Bootstrap inference for fixed-effect models

Ayden Higgins[*]
Faculty of Economics
University of Cambridge

Koen Jochmans[†]
Toulouse School of Economics
University of Toulouse Capitole

This version: April 5, 2022

Abstract

The maximum-likelihood estimator of nonlinear panel data models with fixed effects is asymptotically biased under rectangular-array asymptotics. The literature has devoted substantial effort to devising methods to correct the maximum-likelihood estimator for its bias as a means to salvage standard inferential procedures. We show that the (recursive, parametric) bootstrap replicates the distribution of the (uncorrected) maximum-likelihood estimator in large samples. This justifies the use of confidence sets constructed via conventional bootstrap methods. No adjustment for the presence of bias needs to be made.

JEL Classification: C23

Keywords: bootstrap, fixed effects, incidental parameter problem, inference, panel data

Introduction

The maximum-likelihood estimator of models for panel data is well known to perform poorly when fixed effects are included. The estimator is inconsistent under asymptotics where the number of individuals, n, grows large while the number of time periods, m, is

[*] Address: University of Cambridge, Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DD, United Kingdom. E-mail: amh239@cam.ac.uk.

[†] Address: Toulouse School of Economics, 1 esplanade de l'Université, 31080 Toulouse, France. E-mail: koen.jochmans@tse-fr.eu.

Michael Wolf provided much-appreciated comments on bootstrap terminology and Jin Hahn kindly shared supplementary material from Hahn and Newey (2004) with us.

Support from the European Research Council through grant n° 715787 (MiMo), and from the French Government and the ANR under the Investissements d' Avenir program, grant ANR-17-EURE-0010 is gratefully acknowledged.

held fixed (Neyman and Scott 1948). In fact, many parameters of interest are simply not (point) identified in such a setting (see, e.g., Honoré and Tamer 2006). Maximum likelihood is, however, consistent under so-called rectangular-array asymptotics, where n and m grow large at the same rate (Li, Lindsay and Waterman, 2003). Nevertheless, it is asymptotically biased, in general. This implies that confidence sets based on a naive normal approximation to the distribution of the maximum-likelihood estimator have incorrect coverage, even in large samples.

Over the last two decades substantial effort has been devoted to devising procedures that remove the asymptotic bias, thereby recentering the limit distribution around zero and restoring the validity of conventional inference procedures based on it. A discussion of this literature as well as an overview of many available approaches is given by Arellano and Hahn (2007).[1] Theoretical guidelines on which bias-correction method to use and on how to select their respective tuning parameters are mostly absent. This is inconvenient because, although all proposals lead to estimators with the same (first-order) asymptotic properties, they vary greatly in ease of implementation and in how effective they are at salvaging standard inferential procedures in finite samples.

The current paper shows that, under rectangular-array asymptotics, the parametric bootstrap consistently estimates the distribution of the (uncorrected) maximum-likelihood estimator, including its asymptotic bias. This implies that confidence sets constructed using either the basic bootstrap (also known as the reverse-percentile bootstrap) or the studentized bootstrap (using the terminology of Davison and Hinkley 1997, p. 194) have correct coverage in large samples. Thus, bias correction is not needed. The same conclusion is true for averages over the fixed effects, such as their moments or average marginal effects (Chamberlain 1984).

[1] Approaches to correct the maximum-likelihood estimator, either via analytical formulae or a jackknife, are considered by Hahn and Newey (2004), Hahn and Kuersteiner (2011), and Dhaene and Jochmans (2015b). Adjustments to the (profile) likelihood or score equation have been considered by Hahn and Newey (2004) and Arellano and Hahn (2006). Strategies based on simulation are discussed in Dhaene and Jochmans (2015a) and Kim and Sun (2016).

In its simplest form, inference based on the bootstrap only requires a routine to compute the maximum-likelihood estimator.[2] It is useful to stress that, in spite of the presence of possibly many fixed effects, conventional numerical optimization is, in fact, straightforward, by exploiting the sparsity of the Hessian matrix.[3] Furthermore, because many popular fixed-effect specifications such as probit and tobit models involve likelihood functions that are globally concave, finding the global maximizer requires only a few iterations. Finally, an excellent starting value for the bootstrap maximum-likelihood estimator comes in the form of the maximum-likelihood estimator based on the original data, as the latter is used to generate the bootstrap samples.

In Section 1 we present the setting and state our objectives. In Section 2 we describe our bootstrap procedures. In Section 3 we investigate the performance of the bootstrap in three examples using theoretical calculations and simulations. In Section 4 we discuss numerical computation via an efficient Newton-Raphson routine. In Section 5 we collect all the assumptions and formal results that underlie our claims about the validity of the bootstrap in our setting. Concluding remarks end the paper. An appendix contains proofs. Additional technical derivations are collected in a supplement.

1 Maximum-likelihood estimation

Suppose that we have data on n independent stratified observations $\{y_i, y_{i-}, x_i\}$, with $y_i \coloneqq (y_{i1}, \dots, y_{im})$, $y_{i-} = (y_{i(1-p)}, \dots, y_{i0})$, and $x_i \coloneqq (x_{i1}, \dots, x_{im})$. We consider models

[2] Corrections to the estimator require first estimating the asymptotic bias. The latter depends on moments and cross-moments of higher-order derivatives of the likelihood, which can be cumbersome to derive and compute. Adjustments to the (profile) likelihood have the additional inconvenience that they can be difficult to maximize whereas modified (profile) score equations may have multiple roots. An example where this problem arises is discussed in Dhaene and Jochmans (2016).

[3] The usefulness of partitioned-inverse formulae in models with many parameters has been mentioned before; Prentice and Gloeckler (1978) and Chamberlain (1980) did so in the context of duration models and binary-choice models, respectively. It is not clear that it is widely appreciated, however, as estimation with fixed effects is often said to be computationally demanding or even judged to be infeasible.

where the conditional density of y_i given y_{i-} and x_i (relative to some dominating measure) is given by

$$\prod_{t=1}^{m} f(y_{it}|y_{it-1},\dots,y_{it-p},x_{it};\varphi_0,\eta_{i0}),$$

and f is known up to the finite-dimensional parameters φ_0 and η_{i0}. This framework covers autoregressive processes (of order p), for which y_{i-} serves as the initial condition, as well as models with exogenous covariates, x_i. In what follows we will treat both the initial condition and the covariates as fixed.

It is convenient to introduce the shorthand

$$\ell(\varphi,\eta_i|z_{it}) \coloneqq \log f(y_{it}|y_{it-1},\dots,y_{it-p},x_{it};\varphi,\eta_i),$$

where $z_{it} \coloneqq (y_{it},y_{it-1},\dots,y_{it-p},x_{it})$. The maximum-likelihood estimator is

$$(\hat{\varphi},\hat{\eta}_1,\dots,\hat{\eta}_n) \coloneqq \operatorname*{arg\,max}_{\varphi,\eta_1,\dots,\eta_n} \sum_{i=1}^{n}\sum_{t=1}^{m} \ell(\varphi,\eta_i|z_{it}).$$

In sufficiently regular models we have, as $n,m\to\infty$ with ${}^{n}\!/_{m}\to\gamma^2$ for some $0<\gamma<\infty$, that

$$\sqrt{nm}(\hat{\varphi}-\varphi_0) \xrightarrow{L} N(\gamma\beta,\mathit{\Sigma}), \tag{1.1}$$

where β is a non-random (asymptotic) bias term and the variance is $\mathit{\Sigma} \coloneqq (\lim_{n,m\to\infty}\mathit{\Omega}_{nm})^{-1}$ for

$$\mathit{\Omega}_{nm} \coloneqq -\frac{1}{nm}\sum_{i=1}^{n}\sum_{t=1}^{m}\mathbb{E}\left(\frac{\partial^2\ell(\varphi_0,\eta_{i0}|z_{it})}{\partial\varphi\partial\varphi'} - \rho_{i,m}\frac{\partial^2\ell(\varphi_0,\eta_{i0}|z_{it})}{\partial\eta_i\partial\varphi'}\right),$$

with

$$\rho_{i,m} \coloneqq \left(\frac{1}{m}\sum_{t=1}^{m}\mathbb{E}\left(\frac{\partial^2\ell(\varphi_0,\eta_{i0}|z_{it})}{\partial\varphi\partial\eta_i'}\right)\right)\left(\frac{1}{m}\sum_{t=1}^{m}\mathbb{E}\left(\frac{\partial^2\ell(\varphi_0,\eta_{i0}|z_{it})}{\partial\eta_i\partial\eta_i'}\right)\right)^{-1}.$$

See Hahn and Newey (2004) and Hahn and Kuersteiner (2011) for early derivations of this result in static and dynamic models, respectively.

