Department of Economics

Discussion paper 2014-03

SPLIT-PANEL JACKKNIFE ESTIMATION OF FIXED-EFFECT MODELS

Geert Dhaene
Koen Jochmans

Sciences Po Economics Discussion Papers

SPLIT-PANEL JACKKNIFE ESTIMATION OF FIXED-EFFECT MODELS

Geert Dhaene

University of Leuven, Department of Economics, Naamsestraat 69, B-3000 Leuven, Belgium
geert.dhaene@kuleuven.be

Koen Jochmans

Sciences Po, Department of Economics, 28 rue des Saints-Pères, 75007 Paris, France
koen.jochmans@sciencespo.fr

This version: January 22, 2014

Maximum-likelihood estimation of nonlinear models with fixed effects is subject to the incidental-parameter problem. This typically implies that point estimates suffer from large bias and confidence intervals have poor coverage. This paper presents a jackknife method to reduce this bias and to obtain confidence intervals that are correctly centered under rectangular-array asymptotics. The method is explicitly designed to handle dynamics in the data and yields estimators that are straightforward to implement and that can be readily applied to a range of models and estimands. We provide distribution theory for estimators of index coefficients and average effects, present validity tests for the jackknife, and consider extensions to higher-order bias correction and to two-step estimation problems. An empirical illustration on female labor-force participation is also provided.

Keywords: *bias reduction, dependent data, incidental-parameter problem, jackknife, nonlinear model.*

INTRODUCTION

The analysis of panel data plays an important role in empirical economics. Starting with classic work on investment (Kuh 1959) and production functions (Mundlak 1961; Hoch 1962), panel data have been used to investigate a variety of questions, including the patents-R&D relationship (Hausman, Hall, and Griliches 1984), the dynamics of earnings (Lillard and Willis 1978) and health (Contoyannis, Jones, and Rice 2004), female labor-force participation (Heckman and MaCurdy 1980; Hyslop 1999), consumption and transitory income (Hall and Mishkin 1982), addiction and price effects (Becker, Grossman, and Murphy 1994), legalized abortion and crime (Donohue and Levitt 2001), production frontiers (Schmidt and Sickles 1984), FDI and productivity spillovers (Haddad and Harrison 1993; Javorcik 2004), spatial dynamics of FDI (Blonigen, Davies, Waddell, and Naughton 2007), and cross-country growth convergence (Islam 1995). An important aspect of empirical panel data models is that they typically feature unit-specific effects meant to capture unobserved heterogeneity.

Random-effect approaches to modeling unobserved heterogeneity often specify the distribution of the unit-specific effects and how these relate to the observed covariates, which may result in specification errors. The problem is further complicated in dynamic models because of the initial-condition problem (see, e.g., Heckman 1981b and Wooldridge 2005 for discussions).

Fixed-effect approaches, where the unit-specific effects are treated as parameters to be estimated and inference is performed conditional on the initial observations, are conceptually an attractive alternative. However, in fixed-effect models the incidental-parameter problem arises (Neyman and Scott 1948). That is, maximum-likelihood estimates of the parameters of interest are typically not consistent under asymptotics where the number of units, N, grows large but the number of observations per unit, T, is held fixed. Attempts to solve the incidental-parameter problem have been successful only in a few models, and the solutions generally do not give guidance to estimating average marginal effects, which are quantities of substantial interest. Furthermore, they restrict the fixed effects to be univariate, often entering the model as location parameters. Arellano and Honoré (2001) provide an overview of these methods. Browning and Carro (2007), Browning, Ejrnæs, and Alvarez (2010), and Arellano and Bonhomme (2012) discuss several examples where unit-specific location parameters cannot fully capture the unobserved heterogeneity in the data. Hospido (2012) and Carro and Traferri (2012) present empirical applications using models with multivariate fixed effects.

The incidental-parameter problem is most severe in short panels. Fortunately, in recent decades longer data sets are becoming available. For example, the PSID has been collecting annual waves since 1968 and the BHPS since 1991. They now feature a time-series dimension that can be considered statistically informative about unit-specific parameters. The availability of more observations per unit does not necessarily solve the inference problem, however, because confidence intervals centered at the maximum-likelihood estimate are incorrect under rectangular-array asymptotics, i.e., as $N, T \to \infty$ at the same rate (see, e.g., Li, Lindsay, and Waterman 2003). It has, though, motivated a recent literature in search of bias corrections to maximum likelihood that have desirable properties under rectangular-array asymptotics for a general class of fixed-effect models. Hahn and Newey (2004) and Hahn and Kuersteiner (2011) provide such corrections for static and dynamic models, respectively. Lancaster (2002), Woutersen (2002), Arellano and Hahn (2006), and Arellano and Bonhomme (2009) propose estimators that maximize modified objective functions and enjoy the same type of asymptotic properties. The primary aim of these methods is to remove the leading bias from the maximum-likelihood estimator and, thereby, to recenter its asymptotic distribution. The main difference between the various methods lies in how the bias is estimated. With the exception of the delete-one panel jackknife proposed in Hahn and Newey (2004) for independent data, all existing methods require analytical work that is both model and estimand specific, and may be computationally involved.

In this paper we propose jackknife estimators that correct for incidental-parameter bias in nonlinear dynamic fixed-effect models. In its simplest form, the jackknife estimates (and subsequently removes) the bias by comparing the maximum-likelihood estimate from the full panel with estimates computed from subpanels. Here, subpanels are panels with fewer observations per unit. The subpanels are taken as blocks, so that they preserve the dependency structure of the full panel. This jackknife estimator is very easy to implement. It requires only a routine to compute maximum-likelihood estimates; no analytical work is needed. A key feature of the jackknife is that, unlike analytical approaches to bias correction, the jackknife does not need an explicit characterization of the incidental-parameter bias. Therefore, it can be readily applied to estimate index coefficients, average marginal effects, models with multiple fixed effects per unit, and multiple-equation models. It can also deal with feedback from lagged outcomes on covariates and with generated regressors, which arise when accounting for endogeneity or sample selection, for example. Both types of complications are known to affect the expression of the incidental-parameter bias—see Bun and Kiviet (2006) and Fernández-Val and Vella (2011), respectively—but pose no additional difficulty for the jackknife.

In Section 1 we start with a discussion of the incidental-parameter problem, and we present and motivate our framework. In Section 2 we introduce split-panel jackknife estimators of model parameters and average effects, and provide distribution theory. This section also gives an assessment of the regularity conditions, presents tests of the validity of the jackknife, and compares the jackknife estimators with other bias-correction methods by means of Monte Carlo simulations. Section 3 discusses extensions of the split-panel jackknife to higher-order bias correction and to two-step estimators. Section 4 presents an empirical illustration of bias-corrected estimation in the context of female labor-force participation. We end the paper with some suggestions for future research. Proofs, technical details, and additional results are available as supplementary material.

1. FIXED-EFFECT ESTIMATION AND INCIDENTAL-PARAMETER BIAS

Suppose that we are given data z_{it} for individual units $i = 1, 2, \ldots, N$ and time periods $t = 1, 2, \ldots, T$. Let z_{it} have density $f(z_{it}; \theta_0, \alpha_{i0})$, which is known up to the finite-dimensional parameters $\theta_0 \in \Theta$ and $\alpha_{i0} \in \mathcal{A}$. The fixed-effect estimator of θ_0 is $\widehat{\theta} \equiv \arg\max_{\theta \in \Theta} \widehat{l}(\theta)$, where $\widehat{l}(\theta)$ is the (normalized) profile log-likelihood function, i.e.,

$$\widehat{l}(\theta) \equiv \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \log f(z_{it}; \theta, \widehat{\alpha}_i(\theta)), \qquad \widehat{\alpha}_i(\theta) \equiv \arg\max_{\alpha_i \in \mathcal{A}} \frac{1}{T} \sum_{t=1}^{T} \log f(z_{it}; \theta, \alpha_i).$$

It is well known that $\widehat{\theta}$ is often inconsistent for θ_0 under asymptotics where $N \to \infty$ and T remains fixed. That is, $\theta_T \equiv \text{plim}_{N\to\infty} \widehat{\theta} \neq \theta_0$. This is the incidental-parameter problem (Neyman and Scott 1948). The problem arises because of the estimation noise in $\widehat{\alpha}_i(\theta)$, which vanishes only as $T \to \infty$. Indeed, under regularity conditions,

$$\theta_T = \arg\max_{\theta \in \Theta} l_T(\theta), \qquad l_T(\theta) \equiv \overline{\mathbb{E}}[\log f(z_{it}; \theta, \widehat{\alpha}_i(\theta))],$$

where $\overline{\mathbb{E}}[\cdot] \equiv \lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} \mathbb{E}[\cdot]$, whereas

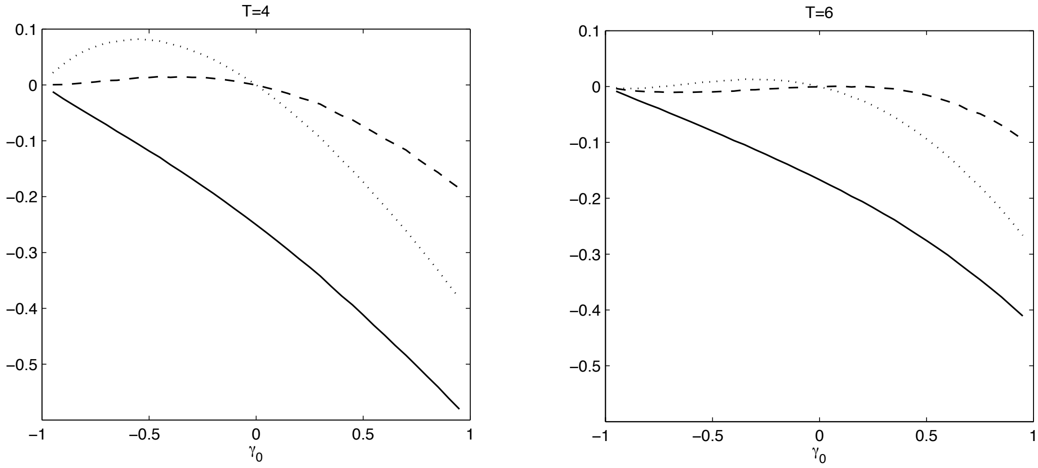
$$\theta_0 = \arg\max_{\theta \in \Theta} l_0(\theta), \qquad l_0(\theta) \equiv \overline{\mathbb{E}}[\log f(z_{it}; \theta, \alpha_i(\theta))],$$

with $\alpha_i(\theta) \equiv \arg\max_{\alpha \in \mathcal{A}} \mathbb{E}[\log f(z_{it}; \theta, \alpha_i)]$. With fixed T, $\widehat{\alpha}_i(\theta) \neq \alpha_i(\theta)$. Hence, the maximands $l_T(\theta)$ and $l_0(\theta)$ are different and so, in general, are their maximizers. The inconsistency (or asymptotic bias) can be large, even with moderately long panels.

Examples help to illustrate the incidental-parameter problem. In the classic example of Neyman and Scott (1948), the z_{it} are independent random variables that are distributed as $z_{it} \sim \mathcal{N}(\alpha_{i0}, \theta_0)$, and the maximum-likelihood estimator of θ_0 converges to $\theta_T = \theta_0 - \theta_0/T$. The inconsistency, $-\theta_0/T$, arises because maximum likelihood fails to make the degrees-of-freedom correction that accounts for replacing $\alpha_{i0} = \mathbb{E}[z_{it}]$ by its estimate $T^{-1} \sum_{t=1}^{T} z_{it}$. If we let $z_{it} = (y_{it}, x_{it})$ and $\theta_0 = (\gamma_0', \sigma_0^2)'$, a regression version of this example is $y_{it} \sim \mathcal{N}(\alpha_{i0} + x_{it}'\gamma_0, \sigma_0^2)$. Here, the maximum-likelihood estimator of γ_0 is the within-group estimator. When $x_{it} = y_{it-1}$ we obtain the Gaussian first-order autoregressive model, for which the incidental-parameter problem has been extensively studied. In this case, when $|\gamma_0| < 1$, $\gamma_T = \gamma_0 - (1+\gamma_0)/T + O(T^{-2})$ (Nickell 1981; Hahn and Kuersteiner 2002). Although these examples are very simple, they illustrate that, in sufficiently regular problems, $\theta_T - \theta_0$ is typically $O(T^{-1})$. Therefore, while $\widehat{\theta}$ will be consistent and asymptotically normal (under regularity conditions) as both $N, T \to \infty$, its asymptotic distribution will be incorrectly centered unless T grows faster than N (Li, Lindsay, and Waterman 2003; Hahn and Newey 2004). As a result, confidence intervals centered at the maximum-likelihood estimate will tend to have poor coverage rates in most microeconometric applications, where T is typically much smaller than N. The jackknife corrections that we introduce below aim to reduce the asymptotic bias of the maximum-likelihood estimator and to recenter its asymptotic distribution. Such an approach is in line with the recent work on nonlinear models for panel data mentioned above.

The jackknife method, which originated as a tool for bias reduction in the seminal work of Quenouille

Figure 1. Inconsistencies in the stationary Gaussian autoregression



Model: $y_{it} = \alpha_{i0} + \gamma_0 y_{it-1} + \varepsilon_{it}$, $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_0^2)$, stationary y_{i0}. Plots: fixed-T inconsistencies of the within-group estimator ($\widehat{\gamma}$, solid) and two jackknife estimators ($\widetilde{\gamma}_{1/2}$, dashed; $\dot{\gamma}_{1/2}$, dotted).

(1949, 1956), exploits variation in the sample size to obtain a nonparametric estimator of the bias. In our context, the (large N, fixed T) bias to be corrected for is $\theta_T - \theta_0$ and the relevant sample size is T, the length of the panel. We will discuss two types of jackknife estimators of θ_0. The first type bias-corrects $\widehat{\theta}$ directly. The second type solves a bias-corrected maximization problem, where the jackknife bias-corrects the objective function $\widehat{l}(\theta)$ prior to maximization. These two types of estimators can be seen as automatic counterparts to the analytical procedures introduced by Hahn and Kuersteiner (2011) and Arellano and Hahn (2006), respectively. The former type is particularly easy to implement as it requires only the computation of a few maximum-likelihood estimates. The latter, while computationally a little more involved, is still generic in terms of applicability and has some advantages, such as equivariance with respect to one-to-one reparameterizations.

The jackknife estimators proposed in this paper differ from the delete-one panel jackknife of Hahn and Newey (2004) in that they allow for dependence between observations on a given unit. Such dependence is natural in most applications and is inherent in dynamic models, such as the Gaussian autoregression or a binary-choice version thereof. The key to handling dynamics is to use subpanels formed by *consecutive* observations for each unit. Of course, some regularity has to be put on the time-series properties of the data. A convenient assumption is to impose stationarity of the individual processes and a sufficient degree of mixing. In applications, however, stationarity may be an unrealistic assumption. Therefore, we will also examine the performance of the jackknife estimators in some specific non-stationary cases and develop tests of the validity of the jackknife corrections.

The jackknife will be shown to remove the $O(T^{-1})$ term of the bias. Hence, in the Neyman and Scott (1948) example, it fully eliminates the bias. More generally, however, the jackknife will only reduce the bias from $O(T^{-1})$ down to $o(T^{-1})$. Nevertheless, for typical sample sizes encountered in practice, this can already be sufficient for a vast reduction in bias and much improved confidence intervals. To illustrate the reduction in bias, Figure 1 plots the inconsistencies of the within-group estimator ($\widehat{\gamma}$, solid) and of the jackknife estimators obtained from correcting $\widehat{\gamma}$ (denoted $\widetilde{\gamma}_{1/2}$, dashed) and from correcting the objective function (denoted $\dot{\gamma}_{1/2}$, dotted), in the stationary Gaussian autoregressive model $y_{it} = \alpha_{i0} + \gamma_0 y_{it-1} + \varepsilon_{it}$. These jackknife estimators will be defined in (2.5) and (2.8) below. The plots show that the jackknife corrections alleviate the Nickell (1981) bias to a large extent, even in short panels ($T = 4, 6$). To gain an idea of the finite-sample performance of bias-corrected estimation, Table 1 shows the results of a small simulation experiment in this model for $\gamma_0 = .5$ and various panel sizes. The biases and the coverage rates of 95% confidence intervals centered at the point estimates are given for $\widehat{\gamma}$, the bias-corrected plug-in estimator $\widetilde{\gamma}_{\mathrm{HK}} = \widehat{\gamma} + (1+\widehat{\gamma})/T$ (see Hahn and Kuersteiner 2002), and the jackknife bias-corrections $\widetilde{\gamma}_{1/2}$ and $\dot{\gamma}_{1/2}$. The inconsistency of the bias-corrected estimators in this model is $O(T^{-2})$. The table also provides results for the optimally-weighted Arellano and Bond (1991) estimator, $\widehat{\gamma}_{\mathrm{AB}}$, which is fixed-$T$ consistent. In line with Figure 1, the results show that bias correction can lead to drastic reductions in small-sample bias. The jackknife corrections are competitive with $\widehat{\gamma}_{\mathrm{AB}}$ in terms of bias (for the sample sizes considered). Furthermore, bias correction leads to much improved coverage rates of the confidence intervals compared with those based on maximum likelihood. The corrections remove enough bias to yield reliable confidence intervals also when T is not small relative to N. Finally, the last two columns of Table 1, $\tilde{t}_{1/2}$ and $\dot{t}_{1/2}$, present the acceptance rates of two 5%-level tests (which will be defined later on) to check the validity of the jackknife corrections. The underlying null hypothesis of the tests is that the jackknife effectively removes the leading bias from the maximum-likelihood estimator. In this example, the acceptance rates are close to the nominal acceptance rate of 95%, confirming that the jackknife is bias-reducing.

Table 1. Small-sample performance in the stationary Gaussian autoregression

		bias					confidence					validity	
N	T	$\widehat{\gamma}$	$\widetilde{\gamma}_{\mathrm{HK}}$	$\widetilde{\gamma}_{1/2}$	$\dot{\gamma}_{1/2}$	$\widehat{\gamma}_{\mathrm{AB}}$	$\widehat{\gamma}$	$\widetilde{\gamma}_{\mathrm{HK}}$	$\widetilde{\gamma}_{1/2}$	$\dot{\gamma}_{1/2}$	$\widehat{\gamma}_{\mathrm{AB}}$	$\tilde{t}_{1/2}$	$\dot{t}_{1/2}$
100	4	−.413	−.141	−.076	−.176	−.054	.000	.495	.682	.273	.923	.953	.735
100	6	−.278	−.074	−.019	−.097	−.047	.000	.702	.815	.509	.910	.966	.878
100	8	−.206	−.044	.001	−.058	−.039	.000	.815	.848	.702	.910	.964	.916
100	12	−.134	−.021	.008	−.027	−.031	.001	.897	.866	.853	.900	.957	.935
20	20	−.081	−.010	.005	−.012	−.089	.595	.947	.903	.935	.613	.956	.951
50	50	−.031	−.002	.001	−.002	−.033	.592	.950	.934	.939	.603	.947	.946
100	100	−.015	.000	.000	.000	−.016	.596	.948	.939	.941	.605	.950	.949

Model: $y_{it} = \alpha_{i0} + \gamma_0 y_{it-1} + \varepsilon_{it}$, $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_0^2)$, stationary y_{i0}. Data generated with $\gamma_0 = .5$, $\sigma_0^2 = 1$, $\alpha_{i0} \sim \mathcal{N}(0, 1)$. 10,000 Monte Carlo replications.

