d}}{EL^i_{ff}}\} \\ \sum_i \{EL^i_{\beta_d\beta_c} - \frac{EL^i_{\beta_d f} EL^i_{f\beta_c}}{EL^i_{ff}}\} & \sum_i \{EL^i_{\beta_d\beta_d} - \frac{EL^i_{\beta_d f} EL^i_{f\beta_d}}{EL^i_{ff}}\} \end{pmatrix}^{-1} \\ = & \begin{pmatrix} NT * G_{cc} & NT * G_{cd} \\ N * G_{dc} & N * G_{dd} \end{pmatrix}^{-1} \end{aligned}$$

which is the asymptotic variance of the integrated moment estimator. *Q.E.D.*

Predetermined Variables

Reasoning similar to Lemma 2, 3 and 4 gives that, under the conditions of Theorem 3, we have

$$\begin{aligned} & E\{\frac{g^i_f L^i_f + g^i_{ff}}{L^i_{ff}}\} \text{ is } O(T^{-1}) \\ & E[\{\frac{L^i_{ff} - (L^i_f)^2}{L^i_{ff}}\}\{\frac{L^i_{fff} g^i_f}{(L^i_{ff})^2}\}] \text{ is } O(T^{-1}) \\ g^*(\beta, \hat{f}) &= g(\beta, \hat{f}) + O_{ms}(T^{-1}), \text{ for } \beta \text{ in a neighborhood } \mathfrak{N} \text{ of } \beta_0, \\ g^{i,*}(\beta_0) &= g^i(\beta_0, f_{i,0}) + O^{\text{zero mean}}_{ms}(1) + O(T^{-1}) \text{ and} \\ \sqrt{NT} * g^*(\beta_0, \hat{f}) &= \sqrt{NT} * g(\beta_0, f_0) + o_{ms}(1). \end{aligned}$$

If $E(L^i_{\mu_{it}\mu_{it}}|x^t_{i,},\beta_0,f_0)$ ='E'$(L^i_{\mu_{it}\mu_{it}}|x^t_{i,},\beta_0,f_0)$, e.g. under correct specification, then G has a simple form,

$$\begin{aligned} G &= Eg^i_\beta(\beta_0,f_0) = \sum_{t=1}^{T-1}[E(x^2_{it}L_{\mu_{it}\mu_{it}}) - E\{x_{it}x_{i,t+1}\frac{\text{`}E\text{'}(L^i_{\mu_{it}\mu_{it}}|x^t_{i,},\beta_0,f_0)E(L^i_{\mu_{i,t+1}\mu_{i,t+1}}|x^{t+1}_i,\beta_0,f_0)}{\text{`}E\text{'}(L^i_{\mu_{i,t+1}\mu_{i,t+1}}|x^{t+1}_i,\beta_0,f_0)}\}] \\ &= \sum_{t=1}^{T-1}[E(x^2_{it}L_{\mu_{it}\mu_{it}}) - E\{x_{it}x_{i,t+1}\text{`}E\text{'}(L^i_{\mu_{it}\mu_{it}}|x^t_{i,},\beta_0,f_0)\}] \\ &= \sum_{t=1}^{T-1}E\{(x^2_{it}-x_{it}x_{i,t+1})E(L^i_{\mu_{it}\mu_{it}}|x^t_{i,},\beta_0,f_0)\}. \end{aligned}$$

The proof of Theorem 3 is identical to the proof of theorem 1 and, therefore, omitted.

