Responses to the incidental parameter problem

Pua, A.A.Y.

Publication date
2016
Document Version
Final published version

Chapter 1

Introduction

1.1 The promise of panel data

In this chapter, I show through a series of examples that panel data offer researchers three broad but sometimes competing advantages – estimating structural or common parameters more precisely, allowing for dynamics and feedback, and control of time-invariant unobserved heterogeneity. I am working within usual panel data context where the cross-sectional units i are independently sampled.

Let $y_i^t = (y_{i1}, \ldots, y_{it})$ and $x_i^t = (x_{i1}, \ldots, x_{it})$ for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. The variable y_{it} is the outcome of interest and x_{it} is a vector of regressors – both of which are observable. We are interested in the conditional distribution of the observables y_i^T given x_i^T, which is indexed by a finite-dimensional parameter θ. Unfortunately, the presence of the unobservable α_i, which is an individual-specific effect capturing time-invariant unobserved heterogeneity potentially correlated with the regressors, obscures our ability to estimate and make inferences about θ. To see this, consider a prototypical panel data model where the previously mentioned elements can be found in the following integral equation, i.e.,

$$f_{y|x}\left(y_i^T|x_i^T;\theta\right) = \int f_{y|x,\alpha}\left(y_i^T|x_i^T,\alpha_i;\theta\right) f_{\alpha|x}\left(\alpha_i|x_i^T\right) d\alpha_i,$$

where $f_{y|x,\alpha}\left(y_i^T|x_i^T,\alpha_i;\theta\right)$ is a conditional model and $f_{\alpha|x}\left(\alpha_i|x_i^T\right)$ is the distribution of time-invariant unobserved heterogeneity.

The integral equation can be modified can allow for x to be strictly exogenous and for y to have dynamics (where y_{i1} plays the role of the initial condition), i.e.,

$$f\left(y_{iT},\ldots,y_{i2}|y_{i1},x_i^T;\theta\right) = \int f_1\left(y_{iT},\ldots,y_{i2}|y_{i1},x_i^T,\alpha_i;\theta\right) f_2\left(\alpha_i|y_{i1},x_i^T\right) d\alpha_i,$$

where the integrand is given by

$$f_1\left(\cdot|y_{i1}, x_i^T, \alpha_i; \theta\right) = g_T\left(y_{iT}|x_i^T, y_i^{T-1}, \alpha_i\right) \times \ldots \times g_2\left(y_{i2}|x_i^T, y_{i1}, \alpha_i\right).$$

The integral equation can also be modified to allow for x to have feedback, i.e.,

$$\begin{aligned} & f\left(y_{iT}, \ldots, y_{i2}, x_{iT}, \ldots, x_{i2}|y_{i1}, x_{i1}; \theta\right) \\ = & \int f_1\left(y_{iT}, \ldots, y_{i2}, x_{iT}, \ldots, x_{i2}|y_{i1}, x_{i1}, \alpha_i; \theta\right) f_2\left(\alpha_i|y_{i1}, x_{i1}\right) d\alpha_i, \end{aligned}$$

where the integrand is given by

$$\begin{aligned} f_1\left(\cdot|y_{i1}, x_{i1}, \alpha_i; \theta\right) = & \; g_T\left(y_{iT}|x_i^T, y_i^{T-1}, \alpha_i\right) h_T\left(x_{iT}|y_i^{T-1}, x_i^{T-1}, \alpha_i\right) \times \ldots \\ & \times g_2\left(y_{i2}|x_i^2, y_{i1}, \alpha_i\right) h_2\left(x_{i2}|y_{i1}, x_{i1}, \alpha_i\right). \end{aligned}$$

It is certainly possible for each of the terms of the above expression to be indexed by some finite-dimensional parameter θ. Furthermore, it is also possible to have a multi-dimensional fixed effect α_i. Note that having the time series dimension provides more degrees of freedom for which to estimate θ but these degrees of freedom may get consumed by considering more and more complex models, even if we retain fully parametric specifications.

A large part of research in panel data econometrics adopts a fully parametric specification for $f_{y|x,\alpha}$ while leaving $f_{\alpha|x}$ unspecified (see the surveys by Chamberlain (1984), Arellano and Honoré (2001), and Arellano and Bonhomme (2011)). Leaving $f_{\alpha|x}$ unspecified is at the core of the fixed-effects approach because one has to account for sources of heterogeneity not always observed by the econometrician. Since there is scarce guidance from economic theory as to the nature of heterogeneity observed units should possess, we start with a widely used notion of heterogeneity – that any differences among observed units are relatively stable over time but are allowed to be correlated with the included regressors. Unfortunately, the presence of individual-specific effects complicates the estimation of common parameters in dynamic nonlinear fixed effects panel data models, as we shall see in the examples in the next section. Alternatively, correlated random effects approaches, where some aspects of the distribution $f_{\alpha|x}$ are specified, can be beneficial as discussed in Example 1.2.6. In practice, either we impose assumptions on the first and second moments of $f_{\alpha|x}$ for linear models or we impose fully parametric assumptions on $f_{\alpha|x}$ for nonlinear models.

The conditional model with $f_{y|x,\alpha}$ fully specified can also be used as a starting point while treating the α_i's as parameters to be estimated. In this case, Neyman and Scott (1948) call θ the structural parameter and α_i the incidental parameter. The distinguishing feature of parametric statistical models with incidental parameters is the presence of a parameter α_i that appears in only a finite number of proba-

bility distributions (in particular, that of ith cross-sectional unit). Neyman and Scott (1948) have shown that the maximum likelihood estimator (MLE) of θ may not be consistent in this case.[1] This unfortunate consequence of using ML have henceforth been referred to as the incidental parameter problem (see Lancaster (2000), Arellano and Honoré (2001), and Arellano and Bonhomme (2011) for surveys of some recent developments).[2]

More formally, this incidental parameter problem arises because the MLE $\widehat{\theta}$ has the following property for fixed T:

$$\begin{aligned}
\widehat{\theta} &= \underset{\theta}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} \log f_{y|x,\alpha}\left(y_i^T | x_i^T, \widehat{\alpha_i}(\theta); \theta\right) && (1.1.1)\\
&\xrightarrow{p} \underset{\theta}{\operatorname{argmax}} \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\log f_{y|x,\alpha}\left(y_i^T | x_i^T, \widehat{\alpha_i}(\theta); \theta\right)\right] \\
&\neq \underset{\theta}{\operatorname{argmax}} \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\log f_{y|x,\alpha}\left(y_i^T | x_i^T, \alpha_i\right)\right] && (1.1.2)
\end{aligned}$$

Note that in (1.1.1), we have substituted an estimator of α_i. Hence, the right hand side of (1.1.1) is called the profile or concentrated likelihood. Plugging in an estimator for a finite-dimensional nuisance parameter usually has an asymptotically negligible effect on the estimator for the parameter of interest. In contrast, when we substitute an estimator $\widehat{\alpha_i}(\theta)$ for α_i in (1.1.1), there is an asymptotically nonnegligible effect. The inconsistency of $\widehat{\theta}$ can be traced to four interrelated reasons: (a) the parameter space grows with n, (b) the finite sample bias of $\widehat{\theta}$ that does not disappear in the limit as seen in (1.1.2), (c) the profile or concentrated likelihood does not correspond to a joint density of the observables, and (d) the profile score, which is the derivative of the profile log-likelihood with respect to θ, is not necessarily an unbiased estimating equation. Since these reasons are interrelated, general purpose solutions (some of which are surveyed from an econometrics perspective by Arellano and Hahn (2007) along with its references and from the statistics perspective by Reid

[1] They also show using the example of estimating a normal mean with variances as incidental parameters that sometimes the MLE can be consistent but is no longer asymptotically efficient. They also propose a bias-adjustment method in the spirit of a profile score adjustment. Finally, they sketch the efficiency losses resulting from the incidental parameter problem.