An implication of (1.1) is that confidence regions based on the limit distribution have to account for the bias term β in order to have correct coverage unless ${}^{n}\!/_{m}$ is close to zero, which is not the case in most applications. Corrections to the estimator have the generic form

$$\hat{\varphi} - \frac{\hat{\beta}}{m},$$

where $\hat{\beta}$ is an estimator of β. Such corrections recenter the estimator's limit distribution around zero, thereby restoring the validity of conventional inference procedures based on it.

We may also be interested in parameters of the form

$$\Delta := \lim_{n,m\to\infty} \frac{1}{nm} \sum_{i=1}^{n} \sum_{t=1}^{m} \mathbb{E}(\mu(z_{it}, \varphi_0, \eta_{i0})),$$

for a chosen function μ. Average marginal effects (as discussed in Chamberlain 1984) or moments of the fixed effects are typical examples. The maximum-likelihood estimator of Δ is

$$\hat{\Delta} := \frac{1}{nm} \sum_{i=1}^{n} \sum_{t=1}^{m} \mu(z_{it}, \hat{\varphi}, \hat{\eta}_i)$$

which, similar to $\hat{\varphi}$, also suffers from asymptotic bias. In particular,

$$\sqrt{nm}(\hat{\Delta} - \Delta) \xrightarrow{L} N(\gamma \nabla, \sigma^2).$$

The form of the bias, ∇, is complicated. The asymptotic variance is

$$\sigma^2 := \lim_{n,m\to\infty} \frac{1}{nm} \sum_{i=1}^{n} \sum_{t=1}^{m} \mathbb{E}\left(\textstyle\sum_{j=-\infty}^{+\infty} \upsilon_{it}\, \upsilon_{it-j} + \omega_{it}^2 \right).$$

Here the term involving $\upsilon_{it} := \mu(z_{it}, \varphi_0, \eta_{i0}) - \mathbb{E}(\mu(z_{it}, \varphi_0, \eta_{i0}))$ is the long-run variance of the infeasible estimator that presumes the parameters to be known. The second term is the variance of

$$\omega_{it} := \varpi' \varSigma \left(\frac{\ell(\varphi_0, \eta_{i0}|z_{it})}{\partial \varphi} - \rho_{i,m} \frac{\ell(\varphi_0, \eta_{i0}|z_{it})}{\partial \eta_i} \right) - \varrho_{i,m} \frac{\partial \ell(\varphi_0, \eta_{i0}|z_{it})}{\partial \eta_i},$$

where

$$\varpi := \lim_{n,m\to\infty} \frac{1}{nm} \sum_{i=1}^{n} \sum_{t=1}^{m} \mathbb{E}\left(\frac{\partial \mu(z_{it}, \varphi_0, \eta_{i0})}{\partial \varphi} - \rho_{i,m} \frac{\partial \mu(z_{it}, \varphi_0, \eta_{i0})}{\partial \eta_i} \right)$$

and

$$\varrho_{i,m} := \left(\frac{1}{m}\sum_{t=1}^{m}\mathbb{E}\left(\frac{\partial\mu(z_{it},\varphi_0,\eta_{i0})}{\partial\eta_i'}\right)\right)\left(\frac{1}{m}\sum_{t=1}^{m}\mathbb{E}\left(\frac{\partial^2\ell(\varphi_0,\eta_{i0}|z_{it})}{\partial\eta_i\partial\eta_i'}\right)\right)^{-1}.$$

The term making up the second contribution to σ^2 reflects the fact that the parameters of the model need to be estimated in a first step to be able to estimate Δ.

2 Bootstrap inference

The (parametric) bootstrap we consider imposes the data generating process implied by the maximum-likelihood estimator. A bootstrap observation $y_i^* := (y_{i1}^*, \ldots, y_{im}^*)$ can be generated recursively by drawing y_{it}^* from the fitted transition density obtained from the original data, i.e.,

$$f(y_{it}^*|y_{it-1}^*, \ldots, y_{it-p}^*, x_{it}; \hat{\varphi}, \hat{\eta}_i).$$

The initial condition, like the covariates, is held fixed, i.e., $y_{i-}^* = y_{i-}$. The associated maximum-likelihood estimator is

$$(\hat{\varphi}^*, \hat{\eta}_1^*, \ldots, \hat{\eta}_n^*) := \underset{\varphi,\eta_1,\ldots,\eta_n}{\arg\max}\sum_{i=1}^{n}\sum_{t=1}^{m}\ell(\varphi,\eta_i|z_{it}^*),$$

with $z_{it}^* := (y_{it}^*, y_{it-1}^*, \ldots, y_{it-p}^*, x_{it})$.

The main observation of this paper is that, in regular situations,

$$\sqrt{nm}(\hat{\varphi}^* - \hat{\varphi}) \overset{L^*}{\to} N(\gamma\beta, \varSigma), \tag{2.2}$$

as $n, m \to \infty$ with $^n/_m \to \gamma^2$. Throughout, we use $\overset{L^*}{\to}$ to denote weak convergence of the bootstrap measure. Equations (1.1) and (2.2) reveal that the bootstrap distribution is consistent for the distribution of the maximum-likelihood estimator. Importantly, the bootstrap mimics the asymptotic bias.

It follows from (2.2) that asymptotically-valid confidence intervals can be constructed by the usual reverse-percentile method, without the need to correct the maximum-likelihood estimator (or, indeed, its bootstrap counterpart) for its bias. For example, for a chosen vector of conformable dimension c,

$$\left\{c'\varphi : c'(\hat{\varphi} - \varphi) \leq q_{1-\alpha}^*\right\}$$

is an upper one-sided confidence interval for the linear combination $c'\varphi_0$ with confidence level $(1-\alpha)$ (in large samples) when setting

$$q^*_\alpha = \inf\left\{q^* : \alpha \leq \mathbb{P}^*(c'(\hat{\varphi}^* - \hat{\varphi}) \leq q^*)\right\}.$$

The notation $\mathbb{P}^*$ refers to a probability computed with respect to the bootstrap measure, i.e, conditional on the sample. Thus, the critical value q^*_α is the α-th quantile of the bootstrap distribution of $c'(\hat{\varphi}^* - \hat{\varphi})$. A two-sided (equal-tailed) confidence interval with the same level of confidence is given by

$$\left\{c'\varphi : c'\hat{\varphi} - q^*_{1-\alpha/2} \leq c'\varphi \leq c'\hat{\varphi} - q^*_{\alpha/2}\right\}.$$

In both cases, construction of the confidence interval only requires a routine to calculate the maximum-likelihood estimator.

The conditions underlying (1.1) and (2.2) imply the consistency of the plug-in estimator $\hat{\varSigma}$ and of its bootstrap counterpart $\hat{\varSigma}^*$ for the inverse Fisher information $\varSigma$. Consequently, we may equally use a studentized bootstrap to perform inference on $c'\varphi_0$. For example, the set

$$\left\{c'\varphi : (c'\,\hat{\varSigma}\,c)^{-1/2}c'(\hat{\varphi}^* - \hat{\varphi}) \leq q^*_{1-\alpha}\right\},$$

now with

$$q^*_\alpha = \inf\left\{q^* : \alpha \leq \mathbb{P}^*\big((c'\,\hat{\varSigma}^*\,c)^{-1/2}c'(\hat{\varphi}^* - \hat{\varphi}) \leq q^*\big)\right\},$$

is an upper one-sided confidence set for $c'\varphi_0$ with confidence level $(1-\alpha)$. For multivariate linear combinations $C'\varphi_0$, where C is a conformable matrix, the set

$$\left\{C'\varphi : (\hat{\varphi}^* - \hat{\varphi})'C\,(C'\hat{\varSigma}\,C)^{-1}C'(\hat{\varphi}^* - \hat{\varphi}) \leq q^*_{1-\alpha}\right\},$$

is based on a quadratic form and, hence, ellipsoidal in shape. Here to ensure coverage of $(1-\alpha)$ in large samples we use

$$q^*_\alpha = \inf\left\{q^* : \alpha \leq \mathbb{P}^*\big((\hat{\varphi}^* - \hat{\varphi})'C\,(C'\,\hat{\varSigma}^*\,C)^{-1}C'(\hat{\varphi}^* - \hat{\varphi}) \leq q^*\big)\right\},$$

which is the α-th quantile of the distribution of the bootstrap version of the quadratic form on which the confidence set is based. Other, non-ellipsoidal constructions are equally possible.