Simulation Results

Consider the panel probit model with exogenous regressors,

$$\Pr(y_{it}=1) = \Phi(x_{it}\beta+f_i).$$

We condition on $0<\sum_t y_{it}<1$,

$$\Pr\{0<\sum_t y_{it}<1\} = 1-\prod_t\Phi(x_{it}\beta+f_i)-\prod_t\bar{\Phi}(x_{it}\beta+f_i).$$

This gives the following conditional log likelihood,

$$\begin{aligned} L &= \sum_t\{y_{it}\ln\Phi(\mu_{it})+(1-y_{it})\ln\bar{\Phi}(\mu_{it})\} \\ &\quad -\ln\{1-\prod_t\Phi(\mu_{it})-\prod_t\bar{\Phi}(\mu_{it})\}, \end{aligned}$$

where $\mu_{it}=x_{it}\beta+f_i$ and $\bar{\Phi}(x_{it}\beta+f_i)=1-\bar{\Phi}(x_{it}\beta+f_i)$.

$$\begin{aligned} L_f &= [\sum_t\{y_{it}\frac{\phi(\mu_{it})}{\Phi(\mu_{it})}-(1-y_{it})\frac{\phi(\mu_{it})}{\bar{\Phi}(\mu_{it})}\} \\ &\quad +\frac{\sum_t\phi(\mu_{it})\{\prod_{s\neq t}\Phi(\mu_{is})-\prod_{s\neq t}\bar{\Phi}(\mu_{is})\}}{1-\prod_t\Phi(\mu_{it})-\prod_t\bar{\Phi}(\mu_{it})}]. \end{aligned}$$

Note that

$$\begin{aligned} L^i_{ff} &= \sum_t [y_{it}\{-\mu_{it}\frac{\phi(\mu_{it})}{\Phi(\mu_{it})} - \frac{\phi(\mu_{it})^2}{\Phi(\mu_{it})^2}\} \\ &\quad +(1-y_{it})\{\mu_{it}\frac{\phi(\mu_{it})}{\bar{\Phi}(\mu_{it})} - \frac{\phi(\mu_{it})^2}{\bar{\Phi}(\mu_{it})^2}\} \\ &\quad -\frac{\mu_{it}\phi(\mu_{it})\{\prod_{s\neq t}\Phi(\mu_{is}) - \prod_{s\neq t}\bar{\Phi}(\mu_{is})\}}{1-\prod_t\Phi(\mu_{it}) - \prod_t\bar{\Phi}(\mu_{it})} \\ &\quad -[\frac{\phi(\mu_{it})\{\prod_{s\neq t}\Phi(\mu_{is}) - \prod_{s\neq t}\bar{\Phi}(\mu_{is})\}}{1-\prod_t\Phi(\mu_{it}) - \prod_t\bar{\Phi}(\mu_{it})}]^2]. \end{aligned}$$

Using 'E'$(y_{it}|\beta, f, x_{it}, 0 < \sum_t y_{it} < 1) = \Phi(\mu_{it})\frac{1-\prod_{s\neq t}\Phi(\mu_{is})}{1-\prod_t\Phi(\mu_{it})-\prod_t\Phi(\mu_{it})}$ gives an expression for 'E'L^i_{ff}.

Average Bias for the fixed effects probit model with exogenous regressors (Heckman (1981b, table 4.1 results).

	$\beta = 1$	$\beta = -0.1$	$\beta = -1$
$\sigma_t^2 = 3$	0.90	−0.10	−0.94
$\sigma_t^2 = 1$	0.91	−0.09	−0.95
$\sigma_t^2 = 0.5$	0.93	−0.10	−0.96

Average Bias for the fixed effects probit model with exogenous regressors as reported by Greene (2002b); simulation design as in Heckman (1981b, table 4.1 results).

	$\beta = 1$	$\beta = -0.1$	$\beta = -1$
$\sigma_t^2 = 3$	1.240	−0.110	−1.224
$\sigma_t^2 = 1$	1.242	−0.1127	−1.200
$\sigma_t^2 = 0.5$	1.225	−0.1230	−1.185

Heckman (1981, table 4.2) reported the following average estimates for β and γ for various values of the parameters.