[2] It would seem that treating α_i's as random variables (or random effects) and treating α_i's as parameters are not different from each other. The former subsumes the usual random effects specification where $f_{\alpha|x} = f_\alpha$. Leaving $f_{\alpha|x}$ unspecified is sometimes called the fixed-effects approach. These two models generate estimators that actually have different distribution theories. Sims (2000) argues that "there is a random effects distribution theory for the fixed effects estimator and vice versa." The measurement error literature has been much more explicit about this distinction with respect to its treatment of the latent variable representing the true value of the measurement. The two models are called structural and functional, respectively. See Moran (1971) for more details. Semiparametric estimation and efficiency theory has also been explicit with respect to the distinction. See Moran (1971), Bickel and Klaassen (1986), Bhanja and Ghosh (1992a; 1992b; 1992c), Bickel, Klaassen, et al. (1993), and Pfanzagl (1993) for more details.

(2013), which contain some of the different likelihoods available in the literature) will tend to focus on directly addressing one of these four reasons.

Because the incidental parameter problem is difficult to handle for many nonlinear panel models, some approaches that weaken the fixed-effects approach have been proposed. Typically, the search for consistent estimators of common parameters depends on a set of auxiliary assumptions. Assumptions include, but are not limited to, correlated random effects strategies where the α_i's are drawn from a known $f_{\alpha|x}$ (a particular approach involving sparsity is explored in Chapter 5), fixed-T or large-T bias corrections that exploit full specification of $f_{y|x,\alpha}$ (some of which are explored further in Chapter 4), and approaches invoking discrete support for $f_{\alpha|x}$ (explored further in a simultaneous equations context in Chapter 3). The next four chapters of this dissertation provide specific theoretical or empirical situations for which these auxiliary assumptions may be appropriate (or inappropriate as will be seen in Chapter 2). Before discussing the rest of the thesis, I first discuss the incidental parameter problem in more detail using seven examples.

1.2 Sketching some of the arguments

In this section, I consider some examples that demonstrate the theoretical and practical relevance of the incidental parameter problem along with some proposed solutions. Example 1.2.1 is the many normal means problem posed in Neyman and Scott (1948) where the parameter of interest is the common variance of the observations. The MLE in this example is inconsistent and model-specific solutions are proposed to remedy this inconsistency.

Example 1.2.2 reconsiders the solutions in Example 1.2.1 when both $n, T \to \infty$. Next, Example 1.2.3 is an illustration of the more general case where the $O\left(T^{-1}\right)$ incidental parameter bias is characterized so that we can pursue a general purpose solution. The model-specific nature of fixed-T solutions is further explored in Examples 1.2.4 and 1.2.5. Sometimes these structural parameters are not of main interest and we want to determine how to recover average marginal effects. Example 1.2.6 contains a discussion of how this can be accomplished in fixed-T and large-T situations. Finally, I consider situations where $f_{\alpha|x}$ has discrete support in Example 1.2.7.

Example 1.2.1. (Neyman and Scott (1948), Waterman (1993), and Hahn and Newey (2004)) Let y_{it} be iid draws from a $N\left(\alpha_{i0}, \sigma_0^2\right)$ distribution for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. The parameter of interest in this classic example is the variance parameter σ_0^2. The model allows for one individual-specific effect and does not contain any time-varying regressors. The log-likelihood for one observation is given by

$$\log f\left(y_{it}; \alpha_i, \sigma^2\right) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{\left(y_{it} - \alpha_i\right)^2}{2\sigma^2}.$$

The MLE satisfies the following first order conditions obtained by taking the derivative of the log-likelihood with respect to σ^2 and α_i:

$$\begin{aligned} \sum_i \sum_t \left[-\frac{1}{2\sigma^2} + \frac{(y_{it} - \alpha_i)^2}{2\sigma^4} \right] &= 0, \\ \sum_t \left(\frac{y_{it} - \alpha_i}{\sigma^2} \right) &= 0. \end{aligned} \tag{1.2.1}$$

Profiling out the α_i's using the second equation above gives

$$\widehat{\alpha_i}\left(\sigma^2\right) = \frac{1}{T}\sum_t y_{it} = \bar{y}_i. \tag{1.2.2}$$

Note that (1.2.2) is written as a function of σ^2 even though σ^2 does not explicitly appear in the expression for this simple setup. In general, however, the profiled α_i is going to depend on the structural parameter. Substituting this into (1.2.1) and solving for σ^2 gives

$$\widehat{\sigma^2} = \frac{1}{nT}\sum_i \sum_t (y_{it} - \bar{y}_i)^2 . \tag{1.2.3}$$

Note that (1.2.2) does not depend on σ_0^2 and both (1.2.2) and (1.2.3) are available in closed form. The normality and independence assumptions imply that

$$\widehat{\alpha_i}\left(\sigma^2\right) = \bar{y}_i \sim N\left(\alpha_{i0}, \sigma_0^2/T\right).$$

Results from normal theory (applied to time series observations for the ith cross sectional unit) allow us to conclude that $\sum_t (y_{it} - \bar{y}_i)^2 \sim \sigma_0^2 \chi^2_{T-1}$ for every i. Since we have independence across i, we can write

$$\sum_i \sum_t (y_{it} - \bar{y}_i)^2 \sim \sigma_0^2 \chi^2_{n(T-1)}.$$

Furthermore, taking the expectation of $\widehat{\sigma^2}$ gives

$$\mathbb{E}\widehat{\sigma^2} = \frac{1}{nT}\mathbb{E}\left[\sigma_0^2 \chi^2_{n(T-1)}\right] = \sigma_0^2\left(1 - \frac{1}{T}\right). \tag{1.2.4}$$

As a consequence, $\widehat{\sigma^2}$ is not an unbiased estimator of σ_0^2 in finite samples.

If we want to determine if this finite sample bias disappears in large samples, we have to think of the dimensions in which sample sizes could grow, i.e., the consistency of $\widehat{\sigma^2}$ will depend on the asymptotic embedding. When $T \to \infty$ and n is fixed, $\widehat{\sigma^2}$ is consistent for σ_0^2. When $n \to \infty$ and T is fixed, however, $\widehat{\sigma^2}$ is inconsistent for σ_0^2 because of (1.2.4). As a result, the finite sample bias does not disappear even if $n \to \infty$. We can correct the finite-sample bias directly by using the bias-corrected

estimator $\widehat{\sigma_c^2} = \frac{T}{T-1}\widehat{\sigma^2}$. The degrees of freedom correction produces an unbiased and consistent estimator in this case. ■

The previous example is practically relevant because it is a restricted version of a static linear panel data model with strictly exogenous covariates. In particular, setting $\beta_0 = 0$ in the model where $y_{it}|x_i^T \overset{iid}{\sim} N\left(\alpha_{i0} + \beta_0 x_{it}, \sigma_0^2\right)$ produces the previous example.

Note that the bias in (1.2.4) arises from the finite T setting. One can argue that we can view this bias as finite sample bias in the time series dimension brought about by our inability to consistently estimate α_i . Letting $T \to \infty$ while fixing n is a solution for panel data typically encountered in financial (and sometimes macroeconomic) situations. In contrast, many existing datasets derived from surveys have a large-n dimension with a relatively small T. Therefore, a slight change in the asymptotic scheme may be fruitful.