The linear autoregressive model is convenient for illustrative purposes because a benchmark is available in the form of the Arellano and Bond (1991) estimator. From a fixed-T perspective there is no theoretical ground to prefer bias-corrected estimators over this estimator. The situation is different under rectangular-array asymptotics, where the bias-corrected estimators are asymptotically efficient and the Arellano and Bond (1991) estimator is asymptotically biased; see Hahn and Kuersteiner (2002) and Alvarez and Arellano (2003), respectively. Furthermore, in nonlinear models, fixed-T approaches are typically not available. For example, in the dynamic binary-choice model where $z_{it} = (y_{it}, y_{it-1})$ and $\Pr[y_{it} = 1 | y_{it-1} = x] = F(\alpha_{i0} + \theta_0 x)$ for $x = 0, 1$ and a given distribution function F, a fixed-T consistent estimator of θ_0 is available when F is logistic (Chamberlain 1985), but when F is Gaussian θ_0 is not point identified for small T (Honoré and Tamer 2006; see also Chamberlain 2010 on the lack of point identification). In such situations, bias-corrected estimation can be an attractive option. To illustrate, Table 2 provides simulation results for the jackknife corrections in the stationary dynamic probit model where $\theta_0 = .5$. Again, the reduction in bias is substantial, and so is the improvement of the 95% confidence intervals.

Table 2. Small-sample performance in the stationary autoregressive probit model

	bias			confidence			validity	
T	$\widehat{\theta}$	$\widetilde{\theta}_{1/2}$	$\dot{\theta}_{1/2}$	$\widehat{\theta}$	$\widetilde{\theta}_{1/2}$	$\dot{\theta}_{1/2}$	$\tilde{t}_{1/2}$	$\dot{t}_{1/2}$
6	$-.618$	.248	$-.272$	.031	.833	.895	.959	.929
8	$-.456$	.078	$-.162$	.079	.917	.889	.956	.951
12	$-.300$	.021	$-.074$	.194	.934	.923	.962	.962
18	$-.197$	.008	$-.031$	.354	.943	.943	.954	.954

Model: $y_{it} = 1(\alpha_{i0} + \theta_0 y_{it-1} + \varepsilon_{it} > 0)$, $\varepsilon_{it} \sim \mathcal{N}(0,1)$, stationary y_{i0}. Data generated with $N = 100$, $\theta_0 = .5$, $\alpha_{i0} \sim \mathcal{N}(0,1)$. 10,000 Monte Carlo replications.

In the next section we will present jackknife estimators of θ_0 and compare them with other approaches available in the literature. We will also present jackknife bias corrections for average (marginal or other) effects, where the averaging is over the fixed effects and, possibly, over covariates (Chamberlain 1984). Averages like this are often parameters of substantial interest. In the Gaussian autoregression, one such quantity would be the survival function at s, i.e.,

$$\int_{-\infty}^{+\infty} \Pr[y_{it} \geq s | y_{it-1} = x, \alpha_{i0} = \alpha] \,\mathrm{d}G(\alpha) = \int_{-\infty}^{+\infty} \Phi\left(\frac{\alpha + x'\gamma_0 - s}{\sigma_0}\right) \,\mathrm{d}G(\alpha),$$

where G denotes the marginal distribution of the α_{i0}. The analog in the dynamic binary-choice model would be the choice probability $F(\alpha_{i0} + x\theta_0)$ averaged against G. Plug-in estimators of such averages based on maximum-likelihood estimates will typically be inconsistent. Again, in regular problems, the asymptotic bias will generally be $O(T^{-1})$. Using a bias-corrected estimate of θ_0 instead of $\widehat{\theta}$ leaves the order of the bias unchanged. Moreover, even if the true θ_0 were used, the bias would remain $O(T^{-1})$ because the α_{i0} are not estimated consistently for small T. However, the idea underlying the jackknife estimators of θ_0 can readily be applied to obtain bias-corrected average-effect estimators.

2. SPLIT-PANEL JACKKNIFE ESTIMATION

In this section we present our jackknife corrections and provide sufficient conditions for them to improve on maximum likelihood. We will work under the following assumption.

Assumption 2.1. *The processes z_{it} are independent across i, and stationary and alpha mixing across t, with mixing coefficients $a_i(m)$ that are uniformly exponentially decreasing, i.e., $\sup_i |a_i(m)| < Cb^m$ for some finite $C > 0$ and b such that $0 < b < 1$, where*

$$a_i(m) \equiv \sup_t \sup_{A \in \mathcal{A}_{it}, B \in \mathcal{B}_{it+m}} |\Pr(A \cap B) - \Pr(A)\Pr(B)|,$$

and $\mathcal{A}_{it} \equiv \sigma(z_{it}, z_{it-1}, \ldots)$ and $\mathcal{B}_{it} \equiv \sigma(z_{it}, z_{it+1}, \ldots)$ are the sigma algebras generated by $z_{it}, z_{it-1}, \ldots$ and $z_{it}, z_{it+1}, \ldots$, respectively. The density of z_{it} given $z_{it-1}, z_{it-2}, \ldots$(relative to some dominating measure) is $f(z_{it}; \theta_0, \alpha_{i0})$ where (θ_0, α_{i0}) is the unique maximizer of $\mathbb{E}[\log f(z_{it}; \theta, \alpha_i)]$ over the Euclidean parameter space $\Theta \times \mathcal{A}$ and is interior to it.

This assumption accommodates dynamic models by letting $z_{it} = (y_{it}, x_{it})$ and $f(z_{it}; \theta, \alpha_i) = f(y_{it}|x_{it}; \theta, \alpha_i)$, where x_{it} may contain past values of the outcome variable y_{it}. The density is assumed to be dynamically complete, but the assumption allows for feedback from past outcomes on covariates. We assume that the data are independent across i. The time-series processes may be heterogeneous across i with a uniform upper bound on the temporal dependencies that decays exponentially. Hahn and Kuersteiner (2011) provide a detailed discussion of the stationarity and mixing assumptions. Hahn and Kuersteiner (2010, 2011) and de Jong and Woutersen (2011) show that they hold under mild conditions in several popular nonlinear models, including dynamic binary-choice models and dynamic tobit models with exogenous covariates. The last part of Assumption 2.1 essentially states that the parameters θ_0 and α_{i0} are identifiable from within-group variation in the data.

Assumption 2.1 is standard in the literature on fixed-effect estimation under rectangular-array asymptotics; compare with Condition 3 in Hahn and Kuersteiner (2011) and Assumption 3 in Arellano and Hahn (2006). As we mentioned, the stationarity assumption may not be realistic in certain applications. For example, it rules out time trends and time dummies, which are often included in empirical models. Accounting for such aggregate time effects is difficult in nonlinear fixed-effect models, even in settings where fixed-T inference would otherwise be feasible (see Honoré and Kyriazidou 2000 and Honoré and Tamer 2006). In recent work, Bai (2009, 2013) deals with time effects in linear panel models under asymptotics where both $N, T \to \infty$. In dynamic models, stationarity further requires that the initial observations are drawn from their respective stationary distributions or, equivalently, that the processes started in the distant past. We will discuss the sensitivity of bias corrections to violations of this assumption below.

2.1. Correcting the estimator

Let $s_{it}(\theta) \equiv \nabla_\theta \log f(z_{it}; \theta, \alpha_i(\theta))$ and $H_{it}(\theta) \equiv \nabla_{\theta\theta'} \log f(z_{it}; \theta, \alpha_i(\theta))$ be the contributions to the infeasible profile score and Hessian matrix, respectively. Let $\Sigma \equiv -\overline{\mathbb{E}}[H_{it}(\theta_0)]$. We will restrict attention to models satisfying the following two conditions.

ASSUMPTION 2.2. *θ_T and Σ exist, and*

$$\sqrt{NT}(\widehat{\theta} - \theta_T) = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma^{-1} s_{it}(\theta_0) + o_p(1)$$

as $N, T \to \infty$.

ASSUMPTION 2.3. *As $T \to \infty$,*

$$\theta_T - \theta_0 = \frac{B_1}{T} + o\left(\frac{1}{T}\right),$$

where B_1 is a constant.

Assumption 2.2 is the usual influence-function representation of the maximum-likelihood estimator when centered around its probability limit, and is a mild requirement. Because $\widehat{\theta}$ is consistent as $T \to \infty$, it holds that $\theta_T - \theta_0 \to 0$ as $T \to \infty$. Assumption 2.3 is a high-level condition on how the bias shrinks. Hahn and Newey (2004) and Hahn and Kuersteiner (2011) provide primitive conditions under which these assumptions are satisfied in static and dynamic models, respectively.

Put together, these assumptions imply that, as $N, T \to \infty$ such that $N/T \to \rho$ for some $\rho \in (0, \infty)$, we have

$$\sqrt{NT}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(B_1\sqrt{\rho}, \Sigma^{-1}).$$

As a result, confidence intervals for θ_0 centered at $\widehat{\theta}$ would be expected to have poor coverage even in panels where T is of the same order of magnitude as N.

We now use the jackknife to obtain a non-parametric estimator of B_1/T, the leading bias term of $\widehat{\theta}$. This bias term generally depends on the data generating process in a complicated way. Hahn and Kuersteiner (2011) derive the exact form of B_1 and present a plug-in estimator of it based on the maximum-likelihood estimator of θ_0 and the α_{i0}. Here we estimate B_1/T by means of a linear combination of $\widehat{\theta}$ and estimators based on subpanels. For our purposes a subpanel is defined as a proper subset $S \subsetneq \{1, 2, \ldots, T\}$ such that the elements of S are consecutive integers and $|S| \geq T_{\min}$, where $|S|$ denotes the cardinality of S and $T_{\min}$ is the smallest T for which θ_T exists. Now, the maximum-likelihood estimator corresponding to subpanel S is

$$\widehat{\theta}_S \equiv \arg\max_{\theta \in \Theta} \widehat{l}_S(\theta) \ , \qquad \widehat{l}_S(\theta) \equiv \frac{1}{N|S|} \sum_{i=1}^{N} \sum_{t \in S} \log f(z_{it}; \theta, \widehat{\alpha}_{iS}(\theta)),$$

where $\widehat{\alpha}_{iS}(\theta) \equiv \arg\max_{\alpha_i \in \mathcal{A}} \frac{1}{|S|} \sum_{t \in S} \log f(z_{it}; \theta, \alpha_i)$. Since, by their very definition, subpanels preserve the dependency structure of the full panel, our assumptions imply that $\mathrm{plim}_{N \to \infty} \widehat{\theta}_S = \theta_{|S|}$ and, as $|S| \to \infty$, $\theta_{|S|}$ can be expanded as in Assumption 2.3, with $|S|$ replacing T. It thus follows that

$$\frac{|S|}{T - |S|}(\theta_S - \theta_T) = \frac{B_1}{T} + o\left(\frac{1}{T}\right), \tag{2.1}$$

and that $\frac{|S|}{T-|S|}(\widehat{\theta}_S - \widehat{\theta})$ is a consistent estimator of B_1/T. Each subpanel S has associated with it an estimator $\widehat{\theta}_S$ that can be combined with $\widehat{\theta}$ to obtain an estimator of the leading bias. Different choices lead to jackknife estimators with different properties, which leads to the question of the optimal choice of subpanels.

Let $g \geq 2$ be an integer such that $T \geq gT_{\min}$. Suppose we split the panel into $\mathcal{S} = \{S_1, S_2, \ldots, S_g\}$, a collection of subpanels partitioning $\{1, 2, \ldots, T\}$ in such a way that the sequence $\min_{S \in \mathcal{S}} |S|/T$ is bounded away from zero as T grows. Then, with

$$\overline{\theta}_{\mathcal{S}} \equiv \sum_{S \in \mathcal{S}} \frac{|S|}{T} \widehat{\theta}_S, \tag{2.2}$$

$\frac{1}{g-1}(\overline{\theta}_{\mathcal{S}} - \widehat{\theta})$ is a consistent estimator of B_1/T based on the collection $\mathcal{S}$. Now, any such collection $\mathcal{S}$ defines an equivalence class $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m\}$ of collections of subpanels partitioning $\{1, 2, \ldots, T\}$ that have the same set of cardinalities as $\mathcal{S}$. Note that $m \leq g!$ and that $m = 1$ when all subpanels in $\mathcal{S}$ have cardinality T/g. Averaging $\frac{1}{g-1}(\overline{\theta}_{\mathcal{S}} - \widehat{\theta})$ over the equivalence class of $\mathcal{S}$ to estimate B_1/T removes any arbitrariness arising from a particular choice of partitioning for given cardinalities of the subpanels. Subtracting this estimate from $\widehat{\theta}$ yields the split-panel jackknife estimator

$$\widetilde{\theta} \equiv \frac{g}{g-1}\widehat{\theta} - \frac{1}{g-1}\overline{\theta}, \qquad \overline{\theta} \equiv \frac{1}{m}\sum_{j=1}^{m} \overline{\theta}_{\mathcal{S}_j}. \tag{2.3}$$

The following theorem gives the asymptotic behavior of this estimator.

THEOREM 2.1. *Let Assumptions 2.1, 2.2, and 2.3 hold. Then* $\mathrm{plim}_{N\to\infty}\widetilde{\theta} = \theta_0 + o(T^{-1})$ *and*

$$\sqrt{NT}(\widetilde{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1})$$

as $N, T \to \infty$ *with* $N/T \to \rho$.

This result states that, under the assumptions made, all members of the class $\widetilde{\theta}$ remove the leading bias from $\widehat{\theta}$ and have a normal limit distribution that is correctly centered under rectangular-array asymptotics. The asymptotic variance is the same as that of the maximum-likelihood estimator. The fact that bias reduction can be achieved without variance inflation is important. It arises here from the way in which the subpanels are combined to estimate the bias term. To see this, note that any $\widehat{\theta}_S$ in (2.2) has an asymptotic variance that is larger than that of $\widehat{\theta}$ because $|S| < T$. However, because each collection partitions $\{1, 2, \ldots, T\}$, averaging the subpanel estimators as in (2.2) brings the variance back down to that of maximum likelihood.

Thus, the split-panel jackknife estimator removes the leading bias from $\widehat{\theta}$ without affecting its asymptotic variance. Like other bias-corrected estimators, it does, however, affect the magnitude of the higher-order bias, i.e., the bias that is not removed. This is because B_1/T is estimated with bias $o(T^{-1})$; recall (2.1). For the split-panel jackknife estimators, the transformation of the higher-order bias is very transparent. To describe it, it is useful to assume for a moment that the inconsistency of $\widehat{\theta}$ can be expanded to a higher order, that is,

$$\theta_T - \theta_0 = \frac{B_1}{T} + \frac{B_2}{T^2} + \cdots + \frac{B_k}{T^k} + o\left(\frac{1}{T^k}\right) \tag{2.4}$$

for some integer k. While $\widetilde{\theta}$ eliminates B_1, it transforms the remaining B_j into B_j', say. Theorem S.2.1 in the supplementary material provides a characterization of this transformation. It shows that $|B_j'| > |B_j|$ for all $j \geq 2$ and that, for given g, any higher-order bias coefficient, B_j', is minimized (in absolute value) if and only if the collections $\mathcal{S}_j$ are almost-equal partitions of $\{1, 2, \ldots, T\}$, i.e., if $\lfloor T/g \rfloor \leq |S| \leq \lceil T/g \rceil$ for all $S \in \mathcal{S}_j$. With almost-equal partitions, the second-order bias term is $-gB_2/T^2$. Minimizing this term over g gives the half-panel jackknife estimator

$$\widetilde{\theta}_{1/2} \equiv 2\widehat{\theta} - \overline{\theta}_{1/2}, \tag{2.5}$$

which also minimizes the magnitude of all higher-order bias terms. Here, $\overline{\theta}_{1/2}$ is the average of $\overline{\theta}_{\mathcal{S}_1}$ and $\overline{\theta}_{\mathcal{S}_2}$ as defined in (2.2), with $\mathcal{S}_1 \equiv \{\{1, \ldots, \lceil T/2 \rceil\}; \{\lceil T/2 \rceil + 1, \ldots, T\}\}$ and $\mathcal{S}_2 \equiv \{\{1, \ldots, \lfloor T/2 \rfloor\}; \{\lfloor T/2 \rfloor + 1, \ldots, T\}\}$. When T is odd, $\mathcal{S}_1$ and $\mathcal{S}_2$ are the two possible ways of splitting the panel into two near half-panels; when T is even, $\mathcal{S}_1 = \mathcal{S}_2$ and the panel is split exactly into half-panels.

The half-panel jackknife estimator is simple to implement, requiring a few maximum-likelihood estimates. To compute these, an efficient algorithm will exploit the sparsity of the Hessian matrix, as suggested by Hall (1978) and Chamberlain (1980). This makes fixed-effect estimation and jackknife-based bias correction straightforward, even when the cross-sectional sample size is large or when α_i is a vector of individual effects. Furthermore, once the full-panel maximum-likelihood estimates have been computed, they are good starting values for computing the subpanel estimates. The asymptotic variance, finally, can be estimated using the point estimates to form a plug-in estimator $\widehat{\Sigma}^{-1}$. In our simulations, we estimated Σ using the Hessian matrix of the profile log-likelihood (estimates based on the variance of the profile score or on the sandwich formula yielded very similar results). For the linear dynamic model we applied a degree-of-freedom correction to account for the estimation of the error variance and, for the half-panel jackknife estimates of θ_0, we estimated Σ as the average of its two-halfpanel estimates.

A drawback of the half-panel jackknife estimator in (2.5) is that it cannot be applied when $T < 2T_{\min}$. One solution, provided that $T_{\min} < T$, is to resort to overlapping subpanels to construct jackknife estimators. Let g be a rational number strictly between 1 and 2 such that T is divisible by g. Let S_1 and S_2 be two overlapping subpanels such that $S_1 \cup S_2 = \{1, 2, \ldots, T\}$ and $|S_1| = |S_2| = T/g$. The estimator

$$\widetilde{\theta}_{1/g} \equiv \frac{g}{g-1}\widehat{\theta} - \frac{1}{g-1}\overline{\theta}_{1/g}, \qquad \overline{\theta}_{1/g} \equiv \frac{1}{2}(\widehat{\theta}_{S_1} + \widehat{\theta}_{S_2}), \tag{2.6}$$

is first-order unbiased. Furthermore, a calculation shows that, as $N, T \to \infty$ with $N/T \to \rho$,

$$\sqrt{\frac{NT}{d_g}}(\widetilde{\theta}_{1/g} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1})$$

where $d_g \equiv \frac{1}{2}g/(g-1)$. A formal derivation is available as Theorem S.3.1 in the supplementary material. The factor d_g is a variance inflation factor. It increases from one to infinity as the fraction of subpanel overlap increases from zero to one. The variance inflation can be interpreted as the price to be paid for bias correction via the jackknife in very short panels.[1] The analytical corrections of, e.g., Hahn and Kuersteiner (2011) and Arellano and Hahn (2006) do not have this drawback.

2.2. Correcting the objective function

As noted above, the incidental-parameter problem arises because the large N, fixed T profile log-likelihood, $l_T(\theta)$, approaches the infeasible objective function $l_0(\theta)$ only as $T \to \infty$. Equivalently, as $N \to \infty$ with fixed T, the profile score $\widehat{s}(\theta) \equiv \nabla_\theta \widehat{l}(\theta)$ converges to $s_T(\theta) \equiv \nabla_\theta l_T(\theta)$, which is generally non-zero at θ_0. Because θ_T solves $s_T(\theta) = 0$, the bias of the profile-score equation can be seen as the source for $\theta_T \neq \theta_0$. This suggests that, rather than correcting $\widehat{\theta}$, one may equally well correct for incidental-parameter bias by maximizing a bias-corrected profile log-likelihood. In the context of inference in the presence of nuisance parameters, such approaches have been the subject of much study in the statistics literature; see Sartori (2003) for a recent account and many references.

We now show that the split-panel jackknife can be applied to correct $\widehat{l}(\theta)$ in the same way as $\widehat{\theta}$. Let $\Delta(\theta) \equiv \lim_{N\to\infty} N^{-1} \sum_{i=1}^{\infty} \sum_{j=-\infty}^{\infty} \operatorname{cov}(s_{it}(\theta), s_{it-j}(\theta))$; note that $\Delta(\theta_0) = \Sigma$, as $s_{it}(\theta_0)$ is a martingale difference sequence and the information matrix equality holds. In analogy to Assumptions 2.2 and 2.3, we will work under the following two conditions.