Inference on Δ may equally be done via the bootstrap. Given a bootstrap sample and the associated maximum-likelihood estimator, we construct the corresponding plug-in estimator

$$\hat{\Delta}^* := \frac{1}{nm} \sum_{i=1}^{n} \sum_{t=1}^{m} \mu(z_{it}^*, \hat{\varphi}^*, \hat{\eta}_i^*).$$

The bootstrap distribution of $\sqrt{nm}(\hat{\Delta}^* - \hat{\Delta})$ mimics the distribution of $\sqrt{nm}(\hat{\Delta} - \Delta)$, in large samples, i.e.,

$$\sqrt{nm}(\hat{\Delta}^* - \hat{\Delta}) \xrightarrow{L^*} N(\gamma\nabla, \sigma^2),$$

as $n, m \to \infty$ with ${}^n\!/_m \to \gamma^2$. The construction of confidence intervals for Δ is then completely analogous to before.

3 Examples

Many normal means In the classic problem of Neyman and Scott (1948) we observe independent variables

$$z_{it} \sim N(\eta_{i0}, \varphi_0).$$

Maximum likelihood estimates the mean parameters by the within-strata sample averages $\overline{z}_i := {}^1\!/_m \sum_{t=1}^m z_{it}$ and the common variance parameter by

$$\hat{\varphi} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{t=1}^{m} (z_{it} - \overline{z}_i)^2.$$

It is well-known that, in this case,

$$\sqrt{nm}(\hat{\varphi} - \varphi_0) \xrightarrow{L} N(-\gamma\varphi_0, 2\varphi_0^2),$$

under rectangular-array asymptotics. Here, starting from the fact that $nm\,\hat{\varphi}/\varphi_0 \sim \chi^2_{n(m-1)}$, the exact distribution of the maximum-likelihood estimator can be derived. We find that

$$\sqrt{nm}(\hat{\varphi} - \varphi_0) \sim \text{Gamma}\left(-\sqrt{nm}\varphi_0, \frac{n(m-1)}{2}, \frac{2\varphi_0}{\sqrt{nm}}\right),$$

where Gamma$(\vartheta_1, \vartheta_2, \vartheta_3)$ refers to the Gamma distribution with location ϑ_1, shape ϑ_2 and scale ϑ_3. It is readily verified that the mean and variance of this distribution are equal to

$$-\sqrt{\frac{n}{m}}\varphi_0, \qquad 2\varphi_0^2\left(1-\frac{1}{m}\right),$$

respectively.

In this example, the bootstrap independently samples $z_{it}^* \sim N(\overline{z}_i, \hat{\varphi})$. The associated maximum-likelihood estimators are $\overline{z}_i^*$ and

$$\hat{\varphi}^* = \frac{1}{nm}\sum_{i=1}^{n}\sum_{t=1}^{m}(z_{it}^* - \overline{z}_i^*)^2.$$

Conditional on the data, the latter estimator follows the same Gamma distribution as above, only with φ_0 replaced by $\hat{\varphi}$. Noting that we can write $\sqrt{nm}(\hat{\varphi}-\varphi_0) = -\sqrt{n/m}\,\varphi_0 + \epsilon$, for a mean-zero random variable $\epsilon = O_P(1)$, this implies that

$$\sqrt{nm}(\hat{\varphi}^* - \hat{\varphi}) \sim \text{Gamma}\left(-\left(\sqrt{nm}\varphi_0 - \sqrt{\frac{n}{m}}\varphi_0 + \epsilon\right), \frac{n(m-1)}{2}, \frac{2\varphi_0}{\sqrt{nm}}\left(1-\frac{1}{m}\right)+\frac{2\epsilon}{nm}\right)$$

conditional on the sample. Its mean and variance are

$$-\sqrt{\frac{n}{m}}\varphi_0 + \frac{1}{m}\left(\sqrt{\frac{n}{m}}\varphi_0 - \epsilon\right), \qquad 2\varphi_0^2\left(1-\frac{2}{m}+\frac{1}{m^2}\right)+O_P\left(\frac{1}{m}\right),$$

which, to first order, agree with the corresponding moments of the maximum-likelihood estimator.

The studentized maximum-likelihood estimator follows a (translated) inverse-Gamma distribution, mirrored about the origin. Moreover,

$$-\sqrt{nm}\,\frac{(\hat{\varphi}-\varphi_0)}{\sqrt{2\hat{\varphi}^2}} \sim \text{Inverse-Gamma}\left(-\sqrt{\frac{nm}{2}}, \frac{n(m-1)}{2}, \sqrt{\frac{nm}{2}}\frac{nm}{2}\right).$$

This distribution is pivotal, and the bootstrap replicates it exactly. Thus, at least in this example, the studentized bootstrap yields confidence intervals whose probability of covering φ_0 can be controlled exactly.

A first-order correction to $\hat{\varphi}$ based on a plug-in estimator of its asymptotic bias is

$$\check{\varphi} \coloneqq \hat{\varphi} + \frac{\hat{\varphi}}{m}.$$

It is interesting to compare the performance of confidence intervals for φ_0 based on bias correction with those obtained via the bootstrap. The bias-correction approach uses the large-sample approximation

$$\sqrt{nm}\,\frac{(\check{\varphi}-\varphi_0)}{\sqrt{2\hat{\varphi}^2}} \overset{L}{\to} N(0,1).$$

Its coverage accuracy can be evaluated for any given sample size from the observation that

$$-\sqrt{nm}\,\frac{(\check{\varphi}-\varphi_0)}{\sqrt{2\hat{\varphi}^2}} \sim \text{Inverse-Gamma}\left(-\sqrt{\frac{nm}{2}}\left(1+\frac{1}{m}\right), \frac{n(m-1)}{2}, \sqrt{\frac{nm}{2}}\,\frac{nm}{2}\right).$$

Notice that this distribution coincides with that of the studentized maximum-likelihood estimator up to the location parameter; the current distribution being located closer to zero. An alternative in this particular example is to studentize the bias-corrected estimator using $\sqrt{2\check{\varphi}^2}$. We find that

$$-\sqrt{nm}\,\frac{(\check{\varphi}-\varphi_0)}{\sqrt{2\check{\varphi}^2}} \sim \text{Inverse-Gamma}\left(-\sqrt{\frac{nm}{2}}, \frac{n(m-1)}{2}, \sqrt{\frac{nm}{2}}\,\frac{nm}{2}\left(\frac{m}{m+1}\right)\right).$$

Here, there is no change in the location parameter (compared to maximum likelihood) but, rather, in the scale parameter. This, then, affects the entire shape of the sampling distribution.

To simplify the presentation we use the shorthand notation

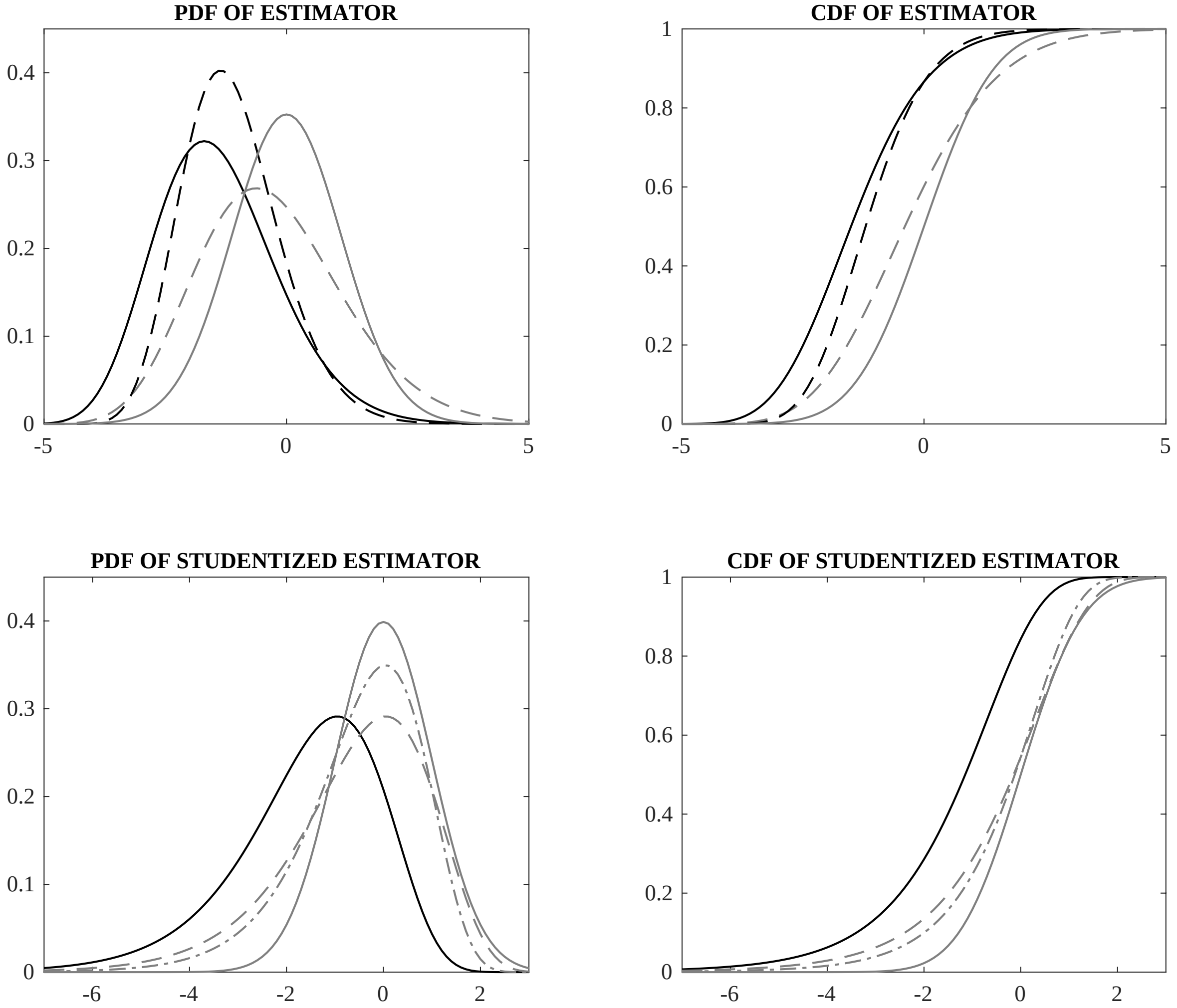
$$\hat{e} := \sqrt{nm}(\hat{\varphi}-\varphi_0), \qquad \hat{s} := 2^{-1/2}\,\hat{e}/\hat{\varphi},$$