Example 1.2.2. (Continuation of Example 1.2.1) Let us return to the earlier example. When both $n, T \to \infty$ at some unspecified rate, $\widehat{\sigma^2}$ will be consistent for σ_0^2. Unfortunately, the limiting distribution of $\widehat{\sigma^2}$ may be incorrectly centered. Consider the limiting distribution of $\sqrt{nT}\left(\widehat{\sigma^2} - \sigma_0^2\right)$. We have

$$\begin{aligned}
\sqrt{nT}\left(\widehat{\sigma^2} - \sigma_0^2\right) &= \sqrt{nT}\left(\frac{1}{nT}\sum_i\sum_t (y_{it} - \bar{y}_i)^2 - \sigma_0^2\right) \\
&= \sqrt{nT}\left(\frac{1}{nT}\sum_i\sum_t (y_{it} - \alpha_{i0} + \alpha_{i0} - \bar{y}_i)^2 - \sigma_0^2\right) \\
&= \underbrace{\sqrt{nT}\left(\frac{1}{nT}\sum_i\sum_t (y_{it} - \alpha_{i0})^2 - \sigma_0^2\right)}_{Z_1} \\
&\quad \underbrace{-\sqrt{nT}\left(\frac{1}{n}\sum_i (\bar{y}_i - \alpha_{i0})^2\right)}_{Z_2}
\end{aligned}$$

where $Z_1 \xrightarrow{d} N\left(0, 2\sigma_0^4\right)$ as $n, T \to \infty$ and

$$Z_2 = \sqrt{\frac{n}{T}}\sigma_0^2\left(\frac{1}{n}\sum_i\left(\frac{\bar{y}_i - \alpha_{i0}}{\sigma_0/\sqrt{T}}\right)^2\right) = \sqrt{\frac{n}{T}}\sigma_0^2\left(\frac{1}{n}\sum_i \chi_1^2\right) \xrightarrow{p} \kappa\sigma_0^2$$

as $n, T \to \infty$ while $n/T \to \kappa^2$ for some finite constant $\kappa > 0$.[3] As a consequence, we have

$$\sqrt{nT}\left(\widehat{\sigma^2} - \sigma_0^2\right) \xrightarrow{d} N\left(-\kappa\sigma_0^2, 2\sigma_0^4\right)$$

[3]The result depends on sequential asymptotics. Here, we have $T \to \infty$ first then $n \to \infty$.

This example shows that the relative growth rates of the two dimensions influence the magnitude of the nonzero center $-\kappa\sigma_0^2$. This nonzero center disappears when $n/T \to 0$. Otherwise, we can remove the nonzero center as follows:

$$\sqrt{nT}\left(\widehat{\sigma^2} - \sigma_0^2\right) + Z_2 = \sqrt{nT}\left(\widehat{\sigma^2} - \sigma_0^2 + \frac{\sigma_0^2}{T}\right) \xrightarrow{d} N\left(0, 2\sigma_0^4\right).$$

By plugging in a consistent estimator for σ_0^2/T under this asymptotic scheme, we are able to bias-correct $\widehat{\sigma^2}$. The bias-corrected estimator $\widehat{\sigma_c^2} = \widehat{\sigma^2} + \widehat{\sigma^2}/T$ will have a limiting distribution that is centered at zero. Interestingly, the asymptotic variance of $\widehat{\sigma_c^2}$ coincides with the asymptotic variance of $\widehat{\sigma^2}$. Finally, note that $\mathbb{E}\widehat{\sigma_c^2} = \sigma_0^2\left(1 - 1/T^2\right)$. As a result, this corrected estimator is different from the degrees of freedom correction considered in Example 1.2.1 because this corrected estimator is biased for fixed T but it no longer has the $O\left(T^{-1}\right)$ bias. ■

The discussion in Examples 1.2.1 and 1.2.2 provides ways in which we can achieve either consistency for fixed T or a correctly centered asymptotic distribution when both $n, T \to \infty$ at rate $n/T \to \kappa^2$. First, we have a closed form solution (1.2.3) for the MLE of the structural parameter and a complete specification of the density of the data. Thus, we can derive the finite-sample distribution of (1.2.3). Second, the bias of the MLE in (1.2.3) also has a closed form and does not depend on α_i (see (1.2.4)). In general, these conditions rarely arise so a general characterization of the nonzero center is needed, as seen in the next example.

Example 1.2.3. (Hahn and Newey (2004), Arellano and Hahn (2007), and Hahn and Kuersteiner (2011)) In the previous example, we have seen an indication that the bias in the estimator for the parameter of interest in a model with incidental parameters is of order $O\left(T^{-1}\right)$. We can think of this bias as time series finite sample bias and consider again the asymptotic setting where both $n, T \to \infty$ and $n/T \to \kappa^2$. This asymptotic setting will allow us to more generally approximate the asymptotic bias in the estimator and then reduce its impact. Assume that $\widehat{\theta}$ is a consistent estimator under this asymptotic setting, i.e. $\lim_{T\to\infty} \theta_T = \theta_0$, where θ_T is the large-n, fixed-T limit of some extremum estimator. Further assume that $\sqrt{nT}\left(\widehat{\theta} - \theta_T\right) \xrightarrow{d} N(0, \Omega)$. Under these assumptions along with a stochastic expansion of θ_T, i.e., $\theta_T = \theta_0 + B/T + O\left(T^{-2}\right)$, we can write

$$\begin{aligned}
\sqrt{nT}\left(\widehat{\theta} - \theta_0\right) &= \sqrt{nT}\left(\widehat{\theta} - \theta_T\right) + \sqrt{nT}\left(\theta_T - \theta_0\right) \\
&= \sqrt{nT}\left(\widehat{\theta} - \theta_T\right) + \sqrt{nT}\frac{B}{T} + \sqrt{nT}O\left(T^{-2}\right) \\
&= \sqrt{nT}\left(\widehat{\theta} - \theta_T\right) + \sqrt{\frac{n}{T}}B + O\left(\sqrt{\frac{n}{T^3}}\right) \\
&\xrightarrow{d} N(B\kappa, \Omega). \qquad (1.2.5)
\end{aligned}$$

Note that (1.2.5) is not centered at 0. In the previous example, we were able to derive that $B = -\sigma_0^2$. To remove the nonzero center in (1.2.5), we need to characterize B and its components of this term because a characterization is essential for the practical purpose of bias reduction and for the theoretical purpose of understanding the sources of incidental parameter bias.

Hahn and Newey (2004) study the case of static panel data models with strictly exogenous regressors. In this example, I highlight the general setting considered by Hahn and Kuersteiner (2011). They show that in panel data models with fully-specified dynamics, the bias term is given by

$$B = -\mathscr{I}^{-1}\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{f_i^{VU^\alpha}}{\mathbb{E}\left(V_{it}^{\alpha_i}\right)} - \frac{1}{2}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{E}\left(U_{it}^{\alpha_i\alpha_i}\right)f_i^{VV}}{\left(\mathbb{E}\left(V_{it}^{\alpha_i}\right)\right)^2}\right),$$

where the components of B involve (a) the information matrix $\mathscr{I}^{-1}$, (b) the cross-covariances of the α_i-score V_{it} and the α_i-derivative of the θ-score U_{it}:[4]

$$f_i^{VU^\alpha} = \sum_{l=-\infty}^{\infty}\mathrm{Cov}\left(V_{it}, U_{i,t-l}^{\alpha_i}\right),$$

(c) the autocovariances of the α_i-score

$$f_i^{VV} = \sum_{l=-\infty}^{\infty}\mathrm{Cov}\left(V_{it}, V_{i,t-l}\right),$$

and (d) the expectation of the second α_i-derivative matrix of the θ-score, denoted by $\mathbb{E}\left(U_{it}^{\alpha_i\alpha_i}\right)$.[5] The other remaining component of B is the α_i-derivative of the α_i-score V_{it}, denoted by $V_{it}^{\alpha_i}$.