Assumption 2.4. *There is a neighborhood $\mathcal{N}_0 \subseteq \Theta$ around θ_0 where both $s_T(\theta)$ and $\Delta(\theta)$ exist, and where*

$$\sqrt{NT}\,(\widehat{s}(\theta) - s_T(\theta)) = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} (s_{it}(\theta) - s_0(\theta)) + o_p(1)$$

as $N, T \to \infty$.

Assumption 2.5. *As $T \to \infty$,*

$$l_T(\theta) - l_0(\theta) = \frac{C_1(\theta)}{T} + o\left(\frac{1}{T}\right),$$

where $C_1(\theta)$ is a continuous function that has a bounded first derivative $C_1'(\theta)$ on $\mathcal{N}_0$.

Assumption 2.4 is an asymptotic-linearity condition on the profile score. Assumption 2.5 states that the bias of the profile log-likelihood has a leading term that is $O(T^{-1})$. Primitive conditions are available in Arellano and Hahn (2006).

These assumptions can be linked to Assumptions 2.2 and 2.3 as follows. A Taylor expansion of $s_T(\theta)$ around θ_0 gives

$$s_T(\theta_T) = s_T(\theta_0) - \Sigma\ (\theta_T - \theta_0) + o(\|\theta_T - \theta_0\|).$$

Because $s_T(\theta) = s_0(\theta) + C_1'(\theta)/T + o(1/T)$ on $\mathcal{N}_0$ and θ_T lies in $\mathcal{N}_0$ with probability approaching one as

[1] On the other hand, overlapping subpanels yield smaller inflation of the higher-order bias. From (2.4) and (2.6) it follows that

$$\operatorname{plim}_{N\to\infty} \widetilde{\theta}_{1/g} - \theta_0 = -gB_2/T^2 - g(1+g)B_3/T^3 - \ldots - g(1+g+\ldots+g^{k-2})B_k/T^k + o(T^{-k}).$$

Each bias term here is smaller (in magnitude) than the corresponding bias term of $\widetilde{\theta}_{1/2}$.

$T \to \infty$, we have

$$\theta_T - \theta_0 = \frac{\Sigma^{-1}\, C_1'(\theta_0)}{T} + o\left(\frac{1}{T}\right), \tag{2.7}$$

using $s_T(\theta_T) = 0$ and $s_0(\theta_0) = 0$. Thus, the leading bias of $\widehat{\theta}$, B_1/T, is the product of a Hessian term with the leading bias of the profile score.

Let $T'_{\min}$ be the least T for which $l_T(\theta)$ exists and is non-constant (we show below that $T'_{\min}$ may be smaller than $T_{\min}$). In analogy to (2.3), consider the split-panel log-likelihood correction

$$\dot{l}(\theta) \equiv \frac{g}{g-1}\widehat{l}(\theta) - \frac{1}{g-1}\bar{l}(\theta), \qquad \bar{l}(\theta) \equiv \frac{1}{m}\sum_{j=1}^{m} \bar{l}_{\mathcal{S}_j}(\theta), \qquad \bar{l}_{\mathcal{S}_j}(\theta) \equiv \sum_{S\in\mathcal{S}_j} \frac{|S|}{T}\widehat{l}_S(\theta),$$

where, as before, $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m\}$ is the equivalence class of a chosen partition $\mathcal{S}$ of the panel into g non-overlapping subpanels (now with $|S| \geq T'_{\min}$ for all $S \in \mathcal{S}$) such that $\min_{S\in\mathcal{S}} |S|/T$ is bounded away from zero as T grows. It is easy to see that $\text{plim}_{N\to\infty}\dot{l}(\theta) = l_0(\theta) + o(T^{-1})$, from which it readily follows that

$$\dot{\theta} \equiv \arg\max_{\theta\in\Theta} \dot{l}(\theta)$$

is a bias-corrected estimator of θ_0.

THEOREM 2.2. *Let Assumptions 2.1, 2.4, and 2.5 hold. Then* $\text{plim}_{N\to\infty}\dot{\theta} = \theta_0 + o(T^{-1})$ *and*

$$\sqrt{NT}(\dot{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1})$$

as $N, T \to \infty$ *with* $N/T \to \rho$.

Thus, $\dot{\theta}$ has the same limit distribution as $\widetilde{\theta}$ under rectangular-array asymptotics. Just as $\widetilde{\theta}$ is a jackknife alternative to the analytical bias correction of Hahn and Kuersteiner (2011), $\dot{\theta}$ is a jackknife alternative to the analytical likelihood correction proposed by Arellano and Hahn (2006). Again, the jackknife estimator estimates the bias term, here $C_1(\theta)/T$, without the need to have an expression for it.

The half-panel likelihood-based jackknife estimator is

$$\dot{\theta}_{1/2} \equiv \arg\max_{\theta\in\Theta} \dot{l}_{1/2}(\theta), \qquad \dot{l}_{1/2}(\theta) \equiv 2\widehat{l}(\theta) - \bar{l}_{1/2}(\theta), \tag{2.8}$$

using obvious notation in analogy to $\widetilde{\theta}_{1/2}$. The motivation for using half-panels is analogous to that in the case of $\widetilde{\theta}_{1/2}$; in the class $\dot{l}(\theta)$, $\dot{l}_{1/2}(\theta)$ minimizes all higher-order bias terms that are not eliminated.

Estimation based on the bias-corrected profile likelihood is computationally somewhat more involved than the simple additive correction $\widetilde{\theta}_{1/2}$ in (2.5). Maximizing $\dot{l}_{1/2}(\theta)$ is equivalent to locating a saddlepoint that involves maximization over θ and the fixed effects implicit in $\widehat{l}(\theta)$, and minimization over two or four separate sets of fixed effects (when T is even or odd, respectively) implicit in $\bar{l}_{1/2}(\theta)$. In our simulations we computed $\dot{\theta}_{1/2}$ using a nested Newton-Raphson algorithm, optimizing over θ in an outer loop and over all sets of fixed effects in an inner loop. We found this to work very reliably and reasonably fast, typically requiring not more than two to three times as much computational time as $\widetilde{\theta}_{1/2}$.

One attractive feature of profile-likelihood corrections is their invariance and equivariance properties. In particular, $\dot{\theta}_{1/2}$ and the associated confidence intervals are equivariant under one-to-one transformations of θ, and the likelihood ratio test is invariant. Corrections of the estimator, such as $\widetilde{\theta}_{1/2}$, do not have these properties.

Another possible advantage of the profile-likelihood correction is that $T'_{\min} \leq T_{\min}$ and, in some models, $T'_{\min} < T_{\min}$. Recall that θ_T maximizes $l_T(\theta)$, so θ_T will not exist when $l_T(\theta)$ does not exist and, therefore,

$T'_{\min} \leq T_{\min}$. An example where $T'_{\min} < T_{\min}$ is the first-order autoregressive binary-choice model. Here, for $T = 2$, $l_T(\theta)$ exists for all θ but is maximized at $-\infty$, so $T'_{\min} = 2$ and $T_{\min} = 3$ (a more detailed derivation is given in the supplementary material).

Finally, bias correction of the profile likelihood extends naturally to unbalanced data, under two conditions: (i) for every unit i the observations form a time series without gaps; (ii) the unbalancedness (for example, attrition) is due to exogenous reasons. Given (i), the unbalanced panel is formed as the union of J independent balanced panels of dimensions $N_j \times T_j$, $j = 1, 2, \ldots, J$. Write $\widehat{l}(\theta; j)$ for the profile log-likelihood for the jth such panel. The profile log-likelihood for the full panel then takes the form of the weighted average

$$\widehat{l}(\theta) = \sum_{j=1}^{J} \omega_j \, \widehat{l}(\theta; j), \qquad \omega_j \equiv \frac{N_j T_j}{\sum_{j=1}^{J} N_j T_j}.$$

Each of the $\widehat{l}(\theta; j)$ may be jackknifed in the usual fashion, giving $\dot{l}(\theta; j)$, say. Now consider asymptotics where, for all $j, j' = 1, 2, \ldots, J$, the ratios $N_j/N_{j'}$ and $T_j/T_{j'}$ remain fixed as $\sum_j N_j$ and $\sum_j T_j$ grow large. It is then immediate that the maximizer of

$$\dot{l}(\theta) \equiv \sum_{j=1}^{J} \omega_j \, \dot{l}(\theta; j), \tag{2.9}$$

will be a bias-corrected estimator of θ_0 that is asymptotically normal and correctly centered provided that $\sum_j N_j / \sum_j T_j \to \rho$.[2] In practical situations, it may occur that some T_j are too small for $\dot{l}(\theta; j)$ to be defined, in which case the corresponding terms have to be dropped from (2.9).

2.3. Discussion

Under our assumptions, all bias-correction estimators remove the leading bias term from $\widehat{\theta}$ and have the same asymptotic distribution as $N, T \to \infty$ with $N/T \to \rho$. Nevertheless, the finite-sample performance of these estimators can be very different, due to the different ways the leading bias is estimated. For the same reason the various methods may react differently to violations of the regularity conditions, in particular to non-stationarity, which we discuss next.

2.3.1. Small-sample comparison Extending Hahn and Newey (2004), Hahn and Kuersteiner (2011) derived the exact expression of B_1/T and gave conditions for consistency of a plug-in estimator. The bias term depends on moments and cross-moments of higher-order derivatives of the likelihood function, evaluated at true parameter values. An estimator can be formed by replacing spectral expectations with sample averages that are truncated via a bandwidth that increases appropriately with T, and replacing θ_0 and the α_{i0} by their maximum-likelihood estimates. Arellano and Hahn (2006) followed a similar strategy in deriving an estimator of $C_1(\theta)/T$, the leading bias of the profile log-likelihood. Just like the jackknife, these ways of estimating the bias introduce statistical noise and alter the remaining higher-order bias. Which of the various approaches delivers the least bias will generally depend on the model at hand and the true parameter values. To gain some insight, we report on the performance of the estimators in simulation experiments. Of course, a Monte Carlo exercise can at best be suggestive. Higher-order expansions of the bias and variance would be needed to

[2] One could also jackknife $\sum_j \omega_j \, \widehat{\theta}(j)$, where $\widehat{\theta}(j)$ is the maximum-likelihood estimator that corresponds to the jth panel. This would yield $\sum_j \omega_j \, \widetilde{\theta}_{1/2}(j)$, say, which is not quite the same as directly jackknifing $\widehat{\theta}$ because, in general, $\widehat{\theta} \neq \sum_j \omega_j \, \widehat{\theta}(j)$. Justifying direct application of the jackknife to $\widehat{\theta}$ would require a proof of a generalized form of the expansion of θ_T in Assumption 2.3.

obtain formal results, as in Pfanzagl and Wefelmeyer (1978) for parametric cross-sectional models. Deriving such expansions is expected to be a difficult task and is left for future research.

The experiment we report on here deals with a dynamic probit model, which we will also use in the empirical illustration below. The design is as follows. The variables (y_{it}, x_{it}) were generated as

$$y_{it} = 1\{\alpha_{i0} + \gamma_0 y_{it-1} + \delta_0 x_{it} \geq \varepsilon_{it}\}, \qquad x_{it} = \eta_{i0} + \pi_0 x_{it-1} + \epsilon_{it},$$

where ε_{it} and ϵ_{it} are independent standard normal. We drew $\alpha_{i0} \sim \mathcal{N}(0,1)$, set $\eta_{i0} = -\sqrt{2/3}\,\alpha_{i0}$ and $\pi_0 = .5$, and generated (y_{i0}, x_{i0}) from their steady-state distribution. We estimate $\theta_0 = (\gamma_0, \delta_0)'$ and report results for $N = 500$, $T = 6, 8, 12, 18$, $\gamma_0 = .5, 1, 1.5$, and $\delta_0 = .5$, in which case the contribution to the variance of y_{it} is the same for α_{i0}, x_{it}, and ε_{it}.

Table 3 below reports the bias, the root mean squared error, the ratio of the estimated standard errors to the standard deviation over the Monte Carlo replications, and the coverage rate of the 95% confidence interval constructed from the Hessian-based estimate of the asymptotic variance. Besides the half-panel jackknife estimators, we considered four analytical bias-correction estimators. The first two of these are the Hahn and Kuersteiner (2011) correction (HK) and the determinant-based version of the Arellano and Hahn (2006) estimator (AH), both implemented with the bandwidth set to one and the latter with a triangular kernel.[3] The two other estimators have been developed especially for the binary-choice model. The first of these, due to Fernández-Val (2009) (F), refines the estimator of the bias of Hahn and Kuersteiner (2011) by using the model structure to replace sample averages by expected quantities. The second, due to Carro (2007) (C), solves a bias-corrected profile-score equation as in Cox and Reid (1987, 1993) (see also Arellano 2003, and Woutersen 2002 for an alternative interpretation). This correction requires recursive calculation of expected likelihood quantities. The use of expected quantities instead of sample averages in the latter two estimators is intuitively attractive. Further, since they use most of the model structure, they may be expected to perform best under correct specification. On the other hand, it is required that these expectations be available in closed form. This is the case in this model, but may not be so in others (see, e.g., Hospido 2012 for such a model).

As is clear from the table, maximum likelihood performs poorly in this model, suffering from substantial bias and confidence intervals with extremely poor coverage. The problem is most severe for the autoregressive parameter, $\widehat{\gamma}$, although the bias is also substantial for $\widehat{\delta}$. The magnitude of the bias is still considerable for large values of T and, all else equal, also increases with the value of γ_0. This is because more state dependence leads to less informative data. All bias-correction approaches considered deliver point estimates with lower bias. In most cases, the reduction in bias is quite substantial, and so is the reduction in root mean squared error. Bias correction also leads to improvements in the coverage rates of the confidence intervals, and so to improved inference. For most design points, $\widetilde{\theta}_{1/2}$ and $\dot{\theta}_{1/2}$ have smaller bias than $\widetilde{\theta}_{\mathrm{HK}}$ and $\widetilde{\theta}_{\mathrm{AH}}$, respectively, although the difference is less pronounced in the latter case. The confidence intervals based on $\widetilde{\theta}_{1/2}$ and $\dot{\theta}_{1/2}$ are also better than those based on $\widetilde{\theta}_{\mathrm{HK}}$ and $\widetilde{\theta}_{\mathrm{AH}}$, respectively. The chief reason for this is their success at removing bias. The plug-in estimator of the asymptotic variance provides a reasonably accurate estimate of the estimators' true variability for most design points. The simulation results further show that replacing sample averages by expectations in the analytical bias-correction methods yields a considerable improvement, as is apparent on comparing $\widetilde{\theta}_{\mathrm{F}}$ with $\widetilde{\theta}_{\mathrm{HK}}$, and $\widetilde{\theta}_{\mathrm{C}}$ with $\widetilde{\theta}_{\mathrm{AH}}$. As the state dependence increases,

[3] The bandwidth is required to grow with T to ensure asymptotic bias reduction. We repeated the experiment with several other choices for the bandwidth. The current choice was found to perform best. Setting the bandwidth too large resulted in estimates with bias of the same order as that of maximum likelihood.

the performance of most estimators of γ_0 worsens, with little bias reduction and hardly improved confidence intervals when $\gamma_0 = 1.5$. Only $\widetilde{\gamma}_{1/2}$ is less sensitive to the value of γ_0, still achieving a substantial bias reduction when the persistence is large.

From this and many other numerical experiments that we conducted, our tentative conclusion is that the jackknife corrections are competitive with the available analytical corrections, and can be a very useful tool for inference in micropanels.

2.3.2. Robustness to non-stationarity The available literature on bias correction in general nonlinear fixed-effect models assumes stationary data. Dealing with potentially non-stationary regressors, trends, or other time effects is complicated when the length of the panel is not treated as fixed. In nonlinear models, a major difficulty is that the maximum-likelihood estimator itself may exhibit non-standard behavior, including a non-standard convergence rate in T and a non-normal limit distribution. In such cases, it is doubtful that the expansions in Assumptions 2.3 or 2.5 will hold. In addition, even in situations where these expansions continue to hold, there may be a concern that the jackknife corrections are potentially more sensitive to violations of the stationarity requirement than the analytical methods because of the necessity to split the panel. For example, when the dynamics of the data are very different in the two half-panels, this could result in half-panel estimates that are very different from each other and lead to a poor estimate of the leading bias.

To infer whether the jackknife estimators yield asymptotically bias-reduced estimates, possibly in non-stationary situations, one can devise validity tests based on comparing subpanel estimates. Let $\mathcal{S} = \{S_1, S_2\}$ partition $\{1, 2, \ldots, T\}$ such that $|S_1| \geq T_{\min}$ and $|S_2| \geq T_{\min}$. Then, using (2.1), we have

$$\frac{|S_1|}{|S_2|}(\widehat{\theta}_{S_1} - \widehat{\theta}) \xrightarrow{p} \frac{B_1}{T} + o\left(\frac{1}{T}\right), \qquad \frac{|S_2|}{|S_1|}(\widehat{\theta}_{S_2} - \widehat{\theta}) \xrightarrow{p} \frac{B_1}{T} + o\left(\frac{1}{T}\right),$$

under the null that the split-panel jackknife estimator based on $\mathcal{S}$ is bias-reducing. It is intuitively clear that a comparison of two subpanel estimates can be informative about the validity of the jackknife corrections. Letting

$$\widehat{r} \equiv \frac{|S_1|}{|S_2|}(\widehat{\theta}_{S_1} - \widehat{\theta}) - \frac{|S_2|}{|S_1|}(\widehat{\theta}_{S_2} - \widehat{\theta}),$$

we can form a Wald test statistic that is asymptotically χ^2 distributed under our assumptions, i.e.,

$$\tilde{t} \equiv \frac{NT}{d}\,\widehat{r}'\,\widehat{\Sigma}\,\widehat{r} \xrightarrow{d} \chi^2_{\dim\theta}, \qquad d \equiv \frac{|S_1|}{|S_2|} + \frac{|S_2|}{|S_1|} + 2. \tag{2.10}$$

The scale factor d accounts for the variance inflation due to the use of subpanels. For example, when T is even, the Wald statistic associated with the half-panel jackknife has $d = 4$.