for the (scaled) sampling error of the maximum-likelihood estimator and for its studentized version, respectively. The bootstrap quantities $\hat{e}^*$ and $\hat{s}^*$ are defined analogously. We similarly let

$$\check{e} := \sqrt{nm}(\check{\varphi}-\varphi_0), \qquad \check{s} := 2^{-1/2}\,\check{e}/\hat{\varphi}, \qquad \tilde{s} := 2^{-1/2}\,\check{e}/\check{\varphi},$$

for the bias-corrected estimator. The left and right plots in Figure 1 contain, respectively, the density and distribution functions of these quantities for $(n,m)=(10,5)$ and $\varphi_0 = 1$. The solid black curves refer to $\hat{e}$. The dashed black curves capture the behavior of $\hat{e}^*$ up to first order (i.e., by setting $\epsilon = 0$, thereby ignoring the randomness induced by its dependence

Figure 1: Many normal means: Sampling densities and distributions



Upper panel: Density functions (left plot) and cumulative distributions (right plot) of $\hat{e}$ (solid black curve), $\hat{e}^*$ (dashed black curve), and $\check{e}$ (dashed grey curve), together with the normal density with zero mean and variance $2\varphi_0^2$ (solid grey curve). Lower panel: Density functions (left plot) and cumulative distributions (right plot) of $\hat{s}$ and $\hat{s}^*$ (solid black curve), and $\check{s}$ (dashed grey curve), and $\tilde{s}$ (dashed-dotted grey curve), along with the standard-normal density (solid grey curve). Plots generated with $\varphi_0 = 1$ and $(n, m) = (10, 5)$.

on the original sample). The solid grey curves, in turn, refer to a mean-zero normal variable with variance $2\varphi_0^2$ while the dashed grey curves depict $\check{e}$, the analytically bias-corrected estimator. Here, the distribution of $\hat{e}^*$ does not have quite enough mass in the left tail,

Table 1: Many normal means: Coverage of two-sided 95% confidence intervals

n	m	MLE	BC1	BC2	BB	SB
10	10	0.765	0.871	0.897	0.918	0.950
20	10	0.682	0.868	0.897	0.918	0.950
40	10	0.535	0.864	0.894	0.916	0.950
100	10	0.235	0.854	0.887	0.911	0.950

compared to the distribution of $\hat{e}$, but mimics the right-tail well. The sampling distribution of $\check{e}$, compared to that of $\hat{e}$, is closer to the normal reference distribution but the sample size is not sufficiently large for the distribution to resemble well its normal approximation. The lower plots in Figure 1 provide corresponding results for the studentized estimators. All these distributions are pivotal and, hence, independent of φ_0. Here, $\hat{s}$ and $\hat{s}^*$ follow exactly the same distribution; it is given by the solid black curve. The dashed grey curves for $\check{s}$ are the same as those for $\hat{s}$ (and $\hat{s}^*$) up to a translation that brings them closer to the standard-normal reference curves (in solid grey). The distribution of $\check{s}$ has considerable excess mass in its left tail so that confidence intervals constructed by treating it as standard normal will be too short. By using an unbiased estimator of the asymptotic variance, $\tilde{s}$ reduces this issue somewhat and yields a sampling distribution that is closer to the normal benchmark.

To complement this graphical illustration, Table 1 gives coverage rates of two-sided 95% confidence intervals for φ_0 across different sample sizes. These rates are invariant to the value of φ_0. The conclusions from the graphical analysis are borne out in the table. The naive normal approximation (MLE) does poorly when applied to maximum likelihood but bootstrapping the maximum-likelihood estimator, both using the basic bootstrap (BB) and the studentized bootstrap (SB), yields reliable inference. Here, the latter gives exact coverage but this will not be true in general. Both bootstrap procedures perform better in terms of coverage than those based on bias correction. The table also confirms the improved approximation of $\tilde{s}$ (BC2) by a standard-normal random variable relative to $\check{s}$

(BC1).

Dynamic logit For our next example we consider the Markov process

$$y_{it} = \begin{cases} 1 & \text{if } \eta_{i0} + \varphi_0 y_{it-1} > \varepsilon_{it} \\ 0 & \text{if not} \end{cases},$$

where the ε_{it} are independent and identically distributed logistic random variables, i.e., $\mathbb{P}(\varepsilon_{it} \leq a) = (1 + e^{-a})^{-1} =: F(a)$. The initial conditions, y_{i0}, are observed and held fixed throughout.

In this example the maximum-likelihood estimator is not available in closed form. Nonetheless, the log-likelihood function is globally concave and numerical optimization via a Newton-Raphson procedure is straightforward (see the next section for details). Given $\hat{\varphi}$ and $\hat{\eta}_1, \ldots, \hat{\eta}_n$ we generate bootstrap samples by recursively drawing y^*_{it} from a Bernoulli distribution with success probability $F(\hat{\eta}_i + \hat{\varphi} y^*_{it-1})$.

The exact distribution of $\hat{\varphi}$ is not known so we resort to simulations. We draw y_{i0} from its stationary distribution,

$$\mathbb{P}(y_{i0} = 1) = \frac{F(\eta_{i0})}{1 - F(\eta_{i0} + \varphi_0) + F(\eta_{i0})},$$

set $\eta_{i0} = 0$ for all the strata, and consider $\varphi_0 \in \{1/2, 1\}$. Table 2 provides the coverage rate of (two-sided) 95% confidence intervals for the autoregressive parameter together with their average length. Results are reported for confidence intervals based on (the naive large-sample approximation to) maximum likelihood (MLE), on the basic bootstrap and studentized bootstrap (BB and SB, respectively), as well as on two procedures that adjust the maximum-likelihood estimator for its bias. The first of these adjustments (BC1) is the analytical correction of Hahn and Kuersteiner (2011). The second adjustment (BC2) is due to Fernández-Val (2009) and exploits the model structure to implement a refined correction that replaces certain sample averages by expected quantities. Both these approaches require a bandwidth choice. We report results for a bandwidth equal to one, which we found was the choice that performed best here. The bootstrap results, in turn, are based on the use

Table 2: Dynamic logit: Properties of two-sided 95% confidence intervals

			COVERAGE					LENGTH				
φ_0	n	m	MLE	BC1	BC2	BB	SB	MLE	BC1	BC2	BB	SB
½	100	10	0.117	0.940	0.970	0.970	0.930	0.567	0.572	0.574	0.629	0.542
½	100	20	0.381	0.958	0.965	0.956	0.951	0.378	0.381	0.381	0.395	0.372
½	250	10	0.001	0.887	0.953	0.963	0.907	0.358	0.362	0.363	0.397	0.344
½	250	20	0.046	0.932	0.949	0.952	0.943	0.239	0.241	0.241	0.250	0.236
1	100	10	0.095	0.878	0.933	0.957	0.907	0.605	0.620	0.623	0.656	0.577
1	100	20	0.329	0.921	0.944	0.953	0.944	0.404	0.410	0.410	0.418	0.398
1	250	10	0.001	0.699	0.891	0.955	0.893	0.383	0.392	0.394	0.413	0.365
1	250	20	0.027	0.866	0.910	0.965	0.943	0.255	0.259	0.259	0.264	0.252

of 999 bootstrap replications. The results in the table are based on 5,000 Monte Carlo replications.