		$\sigma_t^2 = 3$ $\beta = -0.1$	$\sigma_t^2 = 3$ $\beta = 1$	$\sigma_t^2 = 3$ $\beta = 0$	$\sigma_t^2 = 1$ $\beta = -0.1$	$\sigma_t^2 = 1$ $\beta = 1$	$\sigma_t^2 = 1$ $\beta = 0$
$\gamma = 0.5$	$\widehat{\gamma}$	0.14	0.19	0.03	na^d	0.25	0.17
$\gamma = 0.5$	$\widehat{\beta}$	−0.07	1.21	–	na^d	1.17	–
$\gamma = 0.1$	$\widehat{\gamma}$	−0.34	−0.21	−0.04	−0.28	−0.15	−0.01
$\gamma = 0.1$	$\widehat{\beta}$	−0.06	1.14	–	−0.08	1.12	–

7 References

Alvarez, J. and M. Arellano (1998): "The Time Series and Cross-Section Asymptotics of

Dynamic Panel Data Estimators," Working Paper 9808, CEMFI, Madrid.

Anderson, T. W. and C. Hsiao (1981): "Estimation of Dynamic Models with Error Components," *Journal of the American Statistical Society*, 76, 598-606.

Arellano, M. and S. R. Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58, 277-297.

Arellano, M., and B. E. Honoré, (2001): "Panel Data Models: Some Recent Developments," in *Handbook of Econometrics,* Vol. 5, ed. by J. Heckman and E. Leamer, Amsterdam: North-Holland.

Baltagi, B. H. (1995): *Econometric Analysis of Panel Data*, New York: John Wiley and Sons, New York.

Chamberlain, G. (1984): "Panel Data," in *Handbook of Econometrics,* Vol. 2, ed. by Z. Griliches and M. D. Intriligator. Amsterdam: North-Holland.

——— (1985): "Heterogeneity, Omitted Variable Bias, and Duration Dependence," in *Longitudinal Analysis of Labor Market Data*, ed. by J. J. Heckman and B. Singer. Cambridge: Cambridge University Press.

Cox, D. R., and N. Reid (1987): "Parameter Orthogonality and Approximate Conditional Inference (with Discussion)," *Journal of the Royal Statistical Society*, Series B, 49, 1-39.

——— (1993): "A Note on the Calculation of Adjusted Profile Likelihood," *Journal of the Royal Statistical Society*, Series B, 45, 467-471.

Critchley, F. (1987): "Discussion of Parameter Orthogonality and Approximate Conditional Inference (by Cox, D. R., and N. Reid)," *Journal of the Royal Statistical Society*, Series B, 49, 25-26.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Rubin (1995): *Bayesian Data Analysis.* New York: Chapman and Hall.

Greene, W. H. (2002a): "The Behavior of the Fixed Effects Estimator in Nonlinear Models", NYU Stern School of Business working paper.

Greene, W. H. (2002b): *Econometric Analysis*, fifth edition, Prentice Hall, Upper Saddle

River New Jersey.

Griliches, Z. and J. A. Hausman (1986): "Errors in Variables in Panel Data", *Journal of Econometrics*, 31, 93-118.

Hahn J. and W. Newey (2002): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models", MIT manuscript.

Hansen, L.P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, 50, 1029-1054.

Hausman, J. , B. H. Hall and Z. Griliches (1984): "Econometric Models for Count Data with an Application to the Patents R. and D Relationship," *Econometrica*, 52, 909-938.

Heckman, J. J. (1981a): "Statistical Models for Discrete Panel Data," in *Structural Analysis of Discrete Data with Econometric Applications*, edited by C. F. Manski and D. McFadden, pp. 114-178, MIT Press, Cambridge.

Heckman, J. J. (1981b): "The Incidental Parameter Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Dtata Stochastic Process," in *Structural Analysis of Discrete Data with Econometric Applications*, edited by C. F. Manski and D. McFadden, pp. 179-195, MIT Press, Cambridge.

Heckman, J., H. Ichimura, J. Smith and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017-1098.

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

Hsiao, C. (1986): *"Analysis of Panel Data"*. New York: Cambridge University Press.