In the same way, now with $|S_1| \geq T'_{\min}$ and $|S_2| \geq T'_{\min}$, if the expansion in Assumption 2.5 holds for some function $C_1(\theta)$, we have

$$\frac{|S_1|}{|S_2|}(\widehat{s}_{S_1}(\theta) - \widehat{s}(\theta)) \xrightarrow{p} \frac{C_1'(\theta)}{T} + o\left(\frac{1}{T}\right), \qquad \frac{|S_2|}{|S_1|}(\widehat{s}_{S_2}(\theta) - \widehat{s}(\theta)) \xrightarrow{p} \frac{C_1'(\theta)}{T} + o\left(\frac{1}{T}\right),$$

for $\theta \in \mathcal{N}_0$. From this we can form a score test to check the validity of the likelihood-based jackknife correction. A natural value to evaluate the profile scores is the maximum-likelihood estimate of the full panel. Letting

$$\dot{r} \equiv \frac{|S_1|}{|S_2|}\,\widehat{s}_{S_1}(\widehat{\theta}) - \frac{|S_2|}{|S_1|}\,\widehat{s}_{S_2}(\widehat{\theta}),$$

Table 3. Simulation results for a stationary dynamic probit model

					bias								rmse			
T	γ_0	$\widehat{\gamma}$	$\widetilde{\gamma}_{1/2}$	$\widetilde{\gamma}_{\mathrm{HK}}$	$\widetilde{\gamma}_{\mathrm{F}}$	$\dot{\gamma}_{1/2}$	$\dot{\gamma}_{\mathrm{AH}}$	$\dot{\gamma}_{\mathrm{C}}$	$\widehat{\gamma}$	$\widetilde{\gamma}_{1/2}$	$\widetilde{\gamma}_{\mathrm{HK}}$	$\widetilde{\gamma}_{\mathrm{F}}$	$\dot{\gamma}_{1/2}$	$\dot{\gamma}_{\mathrm{AH}}$	$\dot{\gamma}_{\mathrm{C}}$
6	.5	−.531	.315	−.194	−.065	−.218	−.230	−.129	.535	.330	.202	.084	.226	.237	.142
8	.5	−.380	.124	−.119	−.046	−.117	−.144	−.069	.384	.140	.128	.065	.127	.152	.085
12	.5	−.243	.046	−.063	−.032	−.045	−.077	−.028	.246	.065	.073	.048	.061	.086	.048
18	.5	−.158	.019	−.037	−.023	−.017	−.044	−.011	.161	.039	.047	.037	.037	.053	.033
6	1	−.600	.230	−.313	−.197	−.323	−.330	−.209	.605	.255	.319	.205	.331	.337	.219
8	1	−.442	.075	−.219	−.150	−.194	−.236	−.124	.445	.106	.225	.158	.203	.242	.136
12	1	−.288	.026	−.134	−.101	−.085	−.146	−.055	.291	.059	.140	.108	.097	.152	.071
18	1	−.188	.015	−.083	−.068	−.032	−.090	−.022	.191	.042	.089	.075	.049	.096	.042
6	1.5	−.731	.083	−.527	−.392	−.486	−.477	−.355	.737	.164	.532	.398	.494	.490	.364
8	1.5	−.560	−.031	−.400	−.314	−.330	−.381	−.238	.565	.101	.405	.320	.337	.387	.247
12	1.5	−.384	−.038	−.268	−.223	−.177	−.266	−.128	.388	.076	.272	.227	.185	.270	.138
18	1.5	−.260	−.018	−.177	−.153	−.085	−.180	−.063	.264	.052	.181	.158	.096	.184	.077
					se/sd								confidence			
T	γ_0	$\widehat{\gamma}$	$\widetilde{\gamma}_{1/2}$	$\widetilde{\gamma}_{\mathrm{HK}}$	$\widetilde{\gamma}_{\mathrm{F}}$	$\dot{\gamma}_{1/2}$	$\dot{\gamma}_{\mathrm{AH}}$	$\dot{\gamma}_{\mathrm{C}}$	$\widehat{\gamma}$	$\widetilde{\gamma}_{1/2}$	$\widetilde{\gamma}_{\mathrm{HK}}$	$\widetilde{\gamma}_{\mathrm{F}}$	$\dot{\gamma}_{1/2}$	$\dot{\gamma}_{\mathrm{AH}}$	$\dot{\gamma}_{\mathrm{C}}$
6	.5	.987	.952	1.084	1.148	1.520	1.051	1.039	.000	.082	.103	.849	.290	.034	.446
8	.5	.995	.989	1.082	1.115	1.262	1.074	1.015	.000	.521	.334	.880	.578	.171	.726
12	.5	1.017	.984	1.084	1.098	1.110	1.082	1.018	.000	.822	.645	.894	.852	.497	.894
18	.5	1.006	.982	1.054	1.060	1.037	1.053	1.000	.001	.911	.798	.902	.929	.723	.934
6	1	1.010	.980	1.160	1.224	1.577	1.081	1.097	.000	.446	.002	.173	.058	.003	.156
8	1	1.015	1.002	1.143	1.180	1.324	1.115	1.056	.000	.837	.021	.228	.204	.013	.429
12	1	1.011	.981	1.108	1.124	1.136	1.099	1.017	.000	.917	.122	.370	.633	.078	.765
18	1	1.012	.985	1.082	1.087	1.058	1.078	1.004	.001	.932	.332	.520	.880	.259	.901
6	1.5	1.016	1.014	1.256	1.302	1.624	.870	1.171	.000	.919	.000	.002	.015	.011	.021
8	1.5	1.032	1.024	1.227	1.263	1.384	1.131	1.122	.000	.944	.000	.003	.034	.001	.095
12	1.5	1.040	1.025	1.190	1.206	1.218	1.153	1.072	.000	.915	.000	.009	.207	.001	.383
18	1.5	1.013	.998	1.121	1.128	1.095	1.104	1.017	.000	.935	.009	.045	.600	.009	.706
					bias								rmse			
T	γ_0	$\widehat{\delta}$	$\widetilde{\delta}_{1/2}$	$\widetilde{\delta}_{\mathrm{HK}}$	$\widetilde{\delta}_{\mathrm{F}}$	$\dot{\delta}_{1/2}$	$\dot{\delta}_{\mathrm{AH}}$	$\dot{\delta}_{\mathrm{C}}$	$\widehat{\delta}$	$\widetilde{\delta}_{1/2}$	$\widetilde{\delta}_{\mathrm{HK}}$	$\widetilde{\delta}_{\mathrm{F}}$	$\dot{\delta}_{1/2}$	$\dot{\delta}_{\mathrm{AH}}$	$\dot{\delta}_{\mathrm{C}}$
6	.5	.153	−.076	.058	.015	.078	.105	.040	.159	.097	.069	.036	.087	.113	.052
8	.5	.109	−.035	.039	.010	.045	.061	.022	.114	.052	.048	.028	.054	.068	.034
12	.5	.069	−.014	.020	.006	.019	.029	.009	.073	.028	.029	.021	.029	.036	.022
18	.5	.045	−.006	.010	.003	.008	.014	.004	.048	.018	.019	.016	.018	.021	.016
6	1	.182	−.055	.037	.034	.111	.139	.062	.189	.089	.052	.050	.120	.147	.073
8	1	.133	−.023	.033	.025	.069	.087	.038	.138	.050	.045	.039	.078	.094	.048
12	1	.085	−.011	.023	.015	.033	.044	.017	.089	.030	.033	.027	.041	.050	.028
18	1	.056	−.006	.015	.009	.014	.023	.008	.059	.021	.023	.019	.023	.029	.019
6	1.5	.228	−.024	−.034	.061	.158	.195	.095	.236	.097	.049	.075	.169	.215	.106
8	1.5	.171	.000	−.005	.048	.107	.127	.064	.178	.060	.032	.061	.116	.135	.075
12	1.5	.116	.005	.014	.034	.060	.073	.036	.120	.037	.030	.043	.068	.079	.045
18	1.5	.077	.001	.017	.022	.031	.041	.019	.081	.025	.027	.030	.038	.047	.028
					se/sd								confidence			
T	γ_0	$\widehat{\delta}$	$\widetilde{\delta}_{1/2}$	$\widetilde{\delta}_{\mathrm{HK}}$	$\widetilde{\delta}_{\mathrm{F}}$	$\dot{\delta}_{1/2}$	$\dot{\delta}_{\mathrm{AH}}$	$\dot{\delta}_{\mathrm{C}}$	$\widehat{\delta}$	$\widetilde{\delta}_{1/2}$	$\widetilde{\delta}_{\mathrm{HK}}$	$\widetilde{\delta}_{\mathrm{F}}$	$\dot{\delta}_{1/2}$	$\dot{\delta}_{\mathrm{AH}}$	$\dot{\delta}_{\mathrm{C}}$
6	.5	.843	.946	.954	1.052	1.420	.845	1.018	.019	.718	.608	.946	.803	.175	.797
8	.5	.877	1.018	.966	1.032	1.272	.921	1.018	.037	.852	.713	.944	.852	.402	.887
12	.5	.918	1.053	.985	1.022	1.167	.968	1.021	.098	.922	.842	.947	.925	.710	.936
18	.5	.946	1.055	1.001	1.017	1.105	.992	1.018	.215	.946	.912	.952	.950	.864	.952
6	1	.845	.929	1.064	1.087	1.424	.842	1.039	.013	.852	.866	.889	.653	.083	.671
8	1	.867	.999	1.022	1.047	1.282	.912	1.025	.023	.920	.823	.889	.714	.220	.787
12	1	.904	1.035	1.002	1.026	1.175	.952	1.023	.059	.940	.833	.912	.845	.525	.895
18	1	.935	1.038	1.001	1.017	1.109	.980	1.021	.133	.946	.871	.926	.920	.752	.935
6	1.5	.838	.904	1.324	1.117	1.427	.680	1.049	.012	.916	.948	.811	.558	.059	.545
8	1.5	.852	.968	1.180	1.057	1.279	.856	1.027	.016	.940	.977	.788	.559	.123	.643
12	1.5	.890	1.015	1.083	1.029	1.185	.934	1.028	.028	.953	.941	.798	.676	.285	.777
18	1.5	.914	1.034	1.023	1.010	1.126	.953	1.018	.066	.956	.889	.839	.819	.524	.874

Model: $y_{it} = 1\{\alpha_{i0} + \gamma_0 y_{it-1} + \delta_0 x_{it} \geq \varepsilon_{it}\}$, $\varepsilon_{it} \sim \mathcal{N}(0,1)$, stationary (y_{i0}, x_{i0}). Data generated with $N = 500$, $\alpha_{i0} \sim \mathcal{N}(0,1)$, $\delta_0 = .5$, $x_{it} = -\sqrt{2/3}\,\alpha_{i0} + .5x_{it-1} + \epsilon_{it}$, $\epsilon_{it} \sim \mathcal{N}(0,1)$. 10,000 Monte Carlo replications.

it follows under our assumptions that

$$\dot{t} \equiv \frac{NT}{d} \dot{r}\, \widehat{\Sigma}^{-1}\, \dot{r} \xrightarrow{d} \chi^2_{\dim\theta}, \tag{2.11}$$

with the same d as above. When θ_0 is multidimensional it may also be of interest to report component-by-component test statistics.

Let $\widetilde{t}_{1/2}$ and $\dot{t}_{1/2}$ denote the statistics $\widetilde{t}$ and $\dot{t}$ implemented with half-panels. The empirical acceptance rates of the 5%-level validity tests based on $\widetilde{t}_{1/2}$ and $\dot{t}_{1/2}$ were reported in Tables 1 and 2 for the linear autoregressive model and the dynamic probit model. There, the individual time-series processes were indeed stationary, and the empirical acceptance rates are close to the nominal acceptance probability of 95%. For small T, there is some size distortion but it diminishes as T grows.

One realistic departure from Assumption 2.1 is a situation in which the initial observations are not drawn from their respective steady-state distributions. The fixed-T inconsistency of $\widehat{\theta}$ will, in general, depend on the distribution of the initial values, but the processes will still be asymptotically stationary as $T \to \infty$. It is conceivable that this distribution affects the $O(T^{-1})$ bias term (assuming that the leading bias still takes this form), in which case the half-panel jackknife will fail to remove it. This is a potential weakness of the jackknife that the analytical plug-in methods need not share.[4] The test statistics $\widetilde{t}_{1/2}$ and $\dot{t}_{1/2}$ may help to assess the effect of non-stationary initial observations on the jackknife. However, in the event that the jackknife is still bias-reducing, it is natural to expect that the tests will exhibit size distortions that increase with the degree of non-stationarity, although the size distortions should vanish as T increases. Thus, some caution is warranted when the tests are applied in very short panels. To gain some insight in the performance of these tests, we now examine the Gaussian autoregression and the autoregressive probit model in the presence of non-stationary initial observations.

Reconsider the Gaussian autoregression

$$y_{it} = \alpha_{i0} + \gamma_0 y_{it-1} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathcal{N}(0, \sigma_0^2),$$

now with arbitrary initial observations y_{i0}. It is well known that $\gamma_T - \gamma_0$ depends on the joint distribution of (α_{i0}, y_{i0}). However, the first-order bias does not (Hahn and Kuersteiner 2002). In the supplementary material we show that

$$\gamma_T - \gamma_0 = -\frac{1+\gamma_0}{T} - \frac{\gamma_0(1+\gamma_0) + (1-\psi^2)}{(1-\gamma_0)T^2} + O\left(\frac{1}{T^3}\right), \qquad \psi^2 \equiv \overline{\mathbb{E}}\left[\left(y_{i0} - \frac{\alpha_{i0}}{1-\gamma_0}\right)^2 \Big/ \frac{\sigma_0^2}{1-\gamma_0^2}\right].$$

The parameter ψ^2 is a measure of the deviation of the y_{i0} from their stationary distributions, with stationarity implying $\psi^2 = 1$. Because ψ^2 does not show up in the $O(T^{-1})$ bias term, the jackknife will be bias-reducing for arbitrary initial observations. The presence of ψ^2 in the second-order bias term arises from a higher-order expansion of the large N variance of $\widehat{\gamma}$ as $T \to \infty$. This variance appears as the denominator of the fixed-T inconsistency of $\widehat{\gamma}$. With the effect of the initial observations fading out as $T \to \infty$, the asymptotic variance of $\widehat{\gamma}$ under rectangular-array asymptotics is $1-\gamma_0^2$, independently of ψ^2. Similar results may be derived when the model is extended to allow for (incidental) time trends or time-series heteroskedasticity (see Alvarez and Arellano 2004 and Dhaene and Jochmans 2013). The robustness of the jackknife to non-stationary initial observations also holds for the jackknifed profile log-likelihood. Non-stationary initial observations have no effect on the $O(T^{-1})$ bias term of $\widehat{l}(\gamma)$, so the jackknife is bias-reducing (see the supplementary material for

[4] Verifying whether the analytical corrections are immune to non-stationary initial observations would require a proof that the plug-in estimator of the leading bias remains consistent. No general results on this are known to us.

details). One may also work with the profile log-likelihood $\widehat{l}(\gamma, \sigma^2)$, whose $O(T^{-1})$ bias term is, again, free of ψ^2. We found, however, that additionally profiling out σ^2 before jackknifing performs better in terms of bias reduction. The results for $\dot{\gamma}_{1/2}$ presented in Table 4 below and earlier in Figure 1 and Table 1 are based on jackknifing $\widehat{l}(\gamma)$.

Table 4 presents simulation results for the Gaussian autoregression with non-stationary initial observations, where the jackknife is bias-reducing. We generated $y_{i0} \sim \mathcal{N}(\alpha_{i0}/(1-\gamma_0), \psi^2\sigma_0^2/(1-\gamma_0^2))$ with ψ set to 0 and 2. These values correspond, respectively, to inlying and outlying initial observations relative to the steady-state distributions. The results show that the bias-corrected estimators continue to remove most of the small-sample bias from $\widehat{\gamma}$. The jackknife estimator $\widetilde{\gamma}_{1/2}$ generally performs better than the plug-in estimator $\widetilde{\gamma}_{\mathrm{HK}} = \widehat{\gamma} + (1 + \widehat{\gamma})/T$. When $\gamma_0 = .5$, the 5%-level validity tests both overreject the null when T is small, but the overrejection rates decrease as T increases, as predicted by the theory. This is because in the early periods the time-series are moving toward their steady state. This move is bigger as $|\psi|$ is farther away from 1. The impact of ψ vanishes as $\gamma_0 \to 1$ (Dhaene and Jochmans 2013), which explains the much improved acceptance rates for very small T when γ_0 is increased to .95.

Table 4. Small-sample performance in a non-stationary Gaussian autoregression

			bias				confidence				validity	
T	γ_0	ψ	$\widehat{\gamma}$	$\widetilde{\gamma}_{\mathrm{HK}}$	$\widetilde{\gamma}_{1/2}$	$\dot{\gamma}_{1/2}$	$\widehat{\gamma}$	$\widetilde{\gamma}_{\mathrm{HK}}$	$\widetilde{\gamma}_{1/2}$	$\dot{\gamma}_{1/2}$	$\tilde{t}_{1/2}$	$\dot{t}_{1/2}$
4	.5	0	−.537	−.296	−.191	−.239	.000	.022	.304	.090	.601	.458
6	.5	0	−.340	−.147	−.054	−.121	.000	.233	.726	.373	.639	.688
8	.5	0	−.243	−.086	−.012	−.070	.000	.507	.834	.626	.737	.818
12	.5	0	−.151	−.039	.007	−.031	.000	.778	.866	.832	.855	.904
4	.5	2	−.244	.070	.084	−.099	.001	.747	.681	.687	.376	.480
6	.5	2	−.178	.043	.064	−.059	.001	.798	.662	.769	.304	.593
8	.5	2	−.142	.028	.044	−.039	.002	.854	.711	.836	.373	.691
12	.5	2	−.102	.014	.023	−.020	.013	.907	.808	.895	.585	.809
4	.95	0	−.609	−.274	−.220	−.405	.000	.023	.220	.000	.950	.746
6	.95	0	−.441	−.189	−.128	−.290	.000	.016	.332	.000	.945	.870
8	.95	0	−.346	−.146	−.088	−.225	.000	.014	.419	.000	.934	.915
12	.95	0	−.243	−.101	−.051	−.154	.000	.014	.520	.000	.922	.940
4	.95	2	−.511	−.152	−.111	−.330	.000	.370	.620	.001	.947	.729
6	.95	2	−.347	−.079	−.025	−.219	.000	.513	.824	.001	.928	.865
8	.95	2	−.257	−.046	.008	−.159	.000	.660	.850	.002	.909	.896
12	.95	2	−.166	−.017	.028	−.098	.000	.809	.717	.008	.875	.927

Model: $y_{it} = \alpha_{i0} + \gamma_0 y_{it-1} + \varepsilon_{it}$, $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_0^2)$. Data generated with $N = 100$, $\sigma_0^2 = 1$, $\alpha_{i0} \sim \mathcal{N}(0, 1)$, $y_{i0} \sim \mathcal{N}(\alpha_{i0}/(1-\gamma_0), \psi^2\sigma_0^2/(1-\gamma_0^2))$. 10,000 Monte Carlo replications.

In the autoregressive probit model with non-stationary initial observations there are no theoretical results available about the expansions. We approached the question by simulation. Table 5 reports the effect of setting $y_{i0} = 0$ for all i (top panel) and setting $y_{i0} = 1$ for all i (bottom panel), respectively. These are two extreme deviations from stationary initial observations. The bias reduction of the jackknife is manifest. In line with this, the validity tests have acceptance rates close to the nominal rate, even for very short panels. The improved acceptance rates for very small T, compared with those in the linear autoregressive model, are likely to be due to the limited variation in the regressor. The results suggest that non-stationary initial observations in the binary-choice model do not pose problems for bias correction.

We note that flexible modelling can be a way to accommodate certain trends in the data such as increases in cross-sectional variances. For example, when investigating the dynamics of individual earnings, Hospido (2012) allows for worker-specific volatility clustering by specifying a GARCH model for the conditional variance of wages (see also Meghir and Pistaferri 2004). Such a multiple-equation model readily fits into our setup and can easily be estimated via the jackknife.

Table 5. Small-sample performance in a non-stationary autoregressive probit model

	bias			confidence			validity	
T	$\widehat{\theta}$	$\widetilde{\theta}_{1/2}$	$\dot{\theta}_{1/2}$	$\widehat{\theta}$	$\widetilde{\theta}_{1/2}$	$\dot{\theta}_{1/2}$	$\tilde{t}_{1/2}$	$\dot{t}_{1/2}$
				$y_{i0} = 0$				
6	$-.525$	.305	$-.213$	.083	.740	.936	.910	.906
8	$-.394$	.119	$-.126$	.143	.886	.928	.921	.937
12	$-.268$	.038	$-.061$	.259	.930	.944	.936	.948
18	$-.183$	.013	$-.029$	.404	.943	.945	.945	.952
				$y_{i0} = 1$				
6	$-.569$	.273	$-.242$	.054	.791	.914	.945	.921
8	$-.423$	.099	$-.142$	.112	.904	.912	.953	.952
12	$-.282$	.030	$-.066$	.233	.936	.933	.952	.954
18	$-.191$	.008	$-.032$	.375	.940	.944	.951	.953

Model: $y_{it} = 1(\alpha_{i0} + \theta_0 y_{it-1} + \varepsilon_{it} > 0)$, $\varepsilon_{it} \sim \mathcal{N}(0,1)$. Data generated with $N = 100$, $\theta_0 = .5$, $\alpha_{i0} \sim \mathcal{N}(0,1)$. $10,000$ Monte Carlo replications.