The characterization of the nonzero center allows us to develop a bias correction under large-n, large-T asymptotics. Observe that a feasible version of the correction requires us to specify a trimming parameter (called bandwidth) for the infinite sums that form B.

Unfortunately, there are negative results with respect to the point identification of common parameters in fixed-T settings (see Chamberlain (2010)). Honoré and Tamer (2006) show that the common parameters of panel data dynamic discrete choice models are only partially identified. Furthermore, bias correction may fail to provide improvements in fixed-T settings. Given that the MLE is heavily biased without bias correction (as documented by numerous Monte Carlo experiments in the literature), it seems advisable to apply these corrections. In general, it is likely that bias-corrected estimators of the common parameters will be found inside the

[4] In the linear model with strictly exogenous regressors, this cross-covariance is zero. Once dynamics are allowed, this cross-covariance is not necessarily zero.

[5] In the linear model, this expectation is zero regardless of whether the regressors are strictly exogenous or not.

identified set. Although no proof of the previous claim exists, we obtain point identification anyway once T becomes very large. ■

Observe that the examples so far apply to panel data models with strictly exogenous regressors and variables with fully-specified feedback mechanisms. On the other hand, GMM based estimation of linear dynamic panel data methods in the spirit of Arellano and Bond (1991) can in principle allow for regressors whose dynamics are not fully modeled. Unfortunately, these GMM estimators also have an asymptotic distribution with a nonzero center under large-n, large-T asymptotics (see Alvarez and Arellano (2003)). Furthermore, these GMM estimators have been documented to have poor finite sample performance and are susceptible to weak instruments (see Bun and Sarafidis (2015) and its references).

It should not be surprising that there is no uniformly good solution to the incidental parameter problem that would apply to every theoretical or empirical situation. As a result, it helps to look for solutions on a case-by-case basis. One possible approach is to exploit the properties of the chosen parametric family to develop a bias-correction. For instance, in Example 1.2.1, consider transforming the data y_{it} into $y_{it} - \bar{y}_i$. The transformation allows us to eliminate the α_i's because the distribution of the transformed data only depends on σ_0^2. As a result, the likelihood function formed from the transformed data can be used to conduct estimation and inference for σ_0^2. The resulting likelihood is called a marginal likelihood in the statistical literature.[6] Yet, it may be very difficult to find transformations or even subsets of the data that will allow us to construct a marginal likelihood. Despite this, there are successful applications of this idea even outside the likelihood setting as the following example illustrates.

Example 1.2.4. (Honoré, 1992) Consider a linear panel data regression model where $y_{it}^* = \alpha_i + \beta x_{it} + \varepsilon_{it}$ for $i = 1, \ldots, n$ and $t = 1, 2$. For simplicity, assume that x_{it} is scalar. Assume that $\left\{\left(y_{i1}^*, x_{i1}, y_{i2}^*, x_{i2}\right) : i = 1, \ldots, n\right\}$ form a random sample but we only get to observe data on both y and x when $y_{i1}^* > 0$ and $y_{i2}^* > 0$. Further assume that ε_{i1} and ε_{i2} are independent, identically and continuously distributed conditional on $(x_{i1}, x_{i2}, \alpha_i)$ for all i.

Honoré (1992) develops a semiparametric approach in the spirit of a marginal likelihood calculation. The idea is to look for a subset of

$$\left\{\left(y_{i1}^*, y_{i2}^*\right) : y_{i1}^* \in \mathbb{R}, y_{i2}^* \in \mathbb{R}\right\}$$

[6] Some authors call the likelihoods obtained after integrating out the nuisance parameters as marginal likelihoods. See Chamberlain (1980) for an example. To avoid confusion, I will call them integrated likelihoods instead. In contrast, we obtain profile likelihoods by maximizing out the nuisance parameters. These two likelihoods represent different ways of eliminating nuisance parameters (see Basu (1977) and Berger, Liseo, and Wolpert (1999) for more details). The meaning of marginal likelihood I use fits with the notion of marginal inference. See Kalbfleisch and Sprott (1970) and Christensen and Kiefer (2000) for more details. A more recent discussion on the types of likelihood functions can be found in Reid (2013).

that is unaffected by truncation. Observe that such a subset allows us to eliminate α_i by differencing. In other words, we have $y_{i1}^* = y_{i1}$, $y_{i2}^* = y_{i2}$ and both time series observations obey $y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$. Notice that this differencing strategy is exactly the same strategy applied to a linear panel data model (as in Example 1.2.1).

Define $\Delta y_i = y_{i1} - y_{i2}$, $\Delta x_i = x_{i1} - x_{i2}$, and $\Delta \varepsilon_i = \varepsilon_{i1} - \varepsilon_{i2}$. Assume that $\beta \Delta x_i > 0$. Consider the following sets

$$\begin{aligned} A &= \left\{\left(y_{i1}^*, y_{i2}^*\right) : y_{i1}^* > \beta \Delta x_i, y_{i2}^* > y_{i1}^* - \beta \Delta x_i\right\}, \\ B &= \left\{\left(y_{i1}^*, y_{i2}^*\right) : y_{i1}^* > \beta \Delta x_i, 0 < y_{i2}^* < y_{i1}^* - \beta \Delta x_i\right\}. \end{aligned}$$

Notice that whenever $y_{i1}^* > \beta \Delta x_i$, we must have $y_{i2}^* > 0$. Observe that

$$\begin{aligned} & \Pr\left(\left(y_{i1}^*, y_{i2}^*\right) \in A | x_{i1}, x_{i2}, \alpha_i\right) \\ = \; & \Pr\left(y_{i2}^* - y_{i1}^* > -\beta \Delta x_i, y_{i2}^* + y_{i1}^* > \beta \Delta x_i | x_{i1}, x_{i2}, \alpha_i\right) \\ = \; & \Pr\left(\varepsilon_{i2} - \varepsilon_{i1} > 0, \varepsilon_{i2} + \varepsilon_{i1} > -2\alpha_i - 2\beta x_{i2} | x_{i1}, x_{i2}, \alpha_i\right) \\ = \; & \Pr\left(\Delta \varepsilon_i < 0 | x_{i1}, x_{i2}, \alpha_i, \underbrace{\varepsilon_{i2} + \varepsilon_{i1} > -2\alpha_i - 2\beta x_{i2}}_{D_i}\right) \\ & \times \Pr\left(D_i | x_{i1}, x_{i2}, \alpha_i\right). \end{aligned}$$

Similarly, we can write

$$\begin{aligned} & \Pr\left(\left(y_{i1}^*, y_{i2}^*\right) \in B | x_{i1}, x_{i2}, \alpha_i\right) \\ = \; & \Pr\left(y_{i2}^* - y_{i1}^* < -\beta \Delta x_i, y_{i2}^* + y_{i1}^* > \beta \Delta x_i | x_{i1}, x_{i2}, \alpha_i\right) \\ = \; & \Pr\left(\varepsilon_{i2} - \varepsilon_{i1} < 0, \varepsilon_{i2} + \varepsilon_{i1} > -2\alpha_i - 2\beta x_{i2} | x_{i1}, x_{i2}, \alpha_i\right) \\ = \; & \Pr\left(\Delta \varepsilon_i > 0 | x_{i1}, x_{i2}, \alpha_i, D_i\right) \Pr\left(D_i | x_{i1}, x_{i2}, \alpha_i\right) \end{aligned}$$

Under the assumption that the distribution of $\Delta \varepsilon_i$ conditional on $\varepsilon_{i1} + \varepsilon_{i2}$ and on $(x_{i1}, x_{i2}, \alpha_i)$ is symmetric and unimodal around zero,[7] we can then conclude that

$$\Pr\left(\left(y_{i1}^*, y_{i2}^*\right) \in A | x_{i1}, x_{i2}, \alpha_i\right) = \Pr\left(\left(y_{i1}^*, y_{i2}^*\right) \in B | x_{i1}, x_{i2}, \alpha_i\right).$$