Hills, S. E. (1987): "Discussion of Parameter Orthogonality and Approximate Conditional Inference (by Cox, D. R., and N. Reid)," *Journal of the Royal Statistical Society*, Series B, 49, 23-24.

Holtz-Eakin, D., W. Newey, and H. Rosen (1988): "Estimating Vector Autoregressions with Panel Data", *Econometrica*, 56, 1371-1395.

Honoré, B. E. and L. Hu (1999): "Estimation of Censored Regression Models with Endogeneity," unpublished manuscript, Department of Economics, Princeton University.

Honoré, B. E. and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839-874.

Honoré, B. E. and A. Lewbel (2002): "Semiparametric Binary Choice Panel Data Models without Strictly Exogenous Regressors," *Econometrica*, 70, 2053-2063.

Kass, R. E., L. Tierney, and J. B. Kadane (1990): "The Validity of Posterior Expansions Based on Laplace's Method," in *Essays in Honor of George Barnard*, ed. by S. Geiser, S. J. Press, and A. Zellner. Amsterdam: North-Holland.

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

——— (2002): "Orthogonal Parameters and Panel Data, *The Review of Economic Studies,* Vol.69 (3), No.240, 647-666.

LeCam, L. (1953): "On some Asymptotic Properties of Maximum Likelihood Estimates and related Bayes' estimates," *Annals of Mathematical Statistics*, 41, 802.

Mundlak, Y. (1961): "Empirical Production Function Free of Management Bias," *Journal of Farm Economics,* 43, 44-56.

Mundlak, Y. (1978): "On the pooling of time series and cross-section data," *Econometrica*, 46, 69-86.

Nerlove, M. (1971): "Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross Sections," *Econometrica*, 39, 359-382.

Nerlove, M. (2000): "The Future of Panel Data Econometrics," Working Paper, Department of Agriculture and Resource Economics, University of Maryland.

Neyman, J., and E. L. Scott (1948): "Consistent Estimation from Partially Consistent Observations," *Econometrica*, 16, 1-32.

Newey, W. K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99-135.

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics,* Vol. 4, ed. by R. F. Engle and D. MacFadden. Amsterdam: North-Holland.

Newey, W. K., and R. J. Smith (2001): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators" MIT manuscript.

Stuart, A., and J. K. Ord, and S. Arnold (1999): *Kendall's Advanced Theory of Statistics*, Volume 2A. New York: Oxford University Press.

Tierney, L., R. E. Kass and J. B. Kadane (1989): "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Society*, 84, 710-716.

Trognon, A. (2000): "Panel Data Econometrics: A Successful Past and a Promising Future," Working Paper, Genes (INSEE).

Van der Vaart, A. W. (1998): *Asymptotic Statistics,* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, UK.

Wooldridge, J. (2001): *Econometric Analysis of Cross Section and Panel Data,* MIT Press, Cambridge.

Woutersen, T. M. (2000): "Consistent Estimation and Orthogonality," Working paper, Department of Economics, University of Western Ontario.

Notes

[1]For the Poisson model with $Ey_{it} = f_i e^{x_{it}\beta}$, $g(\beta, f)$ gives the conditional maximum likelihood estimator of Hausman, Hall and Griliches (1984); for the linear model, $g(\beta, f)$ produces the difference estimator.

[2]Replacing f by its maximum likelihood estimate for given β implies $L^i_{f_i}(\beta, \hat{f}_i) = 0$ so that $g(\beta, \hat{f}) = L_\beta(\beta, \hat{f})$ which yields the maximum likelihood estimate for β.

[3]Subject to regularity conditions given in the next section, using a prior on f does not change the approximation, i.e.

$$g^{i,*}(\beta) = \frac{\int g(f) e^{L^i} \pi(\beta, f) df}{\int e^{L^i} \pi(\beta, f) df} + O_p(T^{-1}).$$

[4]See section 3.