We end our discussion on non-stationarity by comparing the various bias-correction estimators in the dynamic logit specification of Honoré and Kyriazidou (2000); see also Carro (2007) and Fernández-Val (2009). The data are generated as

$$y_{it} = 1\{\alpha_{i0} + \gamma_0 y_{it-1} + \delta_0 x_{it} \geq \varepsilon_{it}\}, \qquad x_{it} \sim \mathcal{N}(0, \pi^2/3),$$

with ε_{it} logistically distributed and $\delta_0 = 1$. The initial observations are drawn as $x_{i0} \sim \mathcal{N}(0, \pi^2/3)$ and $y_{i0} = 1\{\alpha_{i0} + \delta_0 x_{i0} \geq \varepsilon_{i0}\}$, and the fixed effects are set to $\alpha_{i0} = \frac{1}{4}(x_{i0} + x_{i1} + x_{i2} + x_{i3})$. This design is non-stationary because the (x_{i0}, y_{i0}) are not drawn from the steady-state distributions and, also, because the dependence between the covariate and the fixed effect changes abruptly in the fourth period: the correlation between x_{it} and α_{i0} equals $1/4$ for $t \leq 3$, while α_{i0} and x_{it} are independent once $t > 3$. Table 6 provides simulation results for $N = 500$ and various values of γ_0.

The results are qualitatively similar to those for the probit model reported on above. Again, maximum likelihood is heavily biased and all other estimators reduce this bias, in most cases quite substantially. The non-stationarity has an adverse effect on the jackknife estimator applied directly to the maximum-likelihood estimator for γ_0 when $T = 6$, with only a moderate reduction in bias. Indeed, when $T = 6$, the half-panel estimates would be expected to differ the most from each other, due to the different form of dependence between α_{i0} and x_{it} in the two half-panels. Beyond this, both jackknife corrections tend to perform well compared with the analytical corrections of Hahn and Kuersteiner (2011) and Arellano and Hahn (2006). The model-specific corrections of Fernández-Val (2009) and Carro (2007) again improve on the general analytical corrections. The estimator of Carro (2007), in particular, yields confidence intervals with very good coverage in this design.

Table 6. Simulation results for the Honoré and Kyriazidou (2000) design

		bias							rmse						
T	γ_0	$\widehat{\gamma}$	$\widetilde{\gamma}_{1/2}$	$\widetilde{\gamma}_{\mathrm{HK}}$	$\widetilde{\gamma}_{\mathrm{F}}$	$\dot{\gamma}_{1/2}$	$\dot{\gamma}_{\mathrm{AH}}$	$\dot{\gamma}_{\mathrm{C}}$	$\widehat{\gamma}$	$\widetilde{\gamma}_{1/2}$	$\widetilde{\gamma}_{\mathrm{HK}}$	$\widetilde{\gamma}_{\mathrm{F}}$	$\dot{\gamma}_{1/2}$	$\dot{\gamma}_{\mathrm{AH}}$	$\dot{\gamma}_{\mathrm{C}}$
6	.5	−.905	.747	−.191	−.079	−.327	−.325	−.140	.917	.784	.222	.138	.355	.352	.185
8	.5	−.634	.287	−.127	−.047	−.192	−.193	−.075	.645	.319	.160	.107	.220	.219	.125
12	.5	−.391	.100	−.057	−.022	−.076	−.090	−.027	.400	.134	.094	.077	.111	.118	.082
18	.5	−.249	.038	−.028	−.014	−.032	−.045	−.010	.257	.077	.066	.061	.071	.075	.063
6	1	−.850	.696	−.298	−.164	−.338	−.330	−.181	.863	.736	.318	.199	.365	.357	.218
8	1	−.602	.244	−.187	−.103	−.204	−.213	−.094	.613	.282	.211	.141	.232	.237	.139
12	1	−.377	.077	−.094	−.059	−.087	−.115	−.036	.387	.121	.122	.096	.121	.139	.088
18	1	−.241	.030	−.052	−.038	−.038	−.065	−.014	.250	.075	.080	.071	.075	.089	.064
6	2	−.761	.613	−.636	−.369	−.367	−.356	−.294	.782	.668	.649	.389	.402	.391	.324
8	2	−.563	.175	−.392	−.255	−.242	−.282	−.166	.579	.240	.407	.276	.274	.307	.202
12	2	−.369	.039	−.212	−.159	−.112	−.189	−.070	.382	.115	.229	.181	.150	.209	.115
18	2	−.241	.016	−.123	−.103	−.049	−.122	−.028	.253	.082	.142	.125	.093	.141	.079

		se/sd							confidence						
T	γ_0	$\widehat{\gamma}$	$\widetilde{\gamma}_{1/2}$	$\widetilde{\gamma}_{\mathrm{HK}}$	$\widetilde{\gamma}_{\mathrm{F}}$	$\dot{\gamma}_{1/2}$	$\dot{\gamma}_{\mathrm{AH}}$	$\dot{\gamma}_{\mathrm{C}}$	$\widehat{\gamma}$	$\widetilde{\gamma}_{1/2}$	$\widetilde{\gamma}_{\mathrm{HK}}$	$\widetilde{\gamma}_{\mathrm{F}}$	$\dot{\gamma}_{1/2}$	$\dot{\gamma}_{\mathrm{AH}}$	$\dot{\gamma}_{\mathrm{C}}$
6	.5	.920	.934	1.068	1.077	1.629	.958	1.030	.000	.083	.654	.918	.790	.308	.801
8	.5	.944	1.039	1.039	1.054	1.351	1.002	1.007	.000	.500	.771	.937	.811	.540	.890
12	.5	.983	1.076	1.047	1.058	1.187	1.038	1.011	.002	.835	.898	.954	.919	.802	.938
18	.5	.980	1.029	1.026	1.031	1.079	1.023	.995	.025	.921	.931	.950	.942	.893	.944
6	1	.944	.939	1.109	1.138	1.622	1.003	1.081	.000	.133	.315	.780	.770	.326	.727
8	1	.957	1.045	1.071	1.095	1.342	1.034	1.035	.000	.632	.571	.857	.779	.493	.864
12	1	.973	1.052	1.055	1.065	1.160	1.044	1.013	.008	.894	.806	.901	.889	.732	.928
18	1	1.000	1.044	1.057	1.063	1.096	1.053	1.021	.043	.939	.884	.923	.941	.842	.946
6	2	.919	.943	1.039	1.216	1.509	.985	1.124	.010	.313	.004	.277	.752	.403	.509
8	2	.946	1.032	1.079	1.153	1.298	1.040	1.078	.015	.835	.081	.441	.743	.385	.740
12	2	.970	1.047	1.075	1.099	1.137	1.057	1.040	.043	.948	.373	.616	.853	.493	.894
18	2	.974	1.027	1.051	1.059	1.041	1.044	1.014	.116	.955	.618	.721	.913	.628	.932

		bias							rmse						
T	γ_0	$\widehat{\delta}$	$\widetilde{\delta}_{1/2}$	$\widetilde{\delta}_{\mathrm{HK}}$	$\widetilde{\delta}_{\mathrm{F}}$	$\dot{\delta}_{1/2}$	$\dot{\delta}_{\mathrm{AH}}$	$\dot{\delta}_{\mathrm{C}}$	$\widehat{\delta}$	$\widetilde{\delta}_{1/2}$	$\widetilde{\delta}_{\mathrm{HK}}$	$\widetilde{\delta}_{\mathrm{F}}$	$\dot{\delta}_{1/2}$	$\dot{\delta}_{\mathrm{AH}}$	$\dot{\delta}_{\mathrm{C}}$
6	.5	.317	−.142	−.091	−.026	.134	.199	.020	.326	.174	.110	.048	.149	.211	.051
8	.5	.217	−.114	−.001	−.001	.060	.105	.015	.223	.129	.042	.037	.076	.115	.041
12	.5	.131	−.055	.015	.003	.017	.044	.008	.135	.065	.035	.030	.037	.055	.031
18	.5	.080	−.023	.008	.002	.002	.018	.003	.085	.034	.026	.024	.024	.031	.024
6	1	.319	−.133	−.133	−.020	.144	.204	.022	.328	.169	.146	.046	.159	.216	.052
8	1	.219	−.106	−.019	.000	.068	.109	.016	.225	.122	.046	.038	.084	.119	.043
12	1	.133	−.051	.010	.004	.021	.046	.008	.138	.062	.033	.031	.040	.057	.032
18	1	.082	−.022	.008	.002	.004	.020	.004	.087	.034	.026	.025	.026	.032	.025
6	2	.325	−.111	−.241	−.018	.167	.215	.019	.335	.161	.250	.048	.184	.229	.054
8	2	.229	−.086	−.071	.002	.091	.120	.017	.237	.111	.083	.041	.106	.131	.046
12	2	.142	−.043	−.007	.005	.037	.053	.010	.148	.059	.034	.033	.054	.065	.035
18	2	.090	−.019	.004	.003	.014	.024	.005	.095	.034	.027	.026	.033	.036	.027

		se/sd							confidence						
T	γ_0	$\widehat{\delta}$	$\widetilde{\delta}_{1/2}$	$\widetilde{\delta}_{\mathrm{HK}}$	$\widetilde{\delta}_{\mathrm{F}}$	$\dot{\delta}_{1/2}$	$\dot{\delta}_{\mathrm{AH}}$	$\dot{\delta}_{\mathrm{C}}$	$\widehat{\delta}$	$\widetilde{\delta}_{1/2}$	$\widetilde{\delta}_{\mathrm{HK}}$	$\widetilde{\delta}_{\mathrm{F}}$	$\dot{\delta}_{1/2}$	$\dot{\delta}_{\mathrm{AH}}$	$\dot{\delta}_{\mathrm{C}}$
6	.5	.805	.914	.710	1.124	1.369	.780	1.009	.001	.649	.466	.930	.783	.061	.940
8	.5	.873	1.084	.911	1.030	1.339	.880	1.007	.003	.566	.929	.958	.923	.300	.943
12	.5	.919	1.179	.965	1.002	1.265	.946	.999	.024	.756	.920	.950	.972	.707	.947
18	.5	.948	1.139	.984	.999	1.179	.975	.998	.111	.900	.935	.951	.977	.882	.950
6	1	.818	.914	.703	1.147	1.388	.791	1.035	.001	.700	.231	.951	.757	.067	.944
8	1	.873	1.093	.928	1.041	1.332	.884	1.017	.003	.644	.895	.960	.904	.301	.944
12	1	.910	1.181	.966	1.002	1.231	.942	.997	.027	.802	.934	.949	.959	.700	.943
18	1	.943	1.137	.983	.998	1.160	.974	.998	.113	.904	.940	.951	.971	.882	.950
6	2	.810	.895	.642	1.166	1.364	.774	1.055	.002	.800	.011	.963	.722	.093	.954
8	2	.862	1.041	.955	1.058	1.310	.869	1.031	.006	.782	.570	.963	.843	.301	.946
12	2	.905	1.150	.993	1.016	1.185	.938	1.012	.033	.877	.940	.955	.928	.680	.947
18	2	.941	1.150	.996	1.010	1.085	.977	1.010	.102	.938	.948	.953	.954	.860	.950

Model: $y_{it} = 1\{\alpha_{i0} + \gamma_0 y_{it-1} + \delta_0 x_{it} \geq \varepsilon_{it}\}$, ε_{it} logistically distributed. Data generated with $N = 500$, $\delta_0 = 1$, $x_{it} \sim \mathcal{N}(0, \pi^2/3)$ $(t = 0, 1, \ldots, T)$, $y_{i0} = 1\{\alpha_{i0} + \delta_0 x_{i0} \geq \varepsilon_{i0}\}$, $\alpha_{i0} = (x_{i0} + x_{i1} + x_{i2} + x_{i3})/4$. 10,000 Monte Carlo replications.

2.4. Correcting average effects

The split-panel jackknife can also be used to estimate average marginal or non-marginal effects. Such effects are often parameters of interest, especially in nonlinear models, but have received less attention in the literature. We will look at averages of the form

$$\mu_0 \equiv \overline{\mathbb{E}}[\tau_{it}(\theta_0, \alpha_{i0})], \qquad \tau_{it}(\theta, \alpha_i) \equiv \tau(z_{it}; \theta, \alpha_i),$$

where τ is some known function. Examples of such averages were given above. For notational simplicity we take α_i to be a scalar throughout this subsection. The fixed-effect plug-in estimator of μ_0 is

$$\widehat{\mu} \equiv \widehat{\mu}(\widehat{\theta}), \qquad \widehat{\mu}(\theta) \equiv \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tau_{it}(\theta, \widehat{\alpha}_i(\theta)). \tag{2.12}$$

This estimator is subject to two sources of asymptotic bias, each of order $O(T^{-1})$. The first stems from using $\widehat{\alpha}_i(\theta)$ instead of $\alpha_i(\theta)$. The second arises from using $\widehat{\theta}$ instead of θ_0. Hence, $\text{plim}_{N\to\infty}\widehat{\mu} - \mu_0 = O(T^{-1})$ even if a fixed-T consistent or a bias-corrected estimator of θ_0 were used instead of the maximum-likelihood estimator.

To describe how the jackknife can be applied to average effects it is useful to inspect both sources of bias. We will do so under the following two assumptions.

Assumption 2.6. *For all i, as $T \to \infty$,*

$$\widehat{\alpha}_i(\theta_0) - \alpha_{i0} = \frac{\beta_i}{T} + \frac{1}{T} \sum_{t=1}^{T} \psi_{it} + o_p\left(\frac{1}{T}\right), \qquad \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \psi_{it} \xrightarrow{d} \mathcal{N}(0, \sigma_i^2),$$

where ψ_{it} is a martingale difference sequence, and the bias term β_i and the variance $\sigma_i^2 \equiv \mathbb{E}[\psi_{it}^2]$ are finite.

Assumption 2.7. *The function $\tau_{it}(\theta, \alpha_i)$ is three times continuously-differentiable with respect to (θ, α_i). For all i, $\tau_{it}(\theta_0, \alpha_{i0})$ and its cross-derivatives up to the third order are covariance stationary random variables that have autocovariances that are summable. There exist covariance stationary random variables D_{it}^{α} and D_{it}^{θ} with vanishing autocovariances such that $\sup_{\alpha \in \mathcal{A}} |\nabla_{\alpha_i\alpha_i\alpha_i} \tau_{it}(\theta_0, \alpha)| \leq D_{it}^{\alpha}$ and $\sup_{\theta \in \Theta} \|\nabla_{\theta} \tau_{it}(\theta, \alpha_i(\theta))\| \leq D_{it}^{\theta}$ for all i.*

Assumption 2.6 contains a conventional expansion of $\widehat{\alpha}_i(\theta)$ as $T \to \infty$. This expansion follows from standard higher-order asymptotics (see, e.g., Bao and Ullah 2007) and, in fact, underlies the expansion of the bias of $\widehat{\theta}$ and $\widehat{l}(\theta)$ in Assumptions 2.3 or 2.5 (see Hahn and Newey 2004 and Arellano and Hahn 2006). However, because the jackknife does not require knowledge of the form of this bias, we didn't introduce it up to this point. Assumption 2.7 imposes smoothness on the function τ and demands the existence of suitable moments of τ and its derivatives to justify expansions around true parameter values, and imposes dominance conditions to handle the remainder terms in these expansions.

Under these assumptions we can dissect the inconsistency of $\widehat{\mu}$ into two parts. The first part originates from the estimation noise in the fixed effects. It equals

$$\text{plim}_{N\to\infty}\widehat{\mu}(\theta_0) - \mu_0 = \frac{D}{T} + o\left(\frac{1}{T}\right),$$

where the leading bias term has

$$D \equiv \sum_{j=-\infty}^{+\infty} \overline{\mathbb{E}}[\nabla_{\alpha_i}\tau_{it}(\theta_0, \alpha_{i0})\psi_{it-j}] + \overline{\mathbb{E}}[\nabla_{\alpha_i}\tau_{it}(\theta_0, \alpha_{i0})\beta_i] + \frac{1}{2}\overline{\mathbb{E}}[\nabla_{\alpha_i\alpha_i}\tau_{it}(\theta_0, \alpha_{i0})\sigma_i^2].$$

The additional bias introduced through $\widehat{\theta}$ is the product of a Jacobian term with the first-order bias of $\widehat{\theta}$. Moreover,

$$\text{plim}_{N\to\infty}\widehat{\mu} - \mu_0 = \frac{D+E}{T} + o\left(\frac{1}{T}\right),$$

where $E \equiv \overline{\mathbb{E}}[\nabla_{\theta'}\tau_{it}(\theta_0, \alpha_i(\theta_0))]\, B_1$. A jackknife estimator that removes both sources of bias takes the form

$$\widetilde{\mu} \equiv \frac{g}{g-1}\widehat{\mu} - \frac{1}{g-1}\overline{\mu}, \qquad \overline{\mu} \equiv \frac{1}{m}\sum_{j=1}^{m}\overline{\mu}_{\mathcal{S}_j}, \qquad \overline{\mu}_{\mathcal{S}_j} \equiv \sum_{S\in\mathcal{S}_j}\frac{|S|}{T}\widehat{\mu}_S(\widehat{\theta}_S),$$

where $\widehat{\mu}_S(\theta) \equiv \frac{1}{N|S|}\sum_{i=1}^{N}\sum_{t\in S}\tau_{it}(\theta, \widehat{\alpha}_{iS}(\theta))$. Note that $\overline{\mu}$ is constructed using the corresponding subpanel estimates of θ_0. This estimator complements the corrections for static models in Hahn and Newey (2004) and the analytical correction for dynamic models in Fernández-Val (2009), which build on a plug-in estimator of $D+E$ to remove it.

In contrast to estimators of θ_0, plug-in estimators of average effects of the form in (2.12) do not converge at the rate $(NT)^{-1/2}$ but at the much slower rate of $N^{-1/2}$. To see why, consider the hypothetical situation in which θ_0 and the α_{i0} are known. An estimator of μ_0 for this case would equal

$$\mu_* \equiv \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tau_{it}(\theta_0, \alpha_{i0}),$$

which clearly is both unbiased and consistent. Now,

$$\mu_* = \frac{1}{N}\sum_{i=1}^{n}\mathbb{E}[\tau_{it}(\theta_0, \alpha_{i0})] + \frac{1}{N}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}\tau_{it}(\theta_0, \alpha_{i0}) - \mathbb{E}[\tau_{it}(\theta_0, \alpha_{i0})]\right).$$

The first right-hand side term does not depend on T and convergences to μ_0 at the rate $N^{-1/2}$. The second right-hand side term converges to zero at the rate $(NT)^{-1/2}$ and so is asymptotically negligible under rectangular-array asymptotics. Hence, $\sqrt{N}(\mu_* - \mu_0)$ has a non-degenerate limit distribution. This implies that any feasible average-effect estimator will converge no faster than at the rate $N^{-1/2}$. Furthermore, under our assumptions,

$$\sqrt{N}\,(\widehat{\mu} - \mu_0) = \sqrt{N}(\mu_* - \mu_0) + O\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\frac{1}{\sqrt{T}}\right),$$

so that both the bias and the estimation noise introduced by replacing θ_0 and the α_{i0} by their maximum-likelihood estimates are negligible under rectangular-array asymptotics. This is a surprising result, and leads to the following theorem.[5]

THEOREM 2.3. *Let Assumptions 2.1, 2.2, 2.3, 2.6, and 2.7 hold. Then* $\text{plim}_{N\to\infty}\widehat{\mu}-\mu_0 = (D+E)/T+o(T^{-1})$ *and* $\text{plim}_{N\to\infty}\widetilde{\mu} - \mu_0 = o(T^{-1})$, *and*

$$\sqrt{N}(\widetilde{\mu} - \widehat{\mu}) = o_p(1), \qquad \sqrt{N}(\widehat{\mu} - \mu_0) \xrightarrow{d} \mathcal{N}(0, \text{var}\{\mathbb{E}[\tau_{it}(\theta_0, \alpha_{i0})]\}),$$

as $N, T \to \infty$ *with* $N/T \to \rho$.