Furthermore, these two sets are unaffected by truncation and will be observable (since these sets satisfy $y_{i1}^* > \beta \Delta x_i > 0$ and $y_{i2}^* > 0$). As a result,

$$\Pr\left(\left(y_{i1}, y_{i2}\right) \in A | x_{i1}, x_{i2}\right) = \Pr\left(\left(y_{i1}, y_{i2}\right) \in B | x_{i1}, x_{i2}\right).$$

Therefore, the union of these two sets

$$A \cup B = \left\{\left(y_{i1}, y_{i2}\right) : y_{i1} > \beta \Delta x_i, y_{i2} > 0\right\}$$

[7] See Honoré (1992) for a sufficient condition.

is the basis for constructing a moment condition that only involves the observables but not the fixed effect α_i. Observe further that

$$\begin{aligned}
& \mathbb{E}[\mathbf{1}\{(y_{i1}, y_{i2}) \in A\}\, \Delta\varepsilon_i | x_{i1}, x_{i2}, \alpha_i] \\
= & \int_{\mathbf{1}\{(y_{i1}, y_{i2}) \in A\}} u f_{\Delta\varepsilon|x_1, x_2, \alpha, D}(u)\, du \\
= & \frac{\displaystyle\int_{-\infty}^{0} u f_{\Delta\varepsilon|x_1, x_2, \alpha, D}(u)\, du}{F_{\Delta\varepsilon|x_1, x_2, \alpha, D}(0) \Pr(D_i | x_{i1}, x_{i2}, \alpha_i)} \\
= & \frac{\displaystyle\int_{-\infty}^{0} u f_{\Delta\varepsilon|x_1, x_2, \alpha}(-u)\, du}{\big(1 - F_{\Delta\varepsilon|x_1, x_2, \alpha, D}(0)\big) \Pr(D_i | x_{i1}, x_{i2}, \alpha_i)} \\
= & \frac{\displaystyle -\int_{0}^{\infty} v f_{\Delta\varepsilon|x_1, x_2, \alpha}(v)\, dv}{\big(1 - F_{\Delta\varepsilon|x_1, x_2, \alpha, D}(0)\big) \Pr(D_i | x_{i1}, x_{i2}, \alpha_i)} \\
= & -\mathbb{E}[\mathbf{1}\{(y_{i1}, y_{i2}) \in B\}\, \Delta\varepsilon_i | x_{i1}, x_{i2}, \alpha_i]. \qquad (1.2.6)
\end{aligned}$$

The previous derivation involves the expectation of a truncated random variable and the i.i.d. assumption on the errors. We use the symmetry assumption to obtain the third equality. Using (1.2.6), we are able to show that the moment condition

$$\begin{aligned}
& \mathbb{E}[\mathbf{1}\{(y_{i1}, y_{i2}) \in A \cup B\}\, (\Delta y_i - \beta \Delta x_i)\, \Delta x_i] \\
= & \mathbb{E}[\mathbf{1}\{(y_{i1}, y_{i2}) \in A\}\, \Delta\varepsilon_i \Delta x_i] + \mathbb{E}[\mathbf{1}\{(y_{i1}, y_{i2}) \in B\}\, \Delta\varepsilon_i \Delta x_i] \\
= & \mathbb{E}[\mathbb{E}[\mathbf{1}\{(y_{i1}, y_{i2}) \in A\}\, \Delta\varepsilon_i \Delta x_i | x_{i1}, x_{i2}, \alpha_i]] \\
& + \mathbb{E}[\mathbb{E}[\mathbf{1}\{(y_{i1}, y_{i2}) \in B\}\, \Delta\varepsilon_i \Delta x_i | x_{i1}, x_{i2}, \alpha_i]] \qquad (1.2.7) \\
= & 0
\end{aligned}$$

is satisfied, where $\mathbf{1}(\cdot)$ is the indicator function. The case where $\beta \Delta x_i \leq 0$ is analogous and will be part of the criterion function for estimating β. Notice that without the indicator function in (1.2.7), we have the moment condition for β in the static linear panel data model with strictly exogenous covariates. A least squares objective function can be formed where the resulting first-order condition is exactly the sample analog of (1.2.7).[8] ■

Searching for a suitable subset of the data is what makes marginal likelihood approaches (or any other approach in the same spirit) highly model-specific. Furthermore, assumptions have to be changed in very specific ways to accommodate slight changes in the model. Extensions of the previous example to allow for lagged dependent variables can be found in Honoré (1993) but require a modification of the

[8] See Honoré (1992) for more details.

argument along with the assumptions in the previous example. Abrevaya (1999) proposes an estimator for fixed effects models with an unknown transformation of the dependent variable that also has the flavor of a marginal likelihood approach. Strictly speaking, the estimators discussed here are semiparametric in nature but the common feature is the search for subsets of the data from which to construct moment conditions or likelihoods that do not depend on α_i, but are informative about the structural parameters. Bonhomme (2012) provides a theory that allows any user of a likelihood-based panel data model with strictly exogenous regressors to construct moment conditions that are free of the fixed effects. One can think of the theory as a general treatment of the marginal likelihood approach. Unfortunately, it is possible that certain panel models will not possess moment conditions that are informative of the structural parameters. Aspects of this theory will be discussed further in Example 1.2.7.

Yet another approach is to find appropriate conditioning sets so that a conditional likelihood that does not depend on α_i can be constructed. As a result, the score of the conditional likelihood is itself a moment condition that is free of α_i. The following example illustrates this approach in a dynamic logit model.

Example 1.2.5. (Chamberlain, 1985; Maddala, 1987; Honoré and Kyriazidou, 2000) Consider a dynamic panel logit model with one strictly exogenous regressor. In particular, we have for $i = 1, \ldots, n$ and $t = 1, \ldots, T$:

$$\Pr\left(y_{it} = 1 | x_{i1}, \ldots, x_{iT}, y_{i0}, \ldots, y_{i,T-1}, \alpha_i\right) = \frac{\exp\left(\beta x_{it} + \gamma y_{i,t-1} + \alpha_i\right)}{1 + \exp\left(\beta x_{it} + \gamma y_{i,t-1} + \alpha_i\right)}. \quad (1.2.8)$$

This means that

$$\Pr\left(y_{it} = 0 | x_{i1}, \ldots, x_{iT}, y_{i0}, \ldots, y_{i,T-1}, \alpha_i\right) = \frac{1}{1 + \exp\left(\beta x_{it} + \gamma y_{i,t-1} + \alpha_i\right)}.$$