[5]A dynamic linear model has the following form, $y_{it} = c + x_{it}\beta + \sum_{s=1}^{p} \rho_s y_{i,t-s} + \varepsilon_t$ where one of the coefficients can be individual specific; unlike existing estimators, the integrated moment estimator reaches the Cramér-Rao bound for slowly increasing T.

[6]I thank Jeffrey Wooldridge for his encouragement to include a section about time dummies in the paper.

[7]If all regressors are exogenous, the same estimating functions can be used for an asymptotics in which T increases faster than N by relabelling the time periods as individuals and vice versa; in that case, the common parameter is approximately separated from the time dummies.

[8]However, 'only appearing in the likelihood of a particular period' is not a satisfactory definition since the linear model with a moving average error term has the time dummy of period t appearing in all subsequent periods.

[9]A regressor x is *predetermined* or *weakly exogenous* if

$$P(x_{it}|f_i, y_{i1}, ..., y_{iT}) = P(x_{it}|f_i, y_{i1}, ..., y_{i,t-1}) \text{ for all } t.$$

[10]In the last section, we conditioned on the exogenous x_i^T.

[11]Note that $\frac{\partial \mu_{it}}{\partial f} = 1$.

[12]The Weibull model is a duration model with hazard $\theta(t|f_i, x_{is}) = e^{x_{is}\beta + f_i} \alpha t^{\alpha-1}$; using $g(\alpha, \beta, f)$ gives an estimator that is consistent for fixed T.

[13]This moment function is mentioned by Anderson and Hsiao (1981), Griliches and Hausman (1986), Holtz-Eakin, Newey, and Rosen (1988), and Arellano and Bond (1991) amongst others.

[14]The simulation results are very close to the results obtained by Greene (2002a).

[15]but here we divide by NT

[16]We first consider the case without regressors.

[17]Assumption 1, 8 (i) and 9 (i-iii) are very close to Hansen (1982), assumption 3.1-3.6. Newey and McFadden (1994) is based on Hansen (1982) and the point of Hansen (1982) is to allow for predetermined variables; Newey and McFadden (1994, page 2148) note that "the hypotheses of Theorem 2.6 are only used to make sure that $\hat{\theta} \rightarrow \theta_0$, so that they can be replaced by any other conditions that imply consistency." Proposition 1 implies consistency so all assumptions of theorem 3.4 are satisfied.

[18]see for example Newey and McFadden (1994)

Omitted appendices of Robustness against Incidental Parameters

Appendix A: Simple Models with fixed Effects

1. Poisson Model with exogenous regressors

Hausman, Hall and Griliches (1984) assume that y_{it} has a Poisson distribution with mean $f_i e^{x_{it}\beta}$. The likelihood contribution of individual i has the following form,

$$L^i(\beta, f_i) = -f_i \sum_t e^{x_{it}\beta} + \ln(f_i) \sum_t y_{it} + \sum_t y_{it} x_{it}\beta - \sum_t \ln(y_{it}!).$$

Differentiating gives

$$\begin{aligned}
L^i_\beta(\beta, f_i) &= -f_i \sum_t x_{it} e^{x_{it}\beta} + \sum_t y_{it} x_{it} \\
L^i_{\beta f_i}(\beta, f_i) &= -\sum_t x_{it} e^{x_{it}\beta} \\
L^i_{f_i}(\beta, f_i) &= -\sum_t e^{x_{it}\beta} + \frac{\sum_t y_{it}}{f_i} \\
L^i_{f_i f_i}(\beta, f_i) &= -\frac{\sum_t y_{it}}{f_i^2}
\end{aligned}$$