The naive normal approximation to the sampling distribution of the maximum-likelihood estimator again yields unreliable inference in this problem. Bias correction yields a large improvement in coverage rates and comes with only minor increases in the length of the confidence intervals (which is informative about efficiency). Confidence intervals based on the correction underlying BC2 tend to give better coverage than those based on BC1, with the difference sometimes being considerable (up to 20 percentage points). This highlights the sensitivity of bias-corrected inference to how the bias is being estimated; this is an issue not accounted for by first-order theory. The bootstrap, rather than estimating the bias, mimics it. Both BB and SB are competitive with bias correction, doing at least as well as BC2 in terms of coverage. The studentized bootstrap yields shorter confidence intervals but, for $m = 10$, this comes at the cost of some undercoverage. This problem is essentially resolved for $m = 20$.

Many normal means (cont'd) In our third and final example we reconsider the setup of Neyman and Scott (1948) but change the parameter of interest to

$$\Delta = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \eta_{i0}^2,$$

the second moment of the fixed effects. The plug-in estimator is ${}^1\!/\!_n \sum_{i=1}^n \overline{z}_i^2$. Using the fact that $\overline{z}_i \sim N(\eta_{i0}, {}^{\varphi_0}\!/\!_m)$ by normality of the data it is easy to verify that the plug-in bias due to the estimation of the fixed effects is ${}^{\varphi_0}\!/\!_m$, while the estimator's sampling variance is

$$\frac{2\varphi_0}{nm}\left(2\frac{\sum_{i=1}^n \eta_{i0}^2}{n} + \frac{\varphi_0}{m}\right).$$

The second component in the expression of the variance is of smaller order and not picked up by our general expression for σ^2 given previously.

The exact distribution of the estimator is a complicated mixture and so we again resort to simulations to evaluate the performance of the bootstrap. In our simulations we set $\eta_{i0} = i/n$ so that, in large samples, the distribution of the fixed effects is uniform on $[0,1]$; hence, $\Delta = {}^1\!/\!_3$. Data were generated with $\varphi_0 = 1$. We report results for several choices of (n, m) in Table 3. The bootstrap confidence intervals are again found to yield a large improvement in coverage relative to the ones based on the naive plug-in approach. Again the basic bootstrap does slightly better than the studentized version. The average length of the former's confidence intervals co-incide (up to the fourth decimal digit) with those of maximum likelihood.

4 A note on implementation

In most applications the bootstrap distribution is unknown and needs to be simulated. This, in turn, requires computation of the maximum-likelihood estimator many times. In spite of the presence of a large number of fixed effects, a standard Newton-Raphson procedure is feasible here by exploiting the sparsity of the Hessian matrix. Furthermore, as many popular fixed-effect specifications involve log-likelihood functions that are globally

Table 3: Many normal means: Properties of two-sided 95% confidence intervals for $\lim_{n\to\infty} {}^1\!/\!_n \sum_{i=1}^n \eta_{i0}^2$

		COVERAGE			LENGTH		
n	m	MLE	BB	SB	MLE	BB	SB
50	10	0.545	0.945	0.917	0.232	0.232	0.210
50	20	0.704	0.958	0.934	0.156	0.156	0.145
50	50	0.787	0.946	0.926	0.095	0.095	0.091
100	10	0.258	0.969	0.924	0.164	0.163	0.147
100	20	0.490	0.956	0.933	0.110	0.110	0.102
100	50	0.718	0.935	0.921	0.067	0.067	0.064

concave, such an algorithm is numerically stable and requires only few iterations to locate the global maximizer.

Collect all parameters in $\theta \coloneqq (\varphi, \eta_1, \ldots, \eta_n)$. A Newton step starting at θ is of the form

$$\theta - \ell_{\theta\theta}^{-1}\, \ell_\theta,$$

where ℓ_θ and $\ell_{\theta\theta}$ are the score vector and Hessian matrix. The Hessian matrix is large and so direct inversion can be both slow and numerically inaccurate. Fortunately, the Hessian has a particular block structure. Moreover,

$$\ell_\theta = \begin{pmatrix} \ell_\varphi \\ \ell_{\eta_1} \\ \ell_{\eta_2} \\ \vdots \\ \ell_{\eta_n} \end{pmatrix} \qquad \ell_{\theta\theta} = \begin{pmatrix} \ell_{\varphi\varphi} & \ell_{\varphi\eta_1} & \ell_{\varphi\eta_2} & \cdots & \ell_{\varphi\eta_n} \\ \ell_{\eta_1\varphi} & \ell_{\eta_1\eta_1} & 0 & \cdots & 0 \\ \ell_{\eta_2\varphi} & 0 & \ell_{\eta_2\eta_2} & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \ell_{\eta_n\varphi} & 0 & 0 & \cdots & \ell_{\eta_n\eta_n} \end{pmatrix},$$

where the individual components are

$$\ell_\varphi \coloneqq \sum_{i=1}^n \sum_{t=1}^m \frac{\partial \ell(\varphi, \eta_i | z_{it})}{\partial \varphi}, \qquad \ell_{\eta_i} \coloneqq \sum_{t=1}^m \frac{\partial \ell(\varphi, \eta_i | z_{it})}{\partial \eta_i},$$
$$\ell_{\varphi\varphi} \coloneqq \sum_{i=1}^n \sum_{t=1}^m \frac{\partial^2 \ell(\varphi, \eta_i | z_{it})}{\partial \varphi \partial \varphi'}, \qquad \ell_{\eta_i\eta_i} \coloneqq \sum_{t=1}^m \frac{\partial^2 \ell(\varphi, \eta_i | z_{it})}{\partial \eta_i \partial \eta_i'},$$

and

$$\ell_{\varphi\eta_i} := \sum_{t=1}^{m} \frac{\partial^2 \ell(\varphi, \eta_i | z_{it})}{\partial\varphi\partial\eta_i'} = \ell_{\eta_i\varphi}'.$$

By making use of partitioned-invere formulae we arrive at an expression for $\ell_{\theta\theta}^{-1}$ that can be computed by using only the inverses of the substantially smaller matrices $\ell_{\varphi\varphi}$ and $\ell_{\eta_i\eta_i}$. With

$$\ell_{\theta\theta}^{-1} = \begin{pmatrix} (\ell_{\theta\theta}^{-1})_{\varphi\varphi} & (\ell_{\theta\theta}^{-1})_{\varphi\eta_1} & (\ell_{\theta\theta}^{-1})_{\varphi\eta_2} & \cdots & (\ell_{\theta\theta}^{-1})_{\varphi\eta_n} \\ (\ell_{\theta\theta}^{-1})_{\eta_1\varphi} & (\ell_{\theta\theta}^{-1})_{\eta_1\eta_1} & (\ell_{\theta\theta}^{-1})_{\eta_1\eta_2} & \cdots & (\ell_{\theta\theta}^{-1})_{\eta_1\eta_n} \\ (\ell_{\theta\theta}^{-1})_{\eta_2\varphi} & (\ell_{\theta\theta}^{-1})_{\eta_2\eta_1} & (\ell_{\theta\theta}^{-1})_{\eta_2\eta_2} & \ddots & (\ell_{\theta\theta}^{-1})_{\eta_2\eta_n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ (\ell_{\theta\theta}^{-1})_{\eta_n\varphi} & (\ell_{\theta\theta}^{-1})_{\eta_n\eta_1} & (\ell_{\theta\theta}^{-1})_{\eta_n\eta_2} & \cdots & (\ell_{\theta\theta}^{-1})_{\eta_n\eta_n} \end{pmatrix},$$

we have

$$(\ell_{\theta\theta}^{-1})_{\varphi\varphi} := \left(\ell_{\varphi\varphi} - \sum_{i=1}^{n} \ell_{\varphi\eta_i}\, \ell_{\eta_i\eta_i}^{-1}\, \ell_{\eta_i\varphi}\right)^{-1}, \qquad (\ell_{\theta\theta}^{-1})_{\varphi\eta_i} := -(\ell_{\theta\theta}^{-1})_{\varphi\varphi}\ell_{\varphi\eta_i}\ell_{\eta_i\eta_i}^{-1} = (\ell_{\theta\theta}^{-1})_{\eta_i\varphi}',$$

and, treating the cases where $i = j$ and $i \neq j$ separately for clarity,

$$(\ell_{\theta\theta}^{-1})_{\eta_i\eta_i} := \ell_{\eta_i\eta_i}^{-1} + \ell_{\eta_i\eta_i}^{-1}\, \ell_{\eta_i\varphi}\, (\ell_{\theta\theta}^{-1})_{\varphi\varphi}\, \ell_{\varphi\eta_i}\, \ell_{\eta_i\eta_i}^{-1} \qquad (\ell_{\theta\theta}^{-1})_{\eta_i\eta_j} := \ell_{\eta_i\eta_i}^{-1}\, \ell_{\eta_i\varphi}\, (\ell_{\theta\theta}^{-1})_{\varphi\varphi}\, \ell_{\varphi\eta_j}\, \ell_{\eta_j\eta_j}^{-1}.$$