In the Gaussian autoregression, a parameter of interest would be the average effect on the survival function of a marginal change in lagged outcomes, that is,

$$\int_{-\infty}^{+\infty}\frac{\gamma_0}{\sigma_0}\phi\left(\frac{\alpha + \gamma_0 x - s}{\sigma_0}\right)\mathrm{d}G(\alpha)$$

[5] We note that, recently, Fernández-Val and Weidner (2013) found slow convergence of average-effect estimates when there are both fixed and time effects in the model. The result also holds when there are no time effects.

for given x and s. In the standard non-dynamic regression model with i.i.d. data across t, the plug-in estimator of this effect is consistent for fixed T (Hahn and Newey 2004). This is no longer the case in the dynamic setting considered here. A summary statistic for the population can be obtained by averaging over x. For example, averaging with respect to the distribution function of the data yields

$$\mu_0 = \overline{\mathbb{E}} \left[\frac{\gamma_0}{\sigma_0} \phi \left(\frac{\alpha_{i0} + \gamma_0 y_{it-1} - s}{\sigma_0} \right) \right]$$

as the average effect of interest. Note that, under stationarity, the time-series processes are heterogeneous only in their mean. Thus, the limit distribution of plug-in estimates of μ_0 is degenerate unless the α_{i0} have positive variance. To investigate the finite-sample accuracy of the large-sample results in Theorem 2.3, we estimated μ_0 for $s = 0$ from simulated data with $\gamma_0 = .5$, $\sigma_0 = 1$, and $\alpha_{i0} \sim \mathcal{N}(0, 1)$.

The upper block of Table 7 contains the bias and standard deviation of both the maximum-likelihood estimator and the split-panel jackknife estimator of μ_0, as well as the bias and standard deviation of the infeasible estimators μ_* and $\widehat{\mu}(\theta_0)$. It shows that, in addition to μ_* being unbiased, $\widehat{\mu}(\theta_0)$ has negligible bias, even for very small T, while $\widehat{\mu}$ suffers from downward bias. The jackknife correction removes virtually all of this bias in all cases considered. The second block of Table 7 provides the ratio of the (average of the) estimated standard errors of the estimators to their standard deviation over the $10,000$ Monte Carlo replications. Not surprisingly, when T is small compared to N, use of the asymptotic formula results in considerable underestimation of the true variability of both $\widehat{\mu}$ and $\widetilde{\mu}_{1/2}$. Combined with the bias in $\widehat{\mu}$, this results in maximum-likelihood-based confidence intervals having poor coverage. The results also confirm that, under rectangular-array asymptotics, Theorem 2.3 yields correct inference even without bias correction. Nonetheless, although $\widetilde{\mu}_{1/2}$ is somewhat more variable in small samples, the underestimation of its variability is more than compensated by its reduced small-sample bias in terms of confidence. Even for the larger values of T considered here, $\widetilde{\mu}_{1/2}$ appears preferable over $\widehat{\mu}$.

These results show that, in spite of the results in Theorem 2.3, in small samples one may still want to perform some bias correction. Furthermore, even though the theorem provides an asymptotic justification for inference based on a plug-in estimator of the cross-sectional variance of $\mathbb{E}[\tau_{it}(\theta_0, \alpha_{i0})]$, the within-group variance and the estimation noise in the plug-in estimates of the fixed effects and common parameters may be sizeable for small T and, indeed, may dominate in micropanels. Therefore, it may be useful to consider a variance estimator that accounts for this noise. One possible estimator is a plug-in version of

$$\operatorname{var}\{\mathbb{E}[\tau_{it}(\theta_0, \alpha_{i0})]\} + \frac{1}{T} \sum_{j=-\infty}^{+\infty} \overline{\mathbb{E}}[v_{it} v_{it-j}],$$

where the second term adds an $O(T^{-1})$ correction. A natural choice for v_{it} is

$$(\tau_{it}(\theta_0, \alpha_{i0}) - \mathbb{E}[\tau_{it}(\theta_0, \alpha_{i0})]) + \mathbb{E}[\nabla_{\alpha_i} \tau_{it}(\theta_0, \alpha_{i0})]\, \psi_{it} + \overline{\mathbb{E}}[\nabla_{\theta'} \tau_{it}(\theta_0, \alpha_{i0}) + \nabla_{\alpha_i} \tau_{it}(\theta_0, \alpha_{i0})\, \nabla_{\theta'} \widehat{\alpha}_i(\theta_0)] \Sigma^{-1} s_{it}(\theta_0).$$

Here, the first term captures the within-group variance, and the remaining terms account for the variance in the plug-in estimates of the fixed effects and common parameters, respectively. We experimented with this alternative variance estimator in our Monte Carlo experiment. The results are reported in the last block of Table 7. These adjusted standard errors were computed using a triangular kernel and a bandwidth set to one. For the infeasible estimator μ_*, the correction term consists only of the within-group variance while, for $\widehat{\mu}(\theta_0)$, the correction involves the first two components on v_{it} only. The table shows that, here, the addition of the small-T correction to the variance does fairly little to improve the ratio of standard error to standard deviation for all estimators, and so leads to only a relatively small improvement of the confidence intervals.

Table 7. Average derivative of the survival function at zero

		bias				sd			
N	T	$\widehat{\mu}$	$\widetilde{\mu}_{1/2}$	μ_*	$\widehat{\mu}(\theta_0)$	$\widehat{\mu}$	$\widetilde{\mu}_{1/2}$	μ_*	$\widehat{\mu}(\theta_0)$
100	4	$-.071$	$-.012$	.000	$-.004$	.009	.020	.007	.007
100	8	$-.036$	.000	.000	$-.002$	.007	.012	.006	.006
100	12	$-.023$	.001	.000	$-.002$	.007	.010	.006	.006
100	16	$-.017$	.001	.000	$-.001$	.007	.009	.006	.006
100	24	$-.011$	.001	.000	$-.001$	.007	.008	.006	.006
50	50	$-.005$	.000	.000	$-.001$	.009	.010	.009	.009
100	100	$-.003$	.000	.000	.000	.006	.006	.006	.006
250	250	$-.001$	.000	.000	.000	.004	.004	.004	.004
		se/sd				confidence			
N	T	$\widehat{\mu}$	$\widetilde{\mu}_{1/2}$	μ_*	$\widehat{\mu}(\theta_0)$	$\widehat{\mu}$	$\widetilde{\mu}_{1/2}$	μ_*	$\widehat{\mu}(\theta_0)$
100	4	.175	.305	.993	.992	.000	.386	.945	.903
100	8	.601	.535	1.020	1.025	.001	.706	.951	.936
100	12	.738	.649	.994	.995	.046	.791	.948	.935
100	16	.819	.734	1.006	1.005	.190	.848	.948	.942
100	24	.886	.822	1.003	1.004	.492	.890	.949	.947
50	50	.950	.922	1.004	1.004	.880	.925	.946	.945
100	100	.976	.962	1.001	1.000	.920	.938	.949	.949
250	250	.991	.985	1.000	1.001	.938	.945	.947	.948
		se/sd with correction				confidence with correction			
N	T	$\widehat{\mu}$	$\widetilde{\mu}_{1/2}$	μ_*	$\widehat{\mu}(\theta_0)$	$\widehat{\mu}$	$\widetilde{\mu}_{1/2}$	μ_*	$\widehat{\mu}(\theta_0)$
100	4	.214	.368	1.015	1.009	.000	.452	.950	.909
100	8	.669	.597	1.036	1.055	.002	.761	.955	.943
100	12	.797	.703	1.006	1.022	.057	.829	.950	.942
100	16	.869	.781	1.016	1.028	.215	.874	.951	.947
100	24	.924	.858	1.010	1.022	.522	.906	.950	.950
50	50	.971	.943	1.008	1.014	.886	.931	.947	.947
100	100	.987	.973	1.003	1.005	.922	.941	.949	.951
250	250	.996	.990	1.001	1.003	.939	.946	.947	.948

Model: $y_{it} = \alpha_{i0} + \gamma_0 y_{it-1} + \varepsilon_{it}$, $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_0^2)$, stationary y_{i0}, $\alpha_{i0} \sim \mathcal{N}(0, 1)$, $\gamma_0 = .5$, and $\sigma_0^2 = 1$. 10,000 Monte Carlo replications.

3. EXTENSIONS

3.1. *Higher-order bias correction*

In the previous section we showed how to remove the leading bias from $\widehat{\theta}$ and $\widehat{l}(\theta)$ by means of the jackknife to obtain first-order bias-corrected estimators. It is natural to expect that, in sufficiently smooth models, the inconsistency can be expanded to a higher order, say k, as in (2.4). This raises the question of how to construct estimators that remove the first $h \leq k$ bias terms. Continuing the argument behind the half-panel jackknife readily leads to such estimators. This is another instance of the simplicity of the jackknife that is not shared by the analytical corrections, for which as yet no higher-order generalizations have been obtained. For brevity, we restrict attention to bias corrections applied to the estimator, $\widehat{\theta}$. The development of higher-order corrections of the profile likelihood and average effects is analogous. It is beyond the scope of this paper to derive primitive conditions for the required expansions to hold to the required order, but we discuss two models that are tractable enough to derive θ_T or $l_T(\theta)$ and to establish the existence of their expansions to $o(T^{-k})$ for any positive integer k. Technical details for this subsection are given in the supplementary material.

3.1.1. Higher-order bias correction The h leading terms in (2.4) are simultaneously estimated and removed by suitably combining weighted averages of subpanel estimators associated with collections of subpanels of different length. To illustrate, suppose for a moment that T is divisible by both 2 and 3. Then, using obvious notation for the averages over subpanel estimators, $(1 + a_{1/2} + a_{1/3})\widehat{\theta} - a_{1/2}\overline{\theta}_{1/2} - a_{1/3}\overline{\theta}_{1/3}$ has zero first- and second-order bias if $a_{1/2}$ and $a_{1/3}$ satisfy

$$\left(\frac{1 + a_{1/2} + a_{1/3}}{T} - \frac{a_{1/2}}{T/2} - \frac{a_{1/3}}{T/3}\right) B_1 = 0, \tag{3.1}$$

$$\left(\frac{1 + a_{1/2} + a_{1/3}}{T^2} - \frac{a_{1/2}}{(T/2)^2} - \frac{a_{1/3}}{(T/3)^2}\right) B_2 = 0, \tag{3.2}$$

regardless of B_1 and B_2. This gives $a_{1/2} = 3$ and $a_{1/3} = -1$, leading to the estimator $3\widehat{\theta} - 3\overline{\theta}_{1/2} + \overline{\theta}_{1/3}$, whose inconsistency of $o(T^{-2})$.

Now let $G \equiv \{g_1, g_2, \ldots, g_h\}$ be a non-empty set of integers with $2 \leq g_1 < g_2 < \cdots < g_h$. For $T \geq g_h T_{\min}$ and each $g \in G$, let $\mathcal{S}_g$ be a collection of g non-overlapping subpanels forming an almost equal partition of $\{1, 2, \ldots, T\}$, with equivalence class $\{\mathcal{S}_{gj}; j = 1, 2, \ldots, m_g\}$. Let A be the $h \times h$ matrix with elements

$$[A]_{r,s} \equiv \sum_{S \in \mathcal{S}_{g_s}} \left(\frac{T}{|S|}\right)^{r-1}, \qquad r, s = 1, 2, \ldots, h,$$

and let a_{1/g_r} be the rth element of $(1 - \iota' A^{-1}\iota)^{-1} A^{-1}\iota$, where ι is the $h \times 1$ summation vector. Define the jackknife estimator

$$\widetilde{\theta}_{1/G} \equiv \left(1 + \sum_{g \in G} a_{1/g}\right)\widehat{\theta} - \sum_{g \in G} a_{1/g}\overline{\theta}_{1/g}, \qquad \overline{\theta}_{1/g} \equiv \frac{1}{m_g}\sum_{j=1}^{m_g} \overline{\theta}_{\mathcal{S}_{gj}}, \tag{3.3}$$

with $\overline{\theta}_{\mathcal{S}_{gj}}$ defined by (2.2). The coefficients $a_{1/g}$ solve an $h \times h$ linear-equation system, of which (3.1)–(3.2) is a special case, that ensures that $\widehat{\theta}_{1/G}$ has zero bias up to and including order h. Provided (2.4) holds for $k \geq h$, it will follow from Assumptions 2.1, 2.2, and 2.3 that $\operatorname{plim}_{N\to\infty}\widetilde{\theta}_{1/G} = \theta_0 + o(T^{-h})$ and

$$\sqrt{NT}(\widetilde{\theta}_{1/G} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1})$$

as $N, T \to \infty$ with $N/T \to \rho$. Thus, the higher-order jackknife does not inflate the asymptotic variance.

Like the first-order bias correction, the higher-order bias corrections come at the cost of increasing the higher-order bias terms that are not eliminated. Theorem S.2.2 in the supplementary material characterizes the higher-order bias. It follows from this characterization that, for bias correction of order h, the choice $G = \{2, 3, \ldots, h+1\}$ is optimal in the class $\widetilde{\theta}_{1/G}$ in the sense of minimizing all higher-order terms that are not eliminated. How to choose h optimally in practice is a difficult issue because the choice should also be guided by variance considerations. As such, higher-order asymptotic approximations of both the bias and the variance are needed to answer the question in a satisfactory manner.

3.1.2. Examples Our first example is the Gaussian autoregression, and our focus will be on a higher-order expansion of the Nickell (1981) bias. The model is

$$y_{it} = \alpha_{i0} + \gamma_0 y_{it-1} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathcal{N}(0, \sigma_0^2), \qquad y_{i0} \sim \mathcal{N}\left(\frac{\alpha_{i0}}{1-\gamma_0}, \frac{\sigma_0^2}{1-\gamma_0^2}\right).$$

For $|\gamma_0| < 1$, the inconsistency of the within-group estimator $\widehat{\gamma}$ for fixed T is available in closed form (Nickell, 1981, Equation (18)). It can be expanded as $\gamma_T - \gamma_0 = \sum_{j=1}^{k} B_j/T^j + O(T^{-k-1})$ for any k. The first few terms of this expansion, in the case $|\gamma_0| < 1$, are given by

$$\gamma_T - \gamma_0 = -\frac{1+\gamma_0}{T} - \frac{r\,(1+\gamma_0)}{T^2} + \frac{r\,(1+\gamma_0)}{T^3} + \frac{\left(r + 4r^2 + 2r^3\right)(1+\gamma_0)}{T^4} + O(T^{-5}),$$

with $r \equiv \gamma_0/(1-\gamma_0)$. Consequently, in this model, the jackknife of any order will be asymptotically bias-reducing. Table 8 gives numerical values of the asymptotic biases when $\gamma_0 = .5, .9$ for values of T up to 40 and up to the third-order jackknife. It is clearly seen from the table that the asymptotic bias converges to zero at a faster rate in T as we move to higher-order versions of the jackknife. The table also includes the unit-root case, $\gamma_0 = 1$, where the inconsistency of the within-group estimator is the limit of the Nickell bias,

$$\lim_{\gamma_0 \uparrow 1} \left(\gamma_T - \gamma_0\right) = -\frac{3}{T+1} = -\frac{3}{T} + \frac{3}{T^2} - \frac{3}{T^3} + \dots$$

It follows from this expansion that, interestingly, the jackknife remains a valid tool for bias correction when there is a unit root. Note that the leading bias term is not $\lim_{\gamma_0 \uparrow 1}[-(1+\gamma_0)/T]$, so the plug-in estimator from the stationary case no longer delivers bias-corrected point estimates (see also Hahn and Kuersteiner 2002, Theorems 4 and 5).

Table 8. Asymptotic bias in the Gaussian autoregression

T	4	5	6	8	10	12	16	20	30	40
					$\gamma_0 = .5$					
$\widehat{\gamma}$	−.411	−.331	−.276	−.205	−.162	−.134	−.099	−.079	−.052	−.038
$\widehat{\gamma}_{1/2}$	−.073	−.041	−.016	.002	.007	.008	.007	.005	.003	.002
$\widehat{\gamma}_{1/\{2,3\}}$			.030	.026	.020	.014	.007	.004	.001	.000
$\widehat{\gamma}_{1/\{2,3,4\}}$						.009	.003	.001	.000	.000
					$\gamma_0 = .9$					
$\widehat{\gamma}$	−.560	−.463	−.394	−.302	−.243	−.203	−.151	−.120	−.077	−.056
$\widehat{\gamma}_{1/2}$	−.171	−.123	−.081	−.043	−.023	−.012	−.001	.004	.007	.007
$\widehat{\gamma}_{1/\{2,3\}}$			−.012	.002	.009	.012	.014	.013	.010	.008
$\widehat{\gamma}_{1/\{2,3,4\}}$						.016	.015	.013	.009	.006
					$\gamma_0 = 1$					
$\widehat{\gamma}$	−.600	−.500	−.429	−.333	−.273	−.231	−.176	−.143	−.097	−.073
$\widehat{\gamma}_{1/2}$	−.200	−.150	−.107	−.067	−.045	−.033	−.020	−.013	−.006	−.004
$\widehat{\gamma}_{1/\{2,3\}}$			−.036	−.020	−.011	−.007	−.003	−.002	−.001	.000
$\widehat{\gamma}_{1/\{2,3,4\}}$						−.002	−.001	.000	.000	.000

Model: $y_{it} = \alpha_{i0} + \gamma_0 y_{it-1} + \varepsilon_{it}$, $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_0^2)$, stationary y_{i0} when $\gamma_0 < 1$.

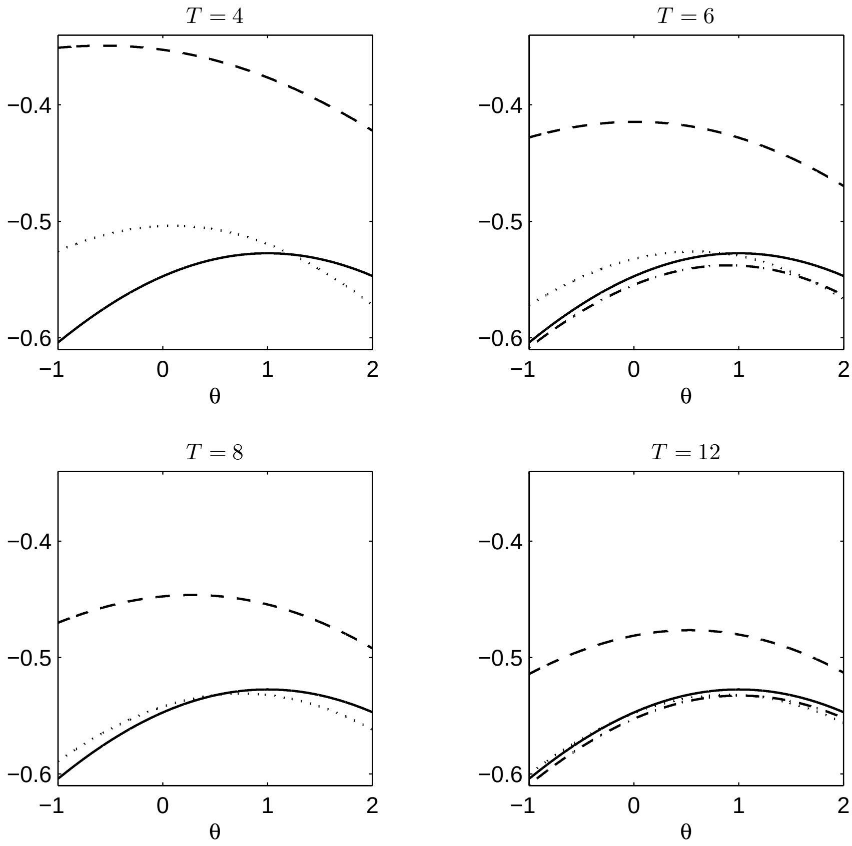
The second example is the stationary autoregressive logit model

$$y_{it} = 1\{\alpha_{i0} + \theta_0 y_{it-1} \geq \varepsilon_{it}\},$$

where the ε_{it} are i.i.d. with distribution function $F(\varepsilon) = e^{\varepsilon}/(1+e^{\varepsilon})$ and the y_{i0} are drawn from their respective steady-state distributions. In this model the bias is much more complicated and depends on the transition probabilities which, in turn, are a function of the α_{i0}. It can be shown that a sufficient condition for $l_T(\theta) - l_0(\theta) = \sum_{j=1}^{k} C_j(\theta)/T^j + O(T^{-k-1})$ to hold for all θ and any k is that the distribution of the fixed effects has bounded support. As a numerical illustration of the convergence properties, we computed the functions $l_0(\theta)$, $l_T(\theta)$, and $l_T(\theta)$ jackknifed up to the third order, for $N = \infty$ and $T = 2, \dots, 40$ when $\theta_0 = 1$ and the fixed effects have a discrete distribution with probability .01 on each of the quantiles $\Phi^{-1}(.01j - .005)$, $j = 1, 2, \dots, 100$, of the standard normal distribution. Figure 2 shows graphs for up to the second-order jackknife for $T = 4, 6, 8, 12$. The infeasible $l_0(\theta)$ (solid) does not depend on T and is maximized at $\theta = \theta_0 = 1$. The difference between $l_T(\theta)$ (dashed) and $l_0(\theta)$ is large and vanishes at the rate T^{-1}. Although T is still relatively small, the half-panel jackknife, $2l_T(\theta) - l_{T/2}(\theta)$ (dotted line), is already much closer to $l_0(\theta)$ and is seen to converge faster to $l_0(\theta)$ than $l_T(\theta)$ does. The second-order jackknife,

$3l_T(\theta) - 3l_{T/2}(\theta) + l_{T/3}(\theta)$ (dashed-dotted; for $T = 6, 12$ only), is even closer to $l_0(\theta)$ and converges still faster. The improved convergence rate as the jackknife order increases is also borne out by the corresponding maximizers, which are given in Table 9 for values of T up to 40 and up to the jackknife correction of the third order.