Thus,

$$\frac{\text{`}E\text{'}L^i_{\beta f}(\beta, f_i)}{\text{`}E\text{'}L^i_{ff}(\beta, f_i)} = \frac{\int L^i_{\beta f}(\beta, f_i) e^{L^i(\beta, f_i)} dt}{\int L^i_{ff}(\beta, f_i) e^{L^i(\beta, f_i)} dt} = \frac{\sum_t x_{it} e^{x_{it}\beta}}{\frac{\sum_t \text{`}E\text{'}y_{it}}{f_i^2}} = f_i \frac{\sum_t x_{it} e^{x_{it}\beta}}{\sum_t e^{x_{it}\beta}}$$

using $\text{`}E\text{'}y_{it} = f_i e^{x_{it}\beta}$. This gives

$$\begin{aligned}
g(\beta, f) &= \frac{\sum_i}{N} \{L^i_\beta(\beta, f_i) - L^i_{f_i}(\beta, f_i) \frac{\text{`}E\text{'}L^i_{\beta f_i}(\beta, f_i)}{\text{`}E\text{'}L^i_{f_i f_i}(\beta, f_i)}\} \\
&= \frac{\sum_i}{N} [f_i \sum_t x_{it} e^{x_{it}\beta} + \sum_t y_{it} x_{it} - \{\sum_t e^{x_{it}\beta} + \frac{\sum_t y_{it}}{f_i}\} f_i \frac{\sum_t x_{it} e^{x_{it}\beta}}{\sum_t e^{x_{it}\beta}}] \\
&= \frac{\sum_i}{N} \{\sum_t y_{it} x_{it} - \frac{\sum_t y_{it} \sum_t x_{it} e^{x_{it}\beta}}{\sum_t e^{x_{it}\beta}}\},
\end{aligned}$$

which is the estimating function first derived by Hausman, Hall and Griliches (1984); see also Lancaster (2002).

2. Weibull Model

The Weibull model is a duration model with hazard

$$\theta(t|f_i,\alpha,\beta,x_{is}) = e^{x_{is}\beta+f_i}\alpha t^{\alpha-1}.$$

First, we assume that x_{is} is exogenous. This gives

$$\begin{aligned}
L^i(\alpha,\beta,f_i) &= Tf_i + T\ln(\alpha) + (\alpha-1)\sum_t \ln(t_{is}) + \sum_t x_{is}\beta - e^{f_i}\sum_s e^{x_{is}\beta}t_{is}^{\alpha} \\
L^i_{\alpha}(\alpha,\beta,f_i) &= \frac{T}{\alpha} + \sum_t \ln(t_{is}) - e^{f_i}\sum_s e^{x_{is}\beta}t_{is}^{\alpha}\ln(t_{is}) \\
L^i_{\beta}(\alpha,\beta,f_i) &= \sum_t x_{is} - e^{f_i}\sum_s x_{is}e^{x_{is}\beta}t_{is}.
\end{aligned}$$

Differentiating with respect to f_i gives

$$\begin{aligned}
L^i_{f_i}(\alpha,\beta,f_i) &= T - e^{f_i}\sum_s e^{x_{is}\beta}t_{is}^{\alpha} \\
L^i_{\alpha f_i}(\alpha,\beta,f_i) &= -e^{f_i}\sum_s e^{x_{is}\beta}t_{is}^{\alpha}\ln(t_{is}) \\
L^i_{\beta f_i}(\alpha,\beta,f_i) &= -e^{f_i}\sum_s x_{is}e^{x_{is}\beta}t_{is}^{\alpha} \\
L^i_{f_if_i}(\alpha,\beta,f_i) &= L^i_{f_if_if_i}(\alpha,\beta,f_i) = -e^{f_i}\sum_s e^{x_{is}\beta}t_{is}^{\alpha}.
\end{aligned}$$

Thus,

$$e^{\hat{f_i}} = \frac{T}{\sum_s e^{x_{is}\beta}t_{is}^{\alpha}}$$

and

$$\begin{aligned}
\text{`}E\text{'}L^i_{\alpha f_i}(\alpha,\beta,f_i) &= -\text{`}E\text{'}\sum_s e^{x_{is}\beta}t_{is}^{\alpha}\ln(t_{is}) = -\frac{T}{\alpha}\{\psi(2) - f_i - \sum_s x_{is}\beta\}, \\
\text{`}E\text{'}L^i_{\beta f_i}(\alpha,\beta,f_i) &= -\sum_s x_{is}.
\end{aligned}$$