The Newton step for φ then simply is

$$\varphi - (\ell_{\theta\theta}^{-1})_{\varphi\varphi}\, \ell_\varphi - \sum_{i=1}^{n} (\ell_{\theta\theta}^{-1})_{\varphi\eta_i}\, \ell_{\eta_i} = \varphi - (\ell_{\theta\theta}^{-1})_{\varphi\varphi} \left(\ell_\varphi - \sum_{i=1}^{n} \ell_{\varphi\eta_i}\, \ell_{\eta_i\eta_i}^{-1}\, \ell_{\eta_i}\right).$$

The corresponding step for each fixed effect η_i is

$$\eta_i - (\ell_{\theta\theta}^{-1})_{\eta_i\varphi}\, \ell_\varphi - \sum_{j=1}^{n} (\ell_{\theta\theta}^{-1})_{\eta_i\eta_j}\, \ell_{\eta_j} = \eta_i - \ell_{\eta_i\eta_i}^{-1} \left(\ell_{\eta_i} - \ell_{\eta_i\varphi}(\ell_{\theta\theta}^{-1})_{\varphi\varphi} \left(\ell_\varphi - \sum_{j=1}^{n} \ell_{\varphi\eta_j}\, \ell_{\eta_j\eta_j}^{-1}\, \ell_{\eta_j}\right)\right).$$

A Newton-Raphson algorithm that uses these updating formulae is feasible even in large data sets. The size of the matrices to be inverted is independent of the sample size. The computational complexity is, therefore, comparable to that of the setting without fixed effects.

5 Asymptotic theory

Our results hold under a set of assumptions that are standard in the literature. The following formulation is mostly borrowed from Kim and Sun (2016). It differs from Hahn and Kuersteiner (2011) in two respects that are worth noting. The first difference is that the individual time series need not be stationary. This is useful because the requirement that the initial condition is a draw from the steady-state distribution, for example, is often hard to justify. The second difference is that certain requirements are assumed to hold uniformly over a neighborhood of the true parameter value. This is useful for the derivation of our results because, like Kim and Sun (2016), we adopt a technique introduced in Andrews (2005) to obtain these. This technique is to first demonstrate a convergence result for the maximum-likelihood estimator uniformly over a set around the true parameter value. Then, as consistency implies that the maximum-likelihood estimator lies in this set with probability approaching one, this allows us the establish the corresponding property for the bootstrap estimator.

In the assumptions (and in the proofs) it is important to make clear under which data generating process certain expectations and probabilities are being computed. We will write $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$ for expectations and probabilities involving data that were generated using parameters $\theta = (\varphi, \eta_1, \ldots, \eta_n)$. Note that some objects, such as $\mathbb{E}_\theta(z_{it})$, only depend on a subset of the elements of θ. For simplicity, however, we do not make this explicit in the notation.

Denote by V_φ and V_η the parameter space for φ and η_i, respectively. Then the parameter space for θ is the Cartesian product $\mathit{\Theta} \coloneqq V_\varphi \times V_\eta \times \cdots \times V_\eta$. We let $\mathit{\Theta}_0$ be a subset of $\mathit{\Theta}$.

Assumption 1.

(i) The function f is continuous in $\varphi \in V_\varphi$ and $\eta_i \in V_\eta$.

(ii) The true parameter value lies in the interior of $\mathit{\Theta}_0$, a subset of the compact set $\mathit{\Theta}$.

For our next assumption, consider the mixing coefficients

$$a_i(\theta, h) \coloneqq \sup_{1 \leq t \leq m} \sup_{A \in \mathcal{A}_{it}(\theta)} \sup_{B \in \mathcal{B}_{it+h}(\theta)} |\mathbb{P}_\theta(A \cap B) - \mathbb{P}_\theta(A)\, \mathbb{P}_\theta(B)|,$$

where $\mathcal{A}_{it}(\theta)$ and $\mathcal{B}_{it}(\theta)$ are the sigma algebras generated by the sequences $z_{it}, z_{it-1}, \ldots$ and $z_{it}, z_{it+1}, \ldots$ when these sequences were generated from our model with the parameter equal to θ.

We will also make use of an open set that covers Θ_0. This set is of the form

$$\Theta_1 := \{\theta \in \Theta : d(\theta, \Theta_0) < \delta\}$$

for some $\delta > 0$, where $d(\theta, \Theta_0) := \inf\{\|\theta - \vartheta\|_2 : \vartheta \in \Theta_0\}$, i.e., the distance between the point θ and the set Θ_0.

Assumption 2. $\sup_{1 \leq i \leq n} \sup_{\theta \in \Theta_1} a_i(\theta, h) = O(r^h)$ *for some constant* $0 < r < 1$.

The next assumption collects smoothness conditions and moment requirements.

Assumption 3.
(i) The function $\ell(\varphi, \eta_i | z_{it})$ *is four times continuously-differentiable in* φ *and* η_i.

(ii) The function $\ell(\varphi, \eta_i | z_{it})$ *and all its cross-derivatives up to fourth order are bounded by a function* $b(z_{it})$ *for which*

$$\sup_{1 \leq i \leq n} \sup_{1 \leq t \leq m} \sup_{\theta \in \Theta_1} \mathbb{E}_\theta(|b(z_{it})|^q) < \infty$$

for some q *such that* $3 + {}^{(\dim(\varphi) + \dim(\eta_i))}\!/_2 < qs$ *with* $0 < s < {}^1\!/_{10}$.

(iii) As $m \to \infty$, ${}^1\!/_m \sum_{t=1}^m \mathbb{E}_\theta(b(z_{it}))$ *converges to* $\lim_{m \to \infty} {}^1\!/_m \sum_{t=1}^m \mathbb{E}_\theta(b(z_{it}))$ *uniformly in* i *and* $\theta \in \Theta_1$.

Let

$$G_i(\varphi, \eta_i | \vartheta) := \lim_{m \to \infty} \frac{1}{m} \sum_{t=1}^m \mathbb{E}_\vartheta(\ell(\varphi, \eta_i | z_{it})).$$

The next assumption ensures that our parameters are identified from time series variation.

Assumption 4. *For each* $\varepsilon > 0$ *there exists a* $\delta_\varepsilon > 0$ *such that*

$$\inf_{1 \leq i \leq n} \inf_{\theta \in \Theta_1} \left(G_i(\varphi, \eta_i | \theta) \; - \sup_{\{(\bar{\varphi}, \bar{\eta}_i) : \|(\bar{\varphi}, \bar{\eta}_i) - (\varphi, \eta_i)\|_2 > \varepsilon\}} G_i(\bar{\varphi}, \bar{\eta}_i | \theta) \right) > \delta_\varepsilon.$$

Assumption 5 states that we are working under rectangular-array asymptotics.

Assumption 5. *As $n, m \to \infty$, ${}^{n}\!/\!{}_{m} \to \gamma^2$ for some $0 < \gamma < \infty$.*

The last assumption ensures a well-defined asymptotic variance for $\hat{\varphi}$. We write $\Omega_{nm,\theta}$ for the matrix defined below (1.1) to highlight its dependence on θ.

Assumption 6. *There exist positive finite constants ϵ_1, ϵ_2 and $\varepsilon_1, \varepsilon_2$ such that, for n and m large enough,*

(i) $$\epsilon_1 \leq \inf_{1 \leq i \leq n} \inf_{\theta \in \Theta_1} \operatorname{mineig} \left(\frac{1}{m} \sum_{t=1}^{m} \mathbb{E}_\theta \left(\frac{\partial^2 \ell(\varphi, \eta_i | z_{it})}{\partial \eta_i \partial \eta_i'} \right) \right)$$
$$\leq \sup_{1 \leq i \leq n} \sup_{\theta \in \Theta_1} \operatorname{maxeig} \left(\frac{1}{m} \sum_{t=1}^{m} \mathbb{E}_\theta \left(\frac{\partial^2 \ell(\varphi, \eta_i | z_{it})}{\partial \eta_i \partial \eta_i'} \right) \right) \leq \epsilon_2,$$

(ii) $\varepsilon_1 < \inf_{\theta \in \Theta_1} \operatorname{mineig}(\Omega_{nm,\theta}) \leq \sup_{\theta \in \Theta_1} \operatorname{maxeig}(\Omega_{nm,\theta}) < \varepsilon_2$.