Figure 2. Asymptotic profile log-likelihoods in the stationary autoregressive logit model



Model: $y_{it} = 1\{\alpha_{i0} + \theta_0 y_{it-1} \geq \varepsilon_{it}\}$, ε_{it} logistically distributed, stationary y_{i0}. True values: $\theta_0 = 1$, α_{i0} approximately $\mathcal{N}(0, 1)$. Plots: $l_0(\theta)$ (solid), $l_T(\theta)$ (dashed), $2l_T(\theta) - l_{T/2}(\theta)$ (dotted), $3l_T(\theta) - 3l_{T/2}(\theta) + l_{T/3}(\theta)$ (dashed-dotted; for $T = 6, 12$ only).

Table 9. Asymptotic bias in the stationary autoregressive logit model

T	4	5	6	8	10	12	16	20	30	40
$\widehat{\theta}$	−1.574	−1.208	−.984	−.720	−.568	−.469	−.348	−.276	−.183	−.136
$\dot{\theta}_{1/2}$	−.903	−.642	−.431	−.245	−.155	−.105	−.057	−.035	−.015	−.008
$\dot{\theta}_{1/\{2,3\}}$			−.100	−.030	.002	.008	.007	.005	.002	.001
$\dot{\theta}_{1/\{2,3,4\}}$						.019	.007	.003	.001	.000

Model: $y_{it} = 1\{\alpha_{i0} + \theta_0 y_{it-1} \geq \varepsilon_{it}\}$, ε_{it} logistically distributed, stationary y_{i0}. True values: $\theta_0 = 1$, α_{i0} approximately $\mathcal{N}(0, 1)$.

3.2. Two-step estimators

Triangular simultaneous-equation models are frequent in microeconometrics. They arise when dealing with endogeneity of covariates or non-random sample selection, for example. Although, in principle, such models can be estimated by full-information maximum likelihood, the use of limited-information methods—i.e., two-step estimators based on control functions (Heckman and Robb 1985)—is more frequent in applied work. One reason for this is that they are typically easier to implement (Rivers and Vuong 1988). Another reason is that such two-step estimators can be generalized to semiparametric settings (Blundell and Powell 2003). Here we discuss how the jackknife can be applied to such estimators.

To describe the setup, let $\lambda_{it}(\theta, \alpha_i) \equiv \lambda(z_{it}; \theta, \alpha_i)$ denote the control function, where the functional form of λ is known. Write $\lambda_{it} \equiv \lambda_{it}(\theta_0, \alpha_{i0})$. In a sample-selection problem, λ_{it} would be a function of the propensity score for observation z_{it} to be selected into the sample, an event typically modeled as a threshold-crossing process such as a probit model. Clearly, this propensity will depend both on the observed covariates and on α_{i0}. Similarly, when a covariate is endogenous, the control function could be the deviation of the endogenous variable from its mean given a set of instrumental variables and fixed effects; we discuss this example in more detail below.

Suppose the main equation of interest has unknown parameters ϑ_0 and η_{i0}, which uniquely maximize an objective function of the form $\mathbb{E}[q(z_{it}; \vartheta, \eta_i, \lambda_{it})]$. Note that, often, this function will not be a log-likelihood. The two-step fixed-effect estimator of ϑ_0 is

$$\widehat{\vartheta} \equiv \arg\max_{\vartheta} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} q(z_{it}, \vartheta, \widehat{\eta}_i(\vartheta), \widehat{\lambda}_{it}), \tag{3.4}$$

where $\widehat{\eta}_i(\vartheta) \equiv \arg\max_{\eta_i} \frac{1}{T}\sum_{t=1}^{T} q(z_{it}, \vartheta, \eta_i, \widehat{\lambda}_{it})$ and $\widehat{\lambda}_{it} \equiv \lambda_{it}(\widehat{\theta}, \widehat{\alpha}_i(\widehat{\theta}))$, the fixed-effect estimator of the control function. As before, typically, $\vartheta_T \equiv \text{plim}_{N\to\infty}\widehat{\vartheta} \neq \vartheta_0$. Under regularity conditions, $\vartheta_T - \vartheta_0$ can again be expanded in powers of T^{-1}. Because $\widehat{\lambda}_{it}$ is a generated regressor which is itself estimated with bias $O(T^{-1})$, however, the bias formula in Hahn and Kuersteiner (2011) will no longer apply to this expansion. Furthermore, the functional form of the leading bias changes if one uses a bias-corrected estimator instead of $\widehat{\theta}$ in the construction of the control function. Fernández-Val and Vella (2011) provide the exact bias expression for this case and extend the analytical bias-correction approach of Hahn and Kuersteiner (2011) to two-step estimators.

The additional complexity of the form of the leading bias of $\widehat{\vartheta}$ due to the presence of generated regressors is substantial. Nonetheless, given that this bias is of the form B/T for some constant B, the jackknife will remove it regardless of where its components arise from. To describe the correction, consider a subpanel S and let

$$\widehat{\vartheta}_S \equiv \arg\max_{\vartheta} \frac{1}{N|S|} \sum_{i=1}^{N} \sum_{t\in S} q(z_{it}, \vartheta, \widehat{\eta}_{iS}(\vartheta), \widehat{\lambda}_{itS}),$$

where $\widehat{\eta}_{iS}(\vartheta) \equiv \arg\max_{\eta_i} \frac{1}{|S|}\sum_{t\in S}^{T} q(z_{it}, \vartheta, \eta_i, \widehat{\lambda}_{itS})$ and $\widehat{\lambda}_{itS} \equiv \lambda_{it}(\widehat{\theta}_S, \widehat{\alpha}_{iS}(\widehat{\theta}_S))$. Observe that the plug-in estimator of the control function, too, uses first-step estimates based on the subpanel. Indeed, the key point to forming a jackknife correction of $\widehat{\vartheta}$ will be that the full two-step estimator has to be computed for each chosen subpanel. The intuition behind this is the presence of estimates of α_{i0} and θ_0 in the first-stage equation and, as such, is analogous to the one behind the jackknife correction of average effects above. The half-panel jackknife estimator for the two-step estimation problem then is

$$\widetilde{\vartheta}_{1/2} \equiv 2\widehat{\vartheta} - \overline{\vartheta}_{1/2},$$

again using obvious notation. Under regularity conditions, $\widetilde{\vartheta}_{1/2}$ will be asymptotically normal and correctly centered as $N/T \to \rho$. Its influence function has the form of that of a conventional two-step estimator (see, e.g., Murphy and Topel 1985) and is omitted here for the sake of brevity. The expression for the asymptotic variance in question is given in Fernández-Val and Vella (2011).

As an illustration, consider a triangular model where (y_{it}, x_{it}) are jointly generated through the structure

$$y_{it} = 1\{\eta_{i0} + \gamma_0 y_{it-1} + \delta_0 x_{it} + u_{it} \geq 0\}, \qquad x_{it} = \alpha_{i0} + \varrho_0 x_{it-1} + \varpi_0 w_{it} + v_{it}, \tag{3.5}$$

where w_{it} is a covariate that is determined exogenously, and (u_{it}, v_{it}) are latent disturbances which are independent and identically distributed as

$$\begin{pmatrix} u_{it} \\ v_{it} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \zeta_0\sigma_0 \\ \zeta_0\sigma_0 & \sigma_0^2 \end{pmatrix} \right) \tag{3.6}$$

for correlation coefficient ζ_0. The model in (3.5)–(3.6) is a routinely referred to as the simultaneous probit model. Its cross-section has received considerable attention in the literature. Here, $\theta_0 = (\varrho_0, \varpi_0, \sigma_0^2)'$ and $\vartheta_0 = (\gamma_0, \delta_0, \zeta_0)'$. The joint likelihood of the data is complicated and full-information maximum likelihood is computationally troublesome (Heckman 1978). Now, the likelihood contribution of an observation factors as

$$\ell_{it}(\vartheta, \eta_i; \theta, \alpha_i) = \ell_{it}(\vartheta, \eta_i | \theta, \alpha_i)\; \ell_{it}(\theta, \alpha_i),$$

say, where $\ell_{it}(\theta, \alpha_i)$ is the contribution to the marginal likelihood of x_{it} and $\ell_{it}(\vartheta, \eta_i|\theta, \alpha_i)$ is the contribution to the conditional likelihood of y_{it} given x_{it}. These contributions are

$$\ell_{it}(\theta, \alpha_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x_{it} - \alpha_i - \varrho x_{it-1} - \varpi w_{it})^2}{\sigma^2} \right),$$

which corresponds to the likelihood for a standard linear model, and

$$\ell_{it}(\vartheta, \eta_i|\theta, \alpha_i) = \Phi\left(\frac{\eta_i + \gamma y_{it-1} + \delta x_{it} + \zeta v_{it}(\theta, \alpha_i)}{\sqrt{1-\zeta^2}} \right)^{y_{it}} \left[1 - \Phi\left(\frac{\eta_i + \gamma y_{it-1} + \delta x_{it} + \zeta\, v_{it}(\theta, \alpha_i)}{\sqrt{1-\zeta^2}} \right) \right]^{1-y_{it}},$$

where $v_{it}(\theta, \alpha_i) \equiv (x_{it} - \alpha_i - \varrho x_{it-1} - \varpi w_{it})/\sigma$. This would be a conventional probit objective function for the rescaled parameter $\vartheta/\sqrt{1-\zeta^2}$ if θ_0 and the α_{i0} were known. Thus, here, $\lambda_{it}(\theta, \alpha_i) = v_{it}(\theta, \alpha_i)$ and, following Smith and Blundell (1986) and Rivers and Vuong (1988), a two-step fixed-effect estimator is a conventional probit estimator, where the residual of a first-stage least-squares regression is added as a regressor. This two-step estimator is very easy to implement.

As another example, consider the reverse situation in which

$$y_{it} = \eta_{i0} + \gamma_0 y_{it-1} + \delta_0 x_{it} + v_{it}, \qquad x_{it} = 1\{\alpha_{i0} + \varrho_0 x_{it-1} + \varpi_0 w_{it} + u_{it} \geq 0\}, \tag{3.7}$$

where (u_{it}, v_{it}) are as before. In this case, for $\theta_0 = (\varrho_0, \varpi_0)'$ and $\vartheta_0 = (\gamma_0, \delta_0, \zeta_0, \sigma_0^2)'$, the joint likelihood has contributions

$$\ell_{it}(\vartheta, \eta_i; \theta, \alpha_i) = \frac{1}{\sigma} \phi\left(v_{it}(\vartheta, \eta_i)\right) \Phi\left(\frac{u_{it}(\theta, \alpha_i) + \zeta\, v_{it}(\vartheta, \eta_i)}{\sqrt{1-\zeta^2}} \right)^{x_{it}} \left[1 - \Phi\left(\frac{u_{it}(\theta, \alpha_i) + \zeta\, v_{it}(\vartheta, \eta_i)}{\sqrt{1-\zeta^2}} \right) \right]^{1-x_{it}},$$

for $v_{it}(\vartheta, \eta_i) \equiv (y_{it} - \eta_i - \gamma y_{it-1} - \delta x_{it})/\sigma$ and $u_{it}(\theta, \alpha_i) \equiv \alpha_i + \varrho x_{it-1} + \varpi w_{it}$. Although a factorization is still possible, it does not readily provide an estimator. However, a simple two-step estimator can be constructed from the observation that

$$\mathbb{E}[y_{it}|y_{it-1}, x_{it}, x_{it-1}, w_{it}, \eta_{i0}, \alpha_{i0}] = \eta_{i0} + \gamma_0 y_{it-1} + \delta_0 x_{it} + \varsigma_0 \lambda_{it},$$

where $\varsigma_0 = \zeta_0 \sigma_0$ and the control function is

$$\lambda_{it}(\theta, \alpha_i) = [x_{it} - \Phi(u_{it}(\theta, \alpha_i))] \frac{\phi(u_{it}(\theta, \alpha_i))}{\Phi(u_{it}(\theta, \alpha_i)) [1 - \Phi(u_{it}(\theta, \alpha_i))]},$$

as can be shown using standard properties of the bivariate normal density. Observe that λ_{it} is the generalized residual (Gouriéroux, Monfort, Renault, and Trognon 1987) from a probit model for the first-stage equation. Therefore, again, a two-step estimator can be easily implemented. First estimate a standard fixed-effect probit model for x_{it} to construct a plug-in estimate of λ_{it}. Next estimate $(\gamma_0, \delta_0, \varsigma_0)$ by running a least-squares regression of y_{it} on a set of unit-specific intercepts, y_{it-1} and x_{it}, and the estimate of the control function.

Table 10. Simulation results for the two-step estimator

		bias						sd					
N	T	$\widehat{\gamma}$	$\widehat{\delta}$	$\widehat{\varsigma}$	$\widetilde{\gamma}_{1/2}$	$\widetilde{\delta}_{1/2}$	$\widetilde{\varsigma}_{1/2}$	$\widehat{\gamma}$	$\widehat{\delta}$	$\widehat{\varsigma}$	$\widetilde{\gamma}_{1/2}$	$\widetilde{\delta}_{1/2}$	$\widetilde{\varsigma}_{1/2}$
500	6	−.226	.113	−.094	−.009	−.056	.038	.017	.111	.069	.027	.196	.125
500	8	−.168	.108	−.084	.002	−.066	.043	.014	.095	.059	.021	.157	.100
500	12	−.109	.087	−.064	.007	−.075	.047	.011	.073	.047	.015	.107	.069
500	18	−.072	.064	−.045	.005	−.039	.026	.009	.056	.036	.011	.072	.046
20	20	−.068	.061	−.042	.003	−.023	.016	.041	.264	.169	.050	.330	.211
50	50	−.026	.023	−.015	.001	−.001	.001	.016	.098	.063	.017	.102	.066
100	100	−.013	.012	−.008	.000	.000	.000	.008	.047	.031	.008	.048	.031
		se/sd						confidence					
N	T	$\widehat{\gamma}$	$\widehat{\delta}$	$\widehat{\varsigma}$	$\widetilde{\gamma}_{1/2}$	$\widetilde{\delta}_{1/2}$	$\widetilde{\varsigma}_{1/2}$	$\widehat{\gamma}$	$\widehat{\delta}$	$\widehat{\varsigma}$	$\widetilde{\gamma}_{1/2}$	$\widetilde{\delta}_{1/2}$	$\widetilde{\varsigma}_{1/2}$
500	6	.894	.792	.809	.635	.552	.497	.000	.692	.574	.756	.704	.653
500	8	.904	.809	.825	.642	.510	.520	.000	.659	.572	.778	.494	.540
500	12	.927	.840	.846	.713	.673	.680	.000	.666	.612	.796	.706	.748
500	18	.938	.880	.886	.791	.791	.790	.000	.712	.680	.849	.852	.862
20	20	.934	.883	.890	.796	.799	.798	.578	.906	.906	.882	.876	.877
50	50	.966	.909	.918	.908	.887	.901	.604	.916	.918	.925	.921	.922
100	100	.975	.933	.938	.943	.938	.929	.601	.924	.926	.935	.935	.932

Model: $y_{it} = \eta_{i0} + \gamma_0 y_{it-1} + \delta_0 x_{it} + v_{it}$ and $x_{it} = 1\{\alpha_{i0} + \varrho_0 x_{it-1} + \varpi_0 w_{it} + u_{it} \geq 0\}$, stationary (y_{i0}, x_{i0}, z_{i0}). Data generated with $w_{it} = -\sqrt{2/3}\,\alpha_{i0} + .5 w_{it-1} + \mathcal{N}(0,1)$, $\varrho_0 = \varpi_0 = \gamma_0 = \delta_0 = \zeta_0 = .5$, $\sigma_0 = 1$, $\alpha_{i0} \sim \mathcal{N}(0,1)$, and $\eta_{i0} \sim \mathcal{N}(0,1)$. $10,000$ Monte Carlo replications.

To check the small-sample behavior of the two-step estimator we simulated data from the model comprised of (3.6)–(3.7). The data generating process for the binary variable x_{it} was identical to the one used to generate the simulation results in Table 3, with the autoregressive parameter fixed at .5, and so we need not restate the results for the first-stage equation here. For the main equation, we drew $\eta_{i0} \sim \mathcal{N}(0,1)$ and set $\delta_0 = 1 - \gamma_0$ to keep the long-run multiplier of x_{it} on y_{it} fixed. In Table 10 we present results for $\gamma_0 = .5$ and $\zeta_0 = .5$, and for various panel sizes. The table shows that the uncorrected two-step fixed-effect estimator is biased, with the bias being largest for the autoregressive parameter. The asymptotic bias in the limit distribution under rectangular-array asymptotics also manifests itself clearly in the coverage rates for the confidence interval. The jackknife removes most of the bias and yields confidence intervals that are correctly centered as $N/T \to \rho$. Because of the reduction in bias, the coverage rates of the jackknife also improve on the uncorrected estimate when T is much smaller than N, although quite some undercoverage remains in such cases. This is so because the plug-in estimator of the asymptotic variance underestimates the finite-sample variability when T is small. Indeed, in short panels, the ratio of the standard errors to the standard deviations is considerably worse for the jackknife.

4. EMPIRICAL ILLUSTRATION: FEMALE LABOR-FORCE PARTICIPATION

Understanding the determinants behind intertemporal labor-supply decisions of women has been the goal of a substantial literature. Classic work on the behavior at the intensive margin—that is, the number of hours worked—includes Heckman and MaCurdy (1980) and Mroz (1987), among others. Heckman (1993) stresses the importance of decisions regarding the extensive margin, that is, the choice of whether or not to participate in the labor market. It is widely recognized that data on such intertemporal participation decisions are characterized by a high degree of serial correlation, and understanding to which degree this correlation is driven by state dependence and unobserved heterogeneity is of great importance (see, e.g., Heckman 1981a). Hyslop (1999) used a simple model of search behavior under uncertainty to specify the participation decision as a threshold-crossing model and estimated a random-effect probit version of this model from the PSID data. He found evidence of strong state dependence and substantial unobserved heterogeneity in the data. Carro (2007) and Fernández-Val (2009) estimated fixed-effect versions of Hyslop's model and confirmed his main findings. Here, we re-estimate the model in Fernández-Val (2009) using the various bias-correction approaches available.

Let y_{it} be a binary indicator for labor-force participation of individual i at time t. The threshold-crossing specification we will estimate assumes that

$$y_{it} = 1\{\alpha_{i0} + \gamma_0 y_{it-1} + x_{it}'\delta_0 \geq \varepsilon_{it}\}, \tag{4.1}$$

where ε_{it} are independent standard-normal innovations and x_{it} is a vector of time-varying covariates. We included the number of children of at most two years of age (# children 0–2), between 3 and 5 years of age (# children 3–5), and between 6 and 17 years of age (# children 6–17), as well as the log of the husband's earnings (log husband income; expressed in thousands of 1995 U.S. dollars), and a quadratic function of age. We do not include time-constant covariates such as race or level of schooling as they are absorbed into the fixed effect. The interaction between labor-market and fertility decisions has been discussed in Browning (1992), among others. In his random-effect setup, Hyslop (1999) is unable to reject exogeneity of fertility decisions once lagged participation decisions are taken into account.[6]

Like Carro (2007) and Fernández-Val (2009) we estimate (4.1) from waves 13 to 22 of the PSID, which span the period 1979–1988. The sample consists of 1461 women aged between 18 and 60 in 1985 who, throughout the sampling period, were married to men who were in the active labor force the whole time. During the sampling period, 664 women changed participation status at least once. Table 11 provides descriptive statistics over both the full sample and the subsample of informative units per year. Women belonging to the latter group are, on average, younger, have more young children, and are married to a husband whose annual income is higher.

The estimation results for the various estimators are collected in Table 12. Estimated standard errors are given in italics below the point estimates. All bias-corrected estimates show significantly larger state dependence than maximum likelihood, with the coefficient estimates of lagged participation being about one third higher. The upward bias correction for the autoregressive coefficient is in line with the Monte Carlo findings above. The jackknife estimate $\widetilde{\theta}_{1/2}$ of lagged participation is somewhat larger than that of the other estimators; $\dot{\theta}_{1/2}$ is very similar to the analytical corrections. This, too, is in accordance with our Monte Carlo results. The bias adjustments for the coefficients on the impact of the number of children is smaller

[6] Similarly, with cross-sectional data, Mroz (1987) finds statistical evidence that allows treating fertility as exogenous to hours worked once participation is controlled for.

Table 11. Descriptive statistics

	mean and standard deviation (in italics) over all units								
year	1980	1981	1982	1983	1984	1985	1986	1987	1988
lagged participation	.722	.707	.695	.692	.711	.746	.740	.741	.735
	.448	*.455*	*.461*	*.462*	*.453*	*.435*	*.439*	*.438*	*.441*
# children 0–2	.323	.331	.318	.261	.211	.190	.168	.137	.103
	.528	*.537*	*.533*	*.494*	*.455*	*.440*	*.408*	*.363*	*.325*
# children 3–5	.307	.313	.321	.330	.337	.320	.268	.215	.185
	.524	*.525*	*.535*	*.524*	*.540*	*.541*	*.505*	*.463*	*.430*
# children 6–17	.934	.960	.973	1.015	1.034	1.077	1.124	1.164	1.165
	1.138	*1.124*	*1.102*	*1.081*	*1.068*	*1.062*	*1.064*	*1.081*	*1.109*
husband income	39.200	39.041	39.115	40.541	43.039	43.572	44.485	45.580	46.038
	23.514	*23.598*	*30.601*	*34.375*	*41.915*	*39.798*	*42.621*	*53.411*	*55.784*
age	33.310	34.251	35.300	36.279	37.288	38.358	39.272	40.309	41.336
	8.841	*8.848*	*8.829*	*8.861*	*8.863*	*8.847*	*8.845*	*8.837*	*8.857*
	mean and standard deviation (in italics) over informative units								
year	1980	1981	1982	1983	1984	1985	1986	1987	1988
lagged participation	.575	.531	.502	.496	.540	.620	.606	.609	.595
	.495	*.499*	*.500*	*.500*	*.499*	*.486*	*.489*	*.488*	*.491*
# children 0–2	.413	.438	.410	.338	.267	.242	.198	.152	.105
	.575	*.597*	*.577*	*.550*	*.503*	*.487*	*.429*	*.385*	*.327*
# children 3–5	.377	.386	.410	.425	.447	.410	.339	.276	.232
	.563	*.565*	*.586*	*.566*	*.587*	*.583*	*.556*	*.519*	*.465*
# children 6–17	.840	.906	.956	1.038	1.097	1.184	1.272	1.363	1.385
	1.107	*1.111*	*1.085*	*1.068*	*1.073*	*1.078*	*1.070*	*1.107*	*1.143*
husband income	40.085	40.406	40.153	42.274	45.354	45.498	46.380	46.811	47.907
	24.110	*24.694*	*33.746*	*41.741*	*50.048*	*44.330*	*49.693*	*45.601*	*54.061*
age	31.545	32.496	33.535	34.526	35.565	36.592	37.512	38.551	39.575
	8.278	*8.282*	*8.259*	*8.303*	*8.315*	*8.277*	*8.277*	*8.272*	*8.295*

Data source: PSID 1979–1988.

and similar for all estimators, taking standard errors into account. Regarding the husband's income and the woman's age, $\dot{\theta}_{\mathrm{AH}}$ deviates from the other estimators, with point estimates that are insignificantly different from zero at conventional significance levels. The other procedures find a significant negative impact of an increase in the husband's income on the participation propensity, and a significant concave response to an increase in the woman's age.

The last two columns of the table provide maximum-likelihood and split-panel jackknife estimates of the average effect for each of the regressors. For lagged participation, the reported effect is the impact of changing y_{it-1} from zero to one on the probability of participation in period t. For the number of children, the effect measures the effect of an additional child in the corresponding age category. The effect for age is defined similarly. For the husband's income, the effect is the derivative of the participation probability. The averaging was done over both the fixed effect and the empirical distribution of the data. The largest impact of adjusting for incidental-parameter bias occurs again for the effect of state dependence, with the estimated marginal effect being adjusted upward by a factor of two. The magnitude of the remaining marginal-effect estimates is adjusted less drastically.

One may express doubt against the underlying assumption of stationarity in this model. It is unlikely that the initial observations on participation are draws from a steady-state distribution. Our investigation into this issue above, however, suggests that this should not be a cause for major concern in this model. Potentially more problematic is that the covariates are not stationary. Obviously, the cross-sectional distributions of age, # children 0–2, # children 3–5, and # children 5–17 change over time, but also the husband's average wage is

Table 12. Female labor-force participation: Estimation results

	index coefficients							average effects (%)	
	$\widehat{\theta}$	$\widetilde{\theta}_{1/2}$	$\widetilde{\theta}_{\mathrm{HK}}$	$\widetilde{\theta}_{\mathrm{F}}$	$\dot{\theta}_{1/2}$	$\dot{\theta}_{\mathrm{AH}}$	$\dot{\theta}_{\mathrm{C}}$	$\widehat{\mu}$	$\widehat{\mu}_{1/2}$
lagged participation	.754	1.335	.993	1.030	1.048	.972	1.083	10.503	21.451
	.043	*.048*	*.043*	*.043*	*.043*	*.043*	*.043*	*.315*	*.337*
# children 0–2	−.561	−.680	−.486	−.441	−.553	−.472	−.416	−6.963	−8.394
	.058	*.063*	*.059*	*.058*	*.059*	*.059*	*.058*	*.213*	*.144*
# children 3–5	−.291	−.370	−.225	−.202	−.262	−.162	−.189	−3.592	−4.458
	.054	*.059*	*.055*	*.055*	*.055*	*.054*	*.055*	*.110*	*.078*
# children 6–17	−.092	−.139	−.072	−.064	−.077	.039	−.054	−1.122	−1.337
	.043	*.047*	*.044*	*.044*	*.044*	*.044*	*.044*	*.035*	*.034*
log husband income	−.235	−.300	−.221	−.199	−.239	.030	−.200	−2.845	−4.382
	.055	*.059*	*.056*	*.056*	*.056*	*.055*	*.056*	*.088*	*.067*
age	2.036	1.542	1.838	1.606	1.813	−.041	1.595	17.681	13.559
	.396	*.445*	*.402*	*.402*	*.405*	*.398*	*.404*	*.647*	*.723*
age squared	−.238	−.176	−.214	−.187	−.214	−.014	−.183	—	—
	.053	*.060*	*.054*	*.054*	*.054*	*.053*	*.054*	—	—

Standard errors in italics. Data source: PSID 1979–1988.

clearly trending upward over the sampling period. This could explain some of the observed differences in the results delivered by the various estimators. Another potential reason is model misspecification, which is likely to show up in the form of diverging estimates across methods. As a robustness check to non-stationarity, we re-estimated the model after including yearly time dummies as additional regressors. Time dummies absorb aggregate time effects and, to some degree, the effect of the changing distribution of the regressors over time. The estimation results were very similar to the ones given here and are available in the supplementary material.

CONCLUDING REMARKS

Our analysis has suggested several routes worth pursuing in future research. First, it would be interesting to further investigate the higher-order properties of bias-corrected estimators. For the jackknife, we derived the higher-order bias in a sequential large N, large T setting. For the analytical bias corrections, the higher-order bias has not yet been derived. A more encompassing analysis should also lead to higher-order variance properties, possibly under joint large N, T asymptotics. This would aid in understanding the differences in small-sample performance between the various bias-correction approaches.

Second, we noticed that inference based on the asymptotic variance can lead to confidence bounds that are too narrow for small T. This is especially so for estimators of average effects and for two-step estimators. In additional Monte Carlo work we found that the nonparametric bootstrap of Efron (1979), applied along the cross-sectional dimension of the panel, can perform much better in such cases. Theoretical results would be very valuable.

Third, it would be worth investigating to what extent the scope of bias correction can be extended beyond the setting of stationary data. We have examined the performance of the jackknife corrections under some common deviations from stationarity and suggested validity tests for the jackknife. In a recent paper, Fernández-Val and Weidner (2013) argue that, under regularity conditions, the introduction of time dummies in a class of linear-index models can be successfully handled by a small modification of the jackknife procedures discussed here.

Lastly, it would be of interest to construct bias-corrected estimators for quantile effects, and to analyze their properties. One technical difficulty to overcome here is the non-smoothness of the moment functions, which implies that the derivation of the relevant expansions must rely on different techniques than those used here.

ACKNOWLEDGMENTS

We are grateful to Manuel Arellano, Stéphane Bonhomme, Victor Chernozhukov, Iván Fernández-Val, Patrick Gagliardini, Bo Honoré, Ulrich Müller, Whitney Newey, Bram Thuysbaert, Frank Windmeijer, Yu Zhu, the editor Enrique Sentana, and two referees for comments and discussion. Iván Fernández-Val also generously shared his code and data with us. Bram Thuysbaert co-authored an early version of this paper, which was circulated under the title "Jackknife bias reduction for nonlinear dynamic panel data models with fixed effects".

REFERENCES

Alvarez, J. and M. Arellano (2003). The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica 71*, 1121–1159.

Alvarez, J. and M. Arellano (2004). Robust likelihood estimation of dynamic panel data models. Working Paper No 0421, CEMFI.

Arellano, M. (2003). Discrete choices with panel data. *Investigaciones Económicas 27*, 423–458.

Arellano, M. and S. 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 S. Bonhomme (2009). Robust priors in nonlinear panel data models. *Econometrica 77*, 489–536.

Arellano, M. and S. Bonhomme (2012). Identifying distributional characteristics in random coefficients panel data models. *Review of Economic Studies 79*, 987–1020.

Arellano, M. and J. Hahn (2006). A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects. Unpublished manuscript.

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

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*, 1229–1279.

Bai, J. (2013). Fixed effects dynamic panel data models, a factor analytical method. *Econometrica 81*, 285–314.

Bao, Y. and A. Ullah (2007). The second-order bias and mean squared error of estimators in time-series models. *Journal of Econometrics 140*, 650–669.

Becker, G. S., M. Grossman, and K. M. Murphy (1994). An empirical analysis of cigarette addiction. *American Economic Review 84*, 396–418.

Blonigen, B. A., R. B. Davies, G. R. Waddell, and H. T. Naughton (2007). FDI in space: Spatial autoregressive relationships in foreign direct investment. *European Economic Review 51*, 1303–1325.

Blundell, R. W. and J. L. Powell (2003). Endogeneity in nonparametric and semiparametric regression models. In M. Dewatripont, L. P. Hansen, and S. J. Turnovsky (Eds.), *Advances In Economics and Econometrics*, Volume II. Econometric Society: Cambridge University Press.

Browning, M. (1992). Children and household economic behavior. *Journal of Economic Literature 30*, 1434–1475.

Browning, M. and J. M. Carro (2007). Heterogeneity and microeconometrics modeling. In R. W. Blundell, W. K. Newey, and T. Persson (Eds.), *Advances In Economics and Econometrics*, Volume III, Chapter 3, pp. 47–74. Cambridge University Press.

Browning, M., M. Ejrnæs, and J. Alvarez (2010). Modeling income processes with lots of heterogeneity. *Review of Economic Studies 77*, 1353–1381.

Bun, M. J. G. and J. F. Kiviet (2006). The effects of dynamic feedbacks on LS and MM estimator accuracy in panel data models. *Journal of Econometrics 132*, 409–444.

Carro, J. M. (2007). Estimating dynamic panel data discrete choice models with fixed effects. *Journal of Econometrics 140*, 503–528.

Carro, J. M. and A. Traferri (2012). State dependence and heterogeneity in health using a bias-corrected fixed-effects estimator. Forthcoming in *Journal of Applied Econometrics*.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies 47*, 225–238.

Chamberlain, G. (1984). Panel data. In Z. Griliches and M. Intriligator (Eds.), *Handbook of Econometrics*, Volume 2 of *Handbook of Econometrics*, Chapter 22, pp. 1247–1315. Elsevier.

Chamberlain, G. (1985). Heterogeneity, omitted variable bias, and duration dependence. In J. J. Heckman and B. Singer (Eds.), *Longitudinal Analysis of Labor Market Data*, Chapter 1, pp. 3–38. Cambridge University Press.

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

Contoyannis, P., A. M. Jones, and N. Rice (2004). The dynamics of health in the British Household Panel Survey. *Journal of Applied Econometrics 19*, 473–503.

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.

Cox, D. R. and N. Reid (1993). A note on the calculation of adjusted profile likelihood. *Journal of the Royal Statistical Society, Series B 55*, 467–471.

de Jong, R. and T. Woutersen (2011). Dynamic time series binary choice. *Econometric Theory 27*, 673–702.

Dhaene, G. and K. Jochmans (2013). Likelihood inference in an autoregression with fixed effects. Discussion Paper No 2013-07, Department of Economics, Sciences Po.

Donohue, J. J. and S. D. Levitt (2001). The impact of legalized abortion on crime. *Quarterly Journal of Economics 116*, 379–420.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*, 1–26.

Fernández-Val, I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics 150*, 71–85.

Fernández-Val, I. and F. Vella (2011). Bias corrections for two-step fixed effects panel data estimators. *Journal of Econometrics 163*, 144–162.

Fernández-Val, I. and M. Weidner (2013). Individual and time effects in nonlinear panel data models with large N, T. CeMMAP Working Paper CWP60/13.

Gouriéroux, C., A. Monfort, E. Renault, and A. Trognon (1987). Generalised residuals. *Journal of Econometrics 34*, 5–32.

Haddad, M. and A. E. Harrison (1993). Positive spillovers from direct foreign investment? Evidence from panel data for Morocco. *Journal of Development Economics 42*, 51–74.

Hahn, J. and G. Kuersteiner (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large. *Econometrica 70*, 1639–1657.

Hahn, J. and G. Kuersteiner (2010). Stationarity and mixing properties of the dynamic tobit model. *Economics Letters 107*, 105–111.

Hahn, J. and G. Kuersteiner (2011). Bias reduction for dynamic nonlinear panel models with fixed effects. *Econometric Theory 27*, 1152–1191.

Hahn, J. and W. K. Newey (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica 72*, 1295–1319.

Hall, B. H. (1978). A general framework for the time series-cross section estimation. *Annales de l'INSEE 30/31*, 177–202.

Hall, R. E. and F. S. Mishkin (1982). The sensitivity of consumption to transitory income: Estimates from panel data on households. *Econometrica 50*, 461–481.

Hausman, J. A., B. H. Hall, and Z. Griliches (1984). Econometric models for count data with an application to the patents-R&D relationship. *Econometrica 52*, 909–938.

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

Heckman, J. J. (1981a). Heterogeneity and state dependence. In S. Rosen (Ed.), *Studies in Labor Markets*, Chapter 3, pp. 91–139. University of Chicago Press.

Heckman, J. J. (1981b). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. In C. F. Manski and D. L. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, Chapter 4, pp. 179–195. MIT Press.

Heckman, J. J. (1993). What has been learned about labor supply in the past twenty years? *American Economic Review 83*, 116–121.

Heckman, J. J. and T. E. MaCurdy (1980). A life cycle model of female labor supply. *Review of Economic Studies 47*, 47–74.

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

Hoch, I. (1962). Estimation of production function parameters combining time-series and cross-section data. *Econometrica 30*, 34–53.

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

Honoré, B. E. and E. Tamer (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica 74*, 611–629.

Hospido, L. (2012). Modelling heterogeneity and dynamics in the volatility of individual wages. *Journal of Applied Econometrics 27*, 386–414.

Hyslop, D. R. (1999). State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. *Econometrica 67*, 1255–1294.

Islam, N. (1995). Growth empirics: A panel data approach. *Quarterly Journal of Economics 110*, 1127–1170.

Javorcik, B. S. (2004). Does foreign direct investment increase the productivity of domestic firms? In search of spillovers through backward linkages. *American Economic Review 94*, 605–627.

Kuh, E. (1959). The validity of cross-sectionally estimated behavior equations in time series applications. *Econometrica 27*, 197–214.

Lancaster, T. (2002). Orthogonal parameters and panel data. *Review of Economic Studies 69*, 647–666.

Li, H., B. Lindsay, and R. Waterman (2003). Efficiency of projected score methods in rectangular array asymptotics. *Journal of the Royal Statistical Society, Series B 65*, 191–208.

Lillard, L. A. and R. J. Willis (1978). Dynamic aspects of earnings mobility. *Econometrica 46*, 985–1012.

Meghir, C. and L. Pistaferri (2004). Income variance dynamics and heterogeneity. *Econometrica 72*, 1–32.

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica 55*, 765–799.

Mundlak, Y. (1961). Empirical production function free of management bias. *Journal of Farm Economics 43*, 44–56.

Murphy, K. M. and R. H. Topel (1985). Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics 3*, 370–379.

Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*, 1–32.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica 49*, 1417–1426.

Pfanzagl, J. and W. Wefelmeyer (1978). A third-order optimum property of the maximum likelihood estimator. *Journal of Multivariate Analysis 8*, 1–29.

Quenouille, H. M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Series B 11*, 68–84.

Quenouille, H. M. (1956). Notes on bias in estimation. *Biometrika 43*, 353–360.

Rivers, D. and Q. Vuong (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics 39*, 347–366.

Sartori, N. (2003). Modified profile likelihood in models with stratum nuisance parameters. *Biometrika 90*, 533–549.

Schmidt, P. and R. C. Sickles (1984). Production frontiers and panel data. *Journal of Business and Economic Statistics 2*, 367–374.

Smith, R. J. and R. W. Blundell (1986). An exogeneity test for simultaneous equations tobit models with an application to labor supply. *Econometrica 54*, 679–686.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics 20*, 39–54.

Woutersen, T. (2002). Robustness against incidental parameters. Working Paper No 20028, Department of Economics, University of Western Ontario.