Let $g(\alpha,\beta,f_i) = \{g_1(\alpha,\beta,f_i), g_2(\alpha,\beta,f_i)\}$ where $g_1(\alpha,\beta,f_i)$ is a scalar and $g_2(\alpha,\beta,f_i)$ has the same dimension as x_{is}.

$$\begin{aligned}
g_1(\alpha,\beta,f_i) &= \frac{T}{\alpha} + \sum_t \ln(t_{is}) - e^{f_i}\sum_s e^{x_{is}\beta}t_{is}^{\alpha}\ln(t_{is}) \\
&\quad -\{T - e^{f_i}\sum_s e^{x_{is}\beta}t_{is}^{\alpha}\}\frac{1}{\alpha}\{\psi(2) - f_i - \sum_s x_{is}\beta\}
\end{aligned}$$

$$
\begin{aligned}
g_{1,f_i}(\alpha,\beta,f_i) &= -e^{f_i}\sum_s e^{x_{is}\beta}t_{is}^{\alpha}\ln(t_{is}) \\
&\quad +e^{f_i}\sum_s e^{x_{is}\beta}t_{is}^{\alpha}\frac{1}{\alpha}\{\psi(2)-f_i-\sum_s x_{is}\beta\} \\
&\quad +\{T-e^{f_i}\sum_s e^{x_{is}\beta}t_{is}^{\alpha}\}\frac{1}{\alpha},
\end{aligned}
$$

and

$$
\begin{aligned}
g_{1,f_if_i}(\alpha,\beta,f_i) &= -e^{f_i}\sum_s e^{x_{is}\beta}t_{is}^{\alpha}\ln(t_{is}) \\
&\quad +e^{f_i}\sum_s e^{x_{is}\beta}t_{is}^{\alpha}\frac{1}{\alpha}\{\psi(2)-f_i-\sum_s x_{is}\beta\} \\
&\quad -\frac{2}{\alpha}e^{f_i}\sum_s e^{x_{is}\beta}t_{is}^{\alpha}.
\end{aligned}
$$

Note that $g_1(\alpha,\beta,\hat{f}_i)=\frac{T}{\alpha}+\sum_t\ln(t_{is})-T\frac{\sum_s e^{x_{is}\beta}t_{is}^{\alpha}\ln(t_{is})}{\sum_s e^{x_{is}\beta}t_{is}^{\alpha}}$ and $g_{1,f_i}(\alpha,\beta,\hat{f}_i)-g_{1,f_if_i}(\alpha,\beta,\hat{f}_i)=\frac{2}{\alpha T}$. This yields

$$
\begin{aligned}
g_1^*(\alpha,\beta) &= \frac{1}{NT}\sum_i\{g^i(\alpha,\beta,\hat{f}_i)-\frac{1}{2}\frac{g^i_{f_if_i}(\alpha,\beta,\hat{f}_i)}{L^i_{f_if_i}(\alpha,\beta,\hat{f}_i)}+\frac{1}{2}\frac{L^i_{fff}(\alpha,\beta,\hat{f}_i)g^i_{f_i}(\alpha,\beta,\hat{f}_i)}{\{L^i_{f_if_i}(\alpha,\beta,\hat{f}_i)\}^2}\} \\
&= \frac{1}{NT}\sum_i[\frac{T}{\alpha}+\sum_t\ln(t_{is})-T\frac{\sum_s e^{x_{is}\beta}t_{is}^{\alpha}\ln(t_{is})}{\sum_s e^{x_{is}\beta}t_{is}^{\alpha}}-\frac{1}{\alpha T}].
\end{aligned}
$$