Assume that y_{i0} is observed and $T = 3$. Hence, we have a total of four observations. Define the sets

$$\begin{aligned} A &= \{y_{i1} = 0, y_{i2} = 1, y_{i3} = d_3\}, \\ B &= \{y_{i1} = 1, y_{i2} = 0, y_{i3} = d_3\}, \end{aligned}$$

where $d_3 \in \{0, 1\}$. Let $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$ and $d_0 \in \{0, 1\}$. We can calculate the following conditional probabilities:

$$\begin{aligned} & \Pr(A | \mathbf{x}_i, y_{i0} = d_0, \alpha_i) \\ = \; & \Pr(y_{i3} = d_3 | \mathbf{x}_i, y_{i0} = d_0, y_{i1} = 0, y_{i2} = 1) \\ & \times \Pr(y_{i2} = 1 | \mathbf{x}_i, y_{i0} = d_0, y_{i1} = 0) \times \Pr(y_{i1} = 0 | \mathbf{x}_i, y_{i0} = d_0) \end{aligned}$$

$$
\begin{aligned}
= & \frac{\exp(d_3(\beta x_{i3}+\gamma+\alpha_i))}{1+\exp(\beta x_{i3}+\gamma+\alpha_i)} \times \frac{\exp(\beta x_{i2}+\alpha_i)}{1+\exp(\beta x_{i2}+\alpha_i)} \\
& \times \frac{1}{1+\exp(\beta x_{i1}+\gamma d_0+\alpha_i)}, \\
& \Pr(B|\mathbf{x}_i, y_{i0}=d_0, \alpha_i) \\
= & \Pr(y_{i3}=d_3|\mathbf{x}_i, y_{i0}=d_0, y_{i1}=1, y_{i2}=0) \\
& \times \Pr(y_{i2}=0|\mathbf{x}_i, y_{i0}=d_0, y_{i1}=1) \times \Pr(y_{i1}=1|\mathbf{x}_i, y_{i0}=d_0) \\
= & \frac{\exp(d_3(\beta x_{i3}+\alpha_i))}{1+\exp(\beta x_{i3}+\alpha_i)} \times \frac{1}{1+\exp(\beta x_{i2}+\gamma+\alpha_i)} \\
& \times \frac{\exp(\beta x_{i1}+\gamma d_0+\alpha_i)}{1+\exp(\beta x_{i1}+\gamma d_0+\alpha_i)}.
\end{aligned}
$$

Choosing $A \cup B$ as a conditioning set and noting that A and B are disjoint sets, the definition of conditional probability allows us to write

$$
\begin{aligned}
& \Pr(A|y_{i0}=d_0, A\cup B, \alpha_i) \\
= & \frac{\Pr(A|\mathbf{x}_i, y_{i0}=d_0, \alpha_i)}{\Pr(A\cup B|\mathbf{x}_i, y_{i0}=d_0, \alpha_i)} \\
= & \frac{\Pr(A|\mathbf{x}_i, y_{i0}=d_0, \alpha_i)}{\Pr(A|\mathbf{x}_i, y_{i0}=d_0, \alpha_i)+\Pr(B|\mathbf{x}_i, y_{i0}=d_0, \alpha_i)}, \qquad (1.2.9) \\
& \Pr(B|y_{i0}=d_0, A\cup B, \alpha_i) \\
= & 1-\Pr(A|\mathbf{x}_i, y_{i0}=d_0, A\cup B, \alpha_i). \qquad (1.2.10)
\end{aligned}
$$

Both probabilities in (1.2.9) and (1.2.10) still depend on α_i.

Consider first the case where $\beta = 0$. Observe that

$$
\begin{aligned}
& \Pr(A\cup B|\mathbf{x}_i, y_{i0}=d_0, \alpha_i) \\
= & \frac{\exp(d_3(\gamma+\alpha_i))}{1+\exp(\gamma+\alpha_i)} \times \frac{\exp(\alpha_i)}{1+\exp(\alpha_i)} \times \frac{1}{1+\exp(\gamma d_0+\alpha_i)} \\
& + \frac{\exp(d_3\alpha_i)}{1+\exp(\alpha_i)} \times \frac{1}{1+\exp(\gamma+\alpha_i)} \times \frac{\exp(\gamma d_0+\alpha_i)}{1+\exp(\gamma d_0+\alpha_i)} \\
= & \frac{\exp(d_3\alpha_i)\exp(\alpha_i)[\exp(d_3\gamma)+\exp(\gamma d_0)]}{[1+\exp(\alpha_i)][1+\exp(\gamma+\alpha_i)][1+\exp(\gamma d_0+\alpha_i)]}.
\end{aligned}
$$

Therefore, we can write (1.2.9) as

$$
\begin{aligned}
& \Pr(A|y_{i0}=d_0, A\cup B, \alpha_i) \\
= & \frac{\dfrac{\exp(d_3(\gamma+\alpha_i))}{1+\exp(\gamma+\alpha_i)} \times \dfrac{\exp(\alpha_i)}{1+\exp(\alpha_i)} \times \dfrac{1}{1+\exp(\gamma d_0+\alpha_i)}}{\dfrac{\exp(d_3\alpha_i)\exp(\alpha_i)[\exp(d_3\gamma)+\exp(d_0\gamma)]}{[1+\exp(\alpha_i)][1+\exp(\gamma+\alpha_i)][1+\exp(\gamma d_0+\alpha_i)]}}
\end{aligned}
$$

$$= \frac{\exp(d_3\gamma)}{\exp(d_3\gamma)+\exp(d_0\gamma)}.$$

Similarly, (1.2.10) can be written as

$$\Pr(B|y_{i0}=d_0, A\cup B, \alpha_i) = \frac{\exp(d_0\gamma)}{\exp(d_3\gamma)+\exp(d_0\gamma)}.$$

Both these conditional probabilities do not depend on α_i and can be used to form a conditional likelihood depending only on γ.

Now, consider the case where $\beta \neq 0$. Honoré and Kyriazidou (2000) show that by further conditioning on the event $\{x_{i2}=x_{i3}\}$, assumed to have positive probability, we can eliminate the dependence of (1.2.9) and (1.2.10) on α_i. In particular, we have

$$\begin{aligned}
\Pr(A|\mathbf{x}_i, y_{i0}=d_0, A\cup B, x_{i2}=x_{i3}, \alpha_i) &= \frac{1}{1+\exp(\beta(x_{i1}-x_{i2})+\gamma(d_0-d_3))}, \\
\Pr(B|\mathbf{x}_i, y_{i0}=d_0, A\cup B, x_{i2}=x_{i3}, \alpha_i) &= \frac{\exp(\beta(x_{i1}-x_{i2})+\gamma(d_0-d_3))}{1+\exp(\beta(x_{i1}-x_{i2})+\gamma(d_0-d_3))}
\end{aligned}$$

and a conditional likelihood will be formed from observations where $x_{i2}=x_{i3}$ and $y_{i1}+y_{i2}=1$, i.e., a conditional MLE can be computed from the following optimization problem:

$$\max_{\beta,\gamma} \sum_{i=1}^{N} \mathbf{1}\{y_{i1}+y_{i2}=1\}\,\mathbf{1}\{x_{i2}=x_{i3}\} \log\left(\frac{[\exp(\beta(x_{i1}-x_{i2})+\gamma(d_0-d_3))]^{y_{i1}}}{1+\exp(\beta(x_{i1}-x_{i2})+\gamma(d_0-d_3))}\right).$$

The condition $x_{i2}=x_{i3}$ is unlikely to be satisfied, so a kernel function replaces the indicator function above. Because we introduce a kernel function, the estimators for the structural parameters converge at a rate slower than the usual parametric rate. We also cannot allow for time dummies because they never satisfy $x_{i2}=x_{i3}$ by definition. Extensions to a semiparametric specification of the probability function (1.2.8) in the spirit of Manski (1987b), the multinomial logit case, and more than four observations for every i are available in Honoré and Kyriazidou (2000). ■

Even though the approaches in Examples 1.2.4 and 1.2.5 are both appealing and insightful, the search for appropriate transformations of the data or appropriate conditioning sets will become cumbersome when T is a bit larger or when we make slight changes to the model.