Our main result is stated in the following theorem.

Theorem 1. *Let Assumptions 1–6 hold. Then*

$$\mathbb{P} \left(\sup_a \left| \mathbb{P}^*(\sqrt{nm}(\hat{\varphi}^* - \hat{\varphi}) \leq a) - \mathbb{P}(\sqrt{nm}(\hat{\varphi} - \varphi) \leq a) \right| > \varepsilon \right) = o(1)$$

for any $\varepsilon > 0$.

Theorem 1 justifies the use of the basic bootstrap for inference.

Next, let $\hat{\Sigma} := \hat{\Omega}_{nm}^{-1}$ where

$$\hat{\Omega}_{nm} := -\frac{1}{nm} \sum_{i=1}^{n} \sum_{t=1}^{m} \left(\frac{\partial^2 \ell(\hat{\varphi}, \hat{\eta}_i | z_{it})}{\partial \varphi \partial \varphi'} - \hat{\rho}_{i,m} \frac{\partial^2 \ell(\hat{\varphi}, \hat{\eta}_i | z_{it})}{\partial \eta_i \partial \varphi'} \right)$$

is the plug-in estimator of Ω_{nm} based on the maximum-likelihood estimator, and we used

$$\hat{\rho}_{i,m} := \left(\frac{1}{m} \sum_{t=1}^{m} \frac{\partial^2 \ell(\hat{\varphi}, \hat{\eta}_i | z_{it})}{\partial \varphi \partial \eta_i'} \right) \left(\frac{1}{m} \sum_{t=1}^{m} \frac{\partial^2 \ell(\hat{\varphi}, \hat{\eta}_i | z_{it})}{\partial \eta_i \partial \eta_i'} \right)^{-1}.$$

A consistency result for this estimator, as well as for its bootstrap counterpart, is given next.

Theorem 2. *Let Assumptions 1–6 hold. Then $\hat{\Sigma} \xrightarrow{P} \Sigma$ and $\hat{\Sigma}^* \xrightarrow{P^*} \Sigma$.*

Both theorems, when taken together, justify an application of the bootstrap to standardized quantities such as the Wald statistic, for example.

Conclusion

The purpose of this paper has been to show that in panel data models with fixed effects, inference based on the bootstrap remains valid under rectangular-array asymptotics. Our results cover quite general nonlinear models and allow for dynamics in the outcome of interest.

The main advantage of the bootstrap is that it avoids the need to correct for the asymptotic bias in the limit distribution of the maximum-likelihood estimator. It is unlikely that, in our context, the bootstrap yields asymptotic refinements in general as the presence of bias renders the limit distribution non-pivotal, even after studentization. It could be of interest to investigate whether refinements can be obtained by combining the bootstrap with bias correction. On the other hand, the parametric bootstrap we consider is restricted to the correctly-specified likelihood setting. While this is arguably the default for nonlinear panel problems, some of the approaches to bias correction can be generalized to other settings, such as partial likelihoods. In related work, Gonçalves and Kaffo (2015) have shown that a version of the wild bootstrap replicates the bias in the setup of Hahn and Kuersteiner (2002). However, their approach is residual-based and is tailored quite specifically to the linear model.

While our attention has been devoted to one-way models, we see no reason why our main message would not carry over to models with two-way fixed effects. The available results on the behavior of the maximum-likelihood estimator of such models are more restrictive, however, in that they impose additive or multiplicative restrictions on the way the fixed effects enter the likelihood; see Fernández-Val and Weidner (2016) for bias expressions (and corrections) in such a setting. In the same way, two-step estimators such as those considered by Fernández-Val and Vella (2011) (to deal with, e.g., the issue of sample selection) should also be amenable to bootstrapping.

Appendix

Proof of Theorem 1. Note that

$$\mathbb{P}\left(\sup_a \left|\mathbb{P}^*(\sqrt{nm}(\hat{\varphi}^* - \hat{\varphi}) \leq a) - \mathbb{P}(\sqrt{nm}(\hat{\varphi} - \varphi_0) \leq a)\right| > \varepsilon\right)$$

is bounded from above by

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(\sup_a \left|\mathbb{P}_{\hat{\theta}}(\sqrt{nm}(\hat{\varphi}^* - \hat{\varphi}) \leq a) - \mathbb{P}_\theta(\sqrt{nm}(\hat{\varphi} - \varphi) \leq a)\right| > \varepsilon\right)$$

which, in turn, is below

$$\begin{aligned} &\sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(\sup_a \left|\mathbb{P}_\theta(\sqrt{nm}(\hat{\varphi} \ \ - \varphi) \leq a) - \mathbb{P}_\theta(v_\theta \leq a)\right| > \frac{\varepsilon}{2}\right) \\ +&\sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(\sup_a \left|\mathbb{P}_{\hat{\theta}}(\sqrt{nm}(\hat{\varphi}^* - \hat{\varphi}) \leq a) - \mathbb{P}_\theta(v_\theta \leq a)\right| > \frac{\varepsilon}{2}\right). \end{aligned} \tag{A.1}$$

Here and later, we let

$$v_\theta \sim N(\gamma\beta_\theta, \Sigma_\theta)$$

for β_θ and Σ_θ the asymptotic bias and asymptotic variance of the maximum-likelihood estimator for data generated with parameter θ. Therefore, it suffices to show that each of the terms in (A.1) is $o(1)$.

In the supplement we show that

$$\sup_{\theta\in\Theta_1} \left|\mathbb{P}_\theta(\sqrt{nm}(\hat{\varphi} \ \ - \varphi) \leq a) - \mathbb{P}_\theta(v_\theta \leq a)\right| = o(1)$$

for any a. Further, because the normal distribution is a continuous function, we have that

$$\sup_{\theta\in\Theta_1} \left(\sup_a \left|\mathbb{P}_\theta(\sqrt{nm}(\hat{\varphi} \ \ - \varphi) \leq a) - \mathbb{P}_\theta(v_\theta \leq a)\right|\right) = o(1) \tag{A.2}$$

by Polya's theorem. This allows us to invoke Lemma A.1 of Andrews (2005) to establish that

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(\sup_a \left|\mathbb{P}_\theta(\sqrt{nm}(\hat{\varphi} \ \ - \varphi) \leq a) - \mathbb{P}_\theta(v_\theta \leq a)\right| > \frac{\varepsilon}{2}\right) = o(1).$$

This handles the first term in (A.1).

Moving on to the second term in (A.1), note that

$$\begin{aligned}
&\sup_{\theta\in\Theta_0}\mathbb{P}_\theta\left(\sup_a\left|\mathbb{P}_{\hat\theta}(\sqrt{nm}(\hat\varphi^*-\hat\varphi)\le a)-\mathbb{P}_\theta(v_\theta\le a)\right|>\frac{\varepsilon}{2}\right)\\
\le&\sup_{\theta\in\Theta_0}\mathbb{P}_\theta\left(\sup_a\left|\mathbb{P}_{\hat\theta}(\sqrt{nm}(\hat\varphi^*-\hat\varphi)\le a)-\mathbb{P}_{\hat\theta}(v_{\hat\theta}\le a)\right|>\frac{\varepsilon}{4}\right)\\
+&\sup_{\theta\in\Theta_0}\mathbb{P}_\theta\left(\sup_a\left|\mathbb{P}_{\hat\theta}(v_{\hat\theta}\le a)-\mathbb{P}_\theta(v_\theta\le a)\right|>\frac{\varepsilon}{4}\right).
\end{aligned}$$

Here, using (A.2), coupled with the consistency result

$$\sup_{\theta\in\Theta_1}\mathbb{P}(\|\hat\theta-\theta\|_2>\epsilon)=o(1) \tag{A.3}$$

(which follows from Theorem 1 of Kim and Sun 2016), by another application of Lemma A.1 of Andrews (2005),

$$\sup_{\theta\in\Theta_0}\mathbb{P}_\theta\left(\sup_a\left|\mathbb{P}_\theta(\sqrt{nm}(\hat\varphi^*-\hat\varphi)\le a)-\mathbb{P}_{\hat\theta}(v_{\hat\theta}\le a)\right|>\frac{\varepsilon}{4}\right)=o(1)$$

while, again using (A.3),

$$\sup_{\theta\in\Theta_0}\mathbb{P}_\theta\left(\sup_a\left|\mathbb{P}_{\hat\theta}(v_{\hat\theta}\le a)-\mathbb{P}_\theta(v_\theta\le a)\right|>\frac{\varepsilon}{4}\right)=o(1)$$

follows from the continuous mapping theorem. This takes care of the second term in (A.1) and completes the proof of the theorem. □