Similary,

$$
\begin{aligned}
g_2^i(\alpha,\beta,f_i) &= -e^{f_i}\sum_s \tilde{x}_{it}e^{x_{is}\beta}t_{is}^{\alpha} \\
g^i_{2,f_i}(\alpha,\beta,f_i) &= g_{2,f_if_i}(\beta,f_i)=-e^{f_i}\sum_s \tilde{x}_{it}e^{x_{is}\beta}t_{is}^{\alpha}.
\end{aligned}
$$

where $\tilde{x}_{it}=x_{it}-\frac{\sum_t x_{it}}{T}$. This yields

$$
\begin{aligned}
g_2^*(\alpha,\beta) &= \frac{1}{NT}\sum_i\{g^i(\alpha,\beta,\hat{f}_i)-\frac{1}{2}\frac{g^i_{f_if_i}(\alpha,\beta,\hat{f}_i)}{L^i_{f_if_i}(\alpha,\beta,\hat{f}_i)}+\frac{1}{2}\frac{L^i_{fff}(\alpha,\beta,\hat{f}_i)g^i_{f_i}(\alpha,\beta,\hat{f}_i)}{\{L^i_{f_if_i}(\alpha,\beta,\hat{f}_i)\}^2}\} \\
&= \frac{1}{NT}\sum_i\frac{\sum_s \tilde{x}_{it}e^{x_{is}\beta}t_{is}^{\alpha}}{\sum_s e^{x_{is}\beta}t_{is}^{\alpha}}.
\end{aligned}
$$

Chamberlain (1985) derived the estimating function $g^*(\alpha,\beta)$ by differencing the logarithms of the durations. If the regressors are predetermined (as opposed to exogenous), then estimating

function $g_2^*(\alpha,\beta)$ needs to be changed in accordance with section (3.2); this yields

$$g_2^*(\alpha,\beta) = \frac{1}{NT}\sum_i \frac{\sum_s x_{is}(e^{x_{is}\beta}t_{is}^{\alpha} - e^{x_{i,s+1}\beta}t_{i,s+1}^{\alpha})}{\sum_s e^{x_{is}\beta}t_{is}^{\alpha}}$$

while $g_1^*(\alpha,\beta)$ remains unchanged. Note $g_1^*(\alpha,\beta)$ only depends on the predetermined regressor x_{is} through $e^{x_{is}\beta}t_{is}^{\alpha}$, where $e^{x_{is}\beta}t_{is}^{\alpha}$ has an exponential distribution and is independent of x_{is}. Also note that $Eg_1^*(\alpha,\beta) = Eg_2^*(\alpha,\beta) = 0$ and that the resulting estimator is consistent for fixed T.

3. Linear Model with exogenous regressors

Assume normality of the error term in order to derive a quasi-likelihood function. This gives the following likelihood contribution of individual i,

$$L^i(\beta, f_i) = -\frac{T}{2}\ln(\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_t (y_{it} - f_i - x_{it}\beta)^2.$$

Reasoning similar to the Poisson model yields $L^i_{\beta f}(\beta, f_i) = \frac{1}{\sigma^2}\sum_t x_{it}$ and a well known moment function,

$$\begin{aligned} g(\beta,\sigma^2,f_i) &= \frac{\sum_i}{N}\frac{1}{\sigma^2}\sum_t \{y_{it} - f_i - x_{it}\beta\}(x_{it} - \frac{\sum_t x_{it}}{T}) \\ &= \frac{\sum_i}{N}\frac{1}{\sigma^2}\sum_t \{\tilde{y}_{it} - \tilde{x}_{it}\beta\}\tilde{x}_{it} = g^*(\beta,\sigma^2). \end{aligned}$$

where $\tilde{x}_{it} = x_{it} - \frac{\sum_t x_{it}}{T}$ and $\tilde{y}_{it} = y_{it} - \frac{\sum_t y_{it}}{T}$.