Some authors like Wooldridge (2005b) and Arellano and Bonhomme (2011) argue that the structural parameters may not be of primary interest especially for policy. Policy parameters are usually of the form $\mathbb{E}\left[m\left(x_i^T, \alpha_i\right)\right]$, where m is some function of the regressors and unobserved heterogeneity. These policy parameters have been called many names depending on the form of m, such as the average structural func-

tion (Blundell and Powell, 2004), quantile structural function (Chernozhukov et al., 2013), average index function (Lewbel, Dong, and Yang, 2012), average marginal effect (Wooldridge, 2005b), and local average response (Altonji and Matzkin, 2005). These policy parameters represent summary measures that describe outcomes of certain thought experiments. One such thought experiment involves a prediction of what m will be when we set x_i^T at some fixed value $\bar{x}$ while holding unobserved heterogeneity constant. Another thought experiment would involve predictions as to how m changes when we change the value $\bar{x}$ while holding unobserved heterogeneity constant. Unfortunately, these policy parameters are hard to identify and require understanding the tradeoffs among competing assumptions as seen in the next example.

Example 1.2.6. (Hoderlein and White (2012)) Consider the following nonseparable model where $Y_{it} = g(X_{it}, \alpha_i, \varepsilon_{it})$ for $i = 1, \ldots, n$ and $t = 1, 2$. An object of interest for policy is how $E(Y_{it}|X_{i1} = x_1, X_{i2} = x_2)$ changes with x_1 or x_2, holding the source of unobserved heterogeneity constant. In other words, the policy parameters of interest or average marginal effects at x_1 and x_2, are given by

$$\begin{aligned} ME_1(x_1, x_2) &= \int\int \frac{\partial g(x_1, a, e)}{\partial x_1} f_{\alpha_i, \varepsilon_{i1}|X_i}(a, e|x_1, x_2)\, da\, de, \\ ME_2(x_1, x_2) &= \int\int \frac{\partial g(x_2, a, e)}{\partial x_2} f_{\alpha_i, \varepsilon_{i2}|X_i}(a, e|x_1, x_2)\, da\, de. \end{aligned}$$

Had we known what $f_{\alpha_i, \varepsilon_{it}|X_i}$ is, then everything becomes straightforward and calculating $ME_1(x)$ and $ME_2(x)$ can be done directly. This situation is really the idea behind the calculation of average marginal effects from fully parametric models with correlated random effects proposed by Chamberlain (1984) and Wooldridge (2005b). If we do not know $f_{\alpha_i, \varepsilon_{it}|X_i}$, we have to indirectly recover $ME_1(x)$ and $ME_2(x)$ somehow. In particular, we have the following

$$\begin{aligned} \mathbb{E}(Y_{i1}|X_{i1} = x_1, X_{i2} = x_2) &= \int\int g(x_1, a, e) f_{\alpha_i, \varepsilon_{i1}|X_i}(a, e|x)\, da\, de, \\ \mathbb{E}(Y_{i2}|X_{i1} = x_1, X_{i2} = x_2) &= \int\int g(x_2, a, e) f_{\alpha_i, \varepsilon_{i2}|X_i}(a, e|x)\, da\, de, \end{aligned}$$

with four derivatives given by

$$\begin{aligned} \frac{\partial \mathbb{E}(Y_{i1}|X_{i1} = x_1, X_{i2} = x_2)}{\partial x_1} &= ME_1(x_1, x_2) \\ &\quad + \int\int g(x_1, a, e) \frac{\partial f_{\alpha_i, \varepsilon_{i1}|X_i}(a, e|x)}{\partial x_1}\, da\, de \\ \frac{\partial \mathbb{E}(Y_{i1}|X_{i1} = x_1, X_{i2} = x_2)}{\partial x_2} &= \int\int g(x_1, \alpha, \varepsilon) \frac{\partial f_{\alpha_i, \varepsilon_{i1}|X_i}(a, e|x)}{\partial x_2}\, da\, de \end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathbb{E}(Y_{i2}|X_{i1}=x_1,X_{i2}=x_2)}{\partial x_1} &= \int\int g(x_2,a,e)\frac{\partial f_{\alpha_i,\varepsilon_{i2}|X_i}(a,e|x)}{\partial x_1}\,da\,de \\
\frac{\partial \mathbb{E}(Y_{i2}|X_{i1}=x_1,X_{i2}=x_2)}{\partial x_2} &= ME_2(x_1,x_2) \\
&\quad + \int\int g(x_2,a,e)\frac{\partial f_{\alpha_i,\varepsilon_{i2}|X_i}(a,e|x)}{\partial x_2}\,da\,de
\end{aligned}$$

The left hand side of the above derivatives are observable from the data. In contrast, the right hand side involves objects that are unknown to the econometrician, specifically the distribution of the errors $f_{\alpha_i,\varepsilon_{it}|X_i}$ and their associated derivatives $\partial f_{\alpha_i,\varepsilon_{it}|X_i}/\partial x$. To recover $ME_1(x_1,x_2)$ and $ME_2(x_1,x_2)$ from the four preceding equations, we have to make further assumptions since there are more unknowns than the number of equations. It is not enough that we assume a form of time homogeneity (which ensures that a repeated measurement will be beneficial with respect to controlling for α_i), i.e.

$$f_{\alpha_i,\varepsilon_{i1}|X_i} = f_{\alpha_i,\varepsilon_{i2}|X_i}$$

because we are still unable to completely remove the distortion caused by the effect of changing x_1 or x_2 on the distribution of the errors. In addition, we have to condition on the set where $X_{i1}=X_{i2}=x$ to completely remove this distortion. As a result, we are able to identify the marginal effects by conditioning on an appropriate set under no assumptions about the nonseparable model and the distribution of the errors (aside from time homogeneity):[9]

$$\begin{aligned}
ME_1(x) &= \frac{\partial \mathbb{E}(Y_{i1}|X_{i1}=X_{i2}=x)}{\partial x_1} - \frac{\partial \mathbb{E}(Y_{i2}|X_{i1}=X_{i2}=x)}{\partial x_1}, \\
ME_2(x) &= \frac{\partial \mathbb{E}(Y_{i2}|X_{i1}=X_{i2}=x)}{\partial x_2} - \frac{\partial \mathbb{E}(Y_{i1}|X_{i1}=X_{i2}=x)}{\partial x_2}.
\end{aligned}$$

Without conditioning on $X_{i1}=X_{i2}=x$, there are multiple avenues to recover the average marginal effect. In general, we can only partially identify the average marginal effect when there are bounds on $g(X_{it},\alpha_i,\varepsilon_{it})$ (see Chernozhukov et al. (2013)).

To avoid conditioning on $X_{i1}=X_{i2}=x$, we may consider correlated random effects strategies that use exchangeability (see Altonji and Matzkin (2005) for more) and dimension reduction to construct "instruments" that allow us to nullify the distortions brought about by the effect of changing x on the distribution of the errors. Bester and Hansen (2009b) show that if there exists a sufficient statistic that could reduce the dimension of the conditioning set $\{X_{i1}=x_1,X_{i2}=x_2,\ldots,X_{iT}=x_T\}$, then it is possible to recover the average marginal effect if $T\geq 3$. Bester and Hansen (2009b) are actually able to weaken the assumption of time homogeneity in this

[9] Imposing further restrictions may help in trading off some assumptions for others. The gains will have to be explored on a case-by-case basis.

case. Testing some of these assumptions is the subject of Ghanem (2015).

Note that the discussion so far focuses on strictly exogenous regressors. Extending the ideas to dynamic models are not very straightforward under the conditions maintained in the earlier discussion. Bounds for the dynamic model are also available in Chernozhukov et al. (2013). Parametric approaches that fully specify the distribution of the errors are available in Wooldridge (2005b). Large-T bias corrections of marginal effects obtained from parametric fixed-effects models can be found in Hahn and Newey (2004), Bester and Hansen (2009a), and Fernandez-Val (2009). ■

In the final example, I show that reducing the support of the distribution of the fixed effects may be helpful in identification of structural parameters. The lack of point identification of structural parameters in nonlinear panel data models has been documented by Honoré and Tamer (2006) and Chamberlain (2010) if we leave the distribution of the fixed effects unspecified. This lack of point identification can also be illustrated in the next example.