Proof of Theorem 2. We introduce the notational shorthand

$$V_{it}:=\begin{pmatrix}V_{it}^{11} & V_{it}^{12}\\ V_{it}^{21} & V_{it}^{22}\end{pmatrix}=\begin{pmatrix}\frac{\partial^2\ell(\varphi,\eta_i|z_{it})}{\partial\varphi\partial\varphi'} & \frac{\partial^2\ell(\varphi,\eta_i|z_{it})}{\partial\varphi\partial\eta_i'}\\ \frac{\partial^2\ell(\varphi,\eta_i|z_{it})}{\partial\eta_i\partial\varphi'} & \frac{\partial^2\ell(\varphi,\eta_i|z_{it})}{\partial\eta_i\partial\eta_i'}\end{pmatrix},$$

where the derivatives are evaluated at the parameter values that were used to generate the data. In the same manner, we write the plug-in estimator constructed using $\hat\varphi,\hat\eta_i$ as $\hat V_{it}$. Then

$$\Omega_{nm,\theta}=-\frac{1}{nm}\sum_{i=1}^n\sum_{t=1}^m\left(\mathbb{E}_\theta(V_{it}^{11})-\left(\frac{1}{m}\sum_{t=1}^m\mathbb{E}_\theta(V_{it}^{12})\right)\left(\frac{1}{m}\sum_{t=1}^m\mathbb{E}_\theta(V_{it}^{22})\right)^{-1}\mathbb{E}_\theta(V_{it}^{21})\right),$$

and its plug-in estimator is

$$\hat{\varOmega}_{nm,\theta} \coloneqq -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{t=1}^{m}\hat{V}_{it}^{11} - \left(\frac{1}{m}\sum_{t=1}^{m}\hat{V}_{it}^{12}\right)\left(\frac{1}{m}\sum_{t=1}^{m}\hat{V}_{it}^{22}\right)^{-1}\frac{1}{m}\sum_{t=1}^{m}\hat{V}_{it}^{21}\right).$$

To show Theorem 2 it suffices to establish that, for all $\varepsilon > 0$,

$$\sup_{\theta\in\varTheta_1}\mathbb{P}_\theta\left(\max_{1\le i\le n}\left\|\frac{1}{m}\sum_{t=1}^{m}(\hat{V}_{it}^{11} - \mathbb{E}_\theta(V_{it}^{11}))\right\|_2 > \varepsilon\right) = o(1),$$

$$\sup_{\theta\in\varTheta_1}\mathbb{P}_\theta\left(\max_{1\le i\le n}\left\|\frac{1}{m}\sum_{t=1}^{m}(\hat{V}_{it}^{12} - \mathbb{E}_\theta(V_{it}^{12}))\right\|_2 > \varepsilon\right) = o(1),$$

$$\sup_{\theta\in\varTheta_1}\mathbb{P}_\theta\left(\max_{1\le i\le n}\left\|\frac{1}{m}\sum_{t=1}^{m}(\hat{V}_{it}^{22} - \mathbb{E}_\theta(V_{it}^{22}))\right\|_2 > \varepsilon\right) = o(1).$$

We can then use Lemma A.1 of Andrews (2005) to verify the consistency of both $\hat{\varSigma}$ and $\hat{\varSigma}^*$ as stated in the theorem. The proof for each of the four terms is similar and so we only provide details for the first of them.

To begin we note that

$$\sup_{\theta\in\varTheta_1}\mathbb{P}_\theta\left(\max_{1\le i\le n}\left\|\frac{1}{m}\sum_{t=1}^{m}(\hat{V}_{it}^{11} - \mathbb{E}_\theta(V_{it}^{11}))\right\|_2 > \varepsilon\right)$$

is bounded from above by

$$\sup_{\theta\in\varTheta_1}\mathbb{P}_\theta\left(\max_{1\le i\le n}\left\|\frac{1}{m}\sum_{t=1}^{m}(\hat{V}_{it}^{11} - V_{it}^{11})\right\|_2 > \frac{\varepsilon}{2}\right) + \sup_{\theta\in\varTheta_1}\mathbb{P}_\theta\left(\max_{1\le i\le n}\left\|\frac{1}{m}\sum_{t=1}^{m}(V_{it}^{11} - \mathbb{E}_\theta(V_{it}^{11}))\right\|_2 > \frac{\varepsilon}{2}\right).$$

To deal with the first of these terms let $\tilde{V}_{it}^{111}$ be the vector that collects all third-order derivatives with respect to φ and let $\tilde{V}_{it}^{112}$ denote derivatives with respect to φ (twice) and η_i. The tilde is used to indicate that these derivatives are evaluated at values $(\tilde{\varphi}, \tilde{\eta}_i)$ that (elementwise) lie between $(\hat{\varphi}, \hat{\eta}_i)$ and (φ, η_i). A mean-value expansion around (φ, η_i) yields

$$\begin{aligned}\left\|\frac{1}{m}\sum_{t=1}^{m}(\hat{V}_{it}^{11} - V_{it}^{11})\right\|_2 &\le \frac{1}{m}\sum_{t=1}^{m}\left\|\hat{V}_{it}^{11} - V_{it}^{11}\right\|_2 \\ &\le \frac{1}{m}\sum_{t=1}^{m}\left\|\tilde{V}_{it}^{111}\right\|_2 \|\hat{\varphi} - \varphi\|_2 + \frac{1}{m}\sum_{t=1}^{m}\left\|\tilde{V}_{it}^{112}\right\|_2 \|\hat{\eta}_i - \eta_i\|_2 \\ &\le \frac{1}{m}\sum_{t=1}^{m}\left\|\tilde{V}_{it}^{111}\right\|_1 \|\hat{\varphi} - \varphi\|_2 + \frac{1}{m}\sum_{t=1}^{m}\left\|\tilde{V}_{it}^{112}\right\|_1 \|\hat{\eta}_i - \eta_i\|_2.\end{aligned}$$

The uniform bound on the derivatives in Assumption 3(ii) implies that

$$\frac{1}{m}\sum_{t=1}^{m}\left\|\tilde{V}_{it}^{111}\right\|_1 \lesssim \frac{1}{m}\sum_{t=1}^{m} b(z_{it}),$$
$$\frac{1}{m}\sum_{t=1}^{m}\left\|\tilde{V}_{it}^{112}\right\|_1 \lesssim \frac{1}{m}\sum_{t=1}^{m} b(z_{it}),$$

where $A \lesssim B$ indicates that there exists a finite constant c such that $A \leq c\, B$. Therefore,

$$\max_{1\leq i\leq n}\left\|\frac{1}{m}\sum_{t=1}^{m}(\hat{V}_{it}^{11} - V_{it}^{11})\right\|_2 \lesssim \left(\max_{1\leq i\leq n}\frac{1}{m}\sum_{t=1}^{m} b(z_{it})\right)\left(\|\hat{\varphi}-\varphi\|_2 + \max_{1\leq i\leq n}\|\hat{\eta}_i - \eta_i\|_2\right).$$

Now, the mixing conditions in Assumption 2 and the moment conditions on the bounding function b in Assumption 3(iii) imply that

$$\sup_{\theta\in\Theta_1}\mathbb{P}_\theta\left(\max_{1\leq i\leq n}\left|\frac{1}{m}\sum_{t=1}^{m}(b(z_{it}) - \mathbb{E}_\theta(b(z_{it})))\right| > \varepsilon\right) = o(1)$$

by an application of Lemma 1 of Hahn and Kuersteiner (2011) (which is easily extended to our setting; see the supplement). Also, ${}^1\!/\!{}_m \sum_{t=1}^{m}\mathbb{E}_\theta(b(z_{it}))$ converges to its limit uniformly over Θ_1 by Assumption 3(iv). At the same time, by Theorem 1 in Kim and Sun (2016) we have that

$$\sup_{\theta\in\Theta_1}\mathbb{P}_\theta\left(\|\hat{\varphi}-\varphi\|_2 > \varepsilon\right) = o(1), \qquad \sup_{\theta\in\Theta_1}\mathbb{P}_\theta\left(\max_{1\leq i\leq n}\|\hat{\eta}_i - \eta_i\|_2 > \varepsilon\right) = o(1).$$

Taken together these results yield

$$\sup_{\theta\