Example 1.2.7. (Bajari et al. (2011) and Bonhomme (2012)) Consider the following panel binary choice model with strictly exogenous regressors x_{it}:

$$\Pr(y_{it} = 1|x_i, \alpha_i) = H(\alpha_i + \beta x_{it}),\ i = 1, \ldots, n;\ t = 1, \ldots, T \tag{1.2.11}$$

where the distribution of the individual-specific fixed effect α_i given $x_i = (x_{i1}, \ldots, x_{iT})$ has finite and discrete support, i.e.

$$\Pr(\alpha_i = \alpha_k|x_i = x) = \pi_{x,k},\ k = 1, \ldots, K.$$

A fixed-effects setup means that we leave the $\pi_{x,k}$'s unspecified and possibly dependent on x. Assume further that the inverse link function H is specified in advance. Since the α_i's are unobservable, we have to look at the full conditional distribution of $y_i = (y_{i1}, \ldots, y_{iT})$ given $x_i = x$ alone. As a consequence of the law of total probability, this full conditional distribution can be written as

$$\Pr(y_i = y|x_i = x) = \sum_{k=1}^{K} \Pr(y_i = y|x_i = x, \alpha_i = \alpha_k; \beta) \Pr(\alpha_i = \alpha_k|x_i = x) \tag{1.2.12}$$

for some binary sequence y . The left hand side of (1.2.12) can be recovered from the data on frequencies of each of the 2^T possible binary sequences. We can collect every (1.2.12) for each possible binary sequence so that we have a matrix equation

$$P_{y|x} = P_x(\beta)\, \pi_x,$$

where $\pi_x = \left(\pi_{x,1}, \ldots, \pi_{x,K}\right)^{\mathrm{T}}$ is a $K \times 1$ vector, $P_x(\beta)$ is a $2^T \times K$ matrix based on the

specification (1.2.11), and $P_{y|x}$ is a $2^T \times 1$ vector of conditional probabilities observed from the data.

Instead of differencing out every α_i which is not generalizable outside linear models, we difference out π_x by annihilating the matrix $P_x(\beta)$, i.e.

$$\left[I - P_x(\beta) P_x(\beta)^-\right] P_{y|x} = \left[I - P_x(\beta) P_x(\beta)^-\right] P_x(\beta) \pi_x = 0. \tag{1.2.13}$$

Note that $P_x(\beta)^-$ is the Moore-Penrose inverse of $P_x(\beta)$. The main message behind (1.2.13) is not that it is possible to construct moment conditions that do not depend on α_i but that the rank of the matrix $P_x(\beta)$ matters. If we know that $K \geq 2^T$, then (1.2.13) is not informative of β at all. On the other hand, considering models for the fixed effects for which $K < 2^T$ may be useful. We can interpret $K < 2^T$ as the support of the fixed effects being less rich than the support of outcomes. In general, we will not know whether $K < 2^T$ or otherwise. There are empirical situations, such as the game-theoretic model estimated by Hahn and Moon (2010) and the one discussed in Chapter 3, where we would know the value of K.

In Hahn and Moon (2010), the reduced support of the fixed effects arises because the fixed effects represent which of the two pure strategy equilibria is selected by players and maintained over time. Further work that allow for time-varying fixed effects with limited support has been studied by Bonhomme and Manresa (2015). The latter paper and Hahn and Moon (2010) have shown that bias correction in a large-n, large-T context like we have seen in Example 1.2.3 is not needed at all. ■

1.3 How should we respond?

The discussion in the previous section comes from a perspective which emphasizes either the elimination of nuisance parameters or the robustification of estimation and inference methods in the presence of nuisance parameters. Furthermore, the distribution of the fixed effects is left unspecified as seen in Examples 1.2.1 to 1.2.5. As models become more complicated, this emphasis may become increasingly untenable, especially when there is meaning to be attached to nuisance parameters or when interest centers on functions of interest and nuisance parameters as seen in Examples 1.2.6 and 1.2.7. Empirically relevant models also have to allow for dynamics and predetermined regressors. Therefore, we need to search for methods that work in slightly complicated settings at the cost of making assumptions that may nevertheless be motivated theoretically or empirically. I now describe the ideas pursued in the succeeding chapters.

Many empirical situations (see Chapter 2 for examples) call for the estimation of a dynamic binary choice model with fixed effects. In Chapter 2, I demonstrate that it is inappropriate to estimate such a model by applying IV to a dynamic linear probability model. Motivations behind the use of a dynamic linear probability model

include the ability to directly recover average marginal effects and the availability of software without additional programming. IV or GMM based estimators of the dynamic linear probability model can also allow for predetermined regressors. We saw the difficulties in recovering average marginal effects and allowing for dynamics in Example 1.2.6. The main results of the chapter actually suggest that IV estimators of the linear probability model converge to an average marginal effect with incorrect weighting. Furthermore, this large-n limit might not even be found inside the large-n limit of the bounds proposed by Chernozhukov et al. (2013). In addition, these IV estimators do not converge to the true average marginal effect even as $T \to \infty$. As a result, this chapter gives an example for which dealing with the incidental parameter problem using IV may not be a good response.

Another empirical situation of interest involves the estimation of simultaneously determined discrete outcomes. Allowing for fixed effects in these models has not been explored fully, since most research has focused on either cross-sectional models, continuous outcomes, or random effects (see for example, the research by Cornwell, Schmidt, and Wyhowski (1992), Leon-Gonzalez (2003), Matzkin (2008), Matzkin (2012), and Masten (2015)). Parameter identification in these models is further complicated by the nonexistence of a unique reduced form. One way of partially resolving the identification problem is to introduce coherency conditions. Unfortunately, the coherency condition needs to be imposed a priori. In Chapter 3, I propose using panel data to estimate such models by allowing the data to determine how the coherency condition will hold. The manner in which the coherency condition holds can be represented as an incidental parameter that has finite support, in the spirit of what we have seen in Example 1.2.7.

The discussion in Example 1.2.3 is an estimator-based bias correction. One will observe that papers proposing an analytical correction of the estimator typically motivate the correction using the score. In Chapter 4, I develop a score-based correction involving projections. This approach is a useful and intuitive alternative when constructing estimating equations for the structural parameters that are relatively insensitive to inconsistent plug-ins. I show that the method can produce familiar estimators in special cases. I also show that projection exploits correct specification to reap the gains from bias reduction especially when T is very small.

Although the notion of time-invariant heterogeneity is hardly unique, a large gap exists between specifications where we allow for full heterogeneity (i.e., acknowledging that all units are different from each other) and full homogeneity (i.e., acknowledging that all units are the same). This gap enables us to explore different notions of partial pooling. Researchers acknowledge that units might be different from one another yet they may believe that some units are more alike than others. Despite this, they might be unwilling to specify which units are different from each other and which units are similar to one another. I formalize the preceding intuition by allowing some incidental parameters to take on the same value, namely zero.

In Chapter 5, I demonstrate that some notion of sparsity of the incidental parameters may be useful in constructing fixed-T consistent estimators that converge at the root-n rate. In particular, I tune the lasso (see Tibshirani (1996; 2011) and Chapter 5 for more) so that it will be able to detect the non-zero incidental parameters. A subsample for which the incidental parameters are set to zero can then be used for estimation and inference. This is in contrast to the machine learning and big data literature where the main developments have concentrated on uncovering non-zero effects in a sea of zero effects.

The four essays included in this thesis demonstrate several different ways to cope with the incidental parameter problem. None of these essays offer a general solution. Instead, these essays provide situations for which the incidental parameter problem may not be a serious impediment to theoretical and empirical work. However, I restrict myself to parametric situations and leave the nonparametric situations to future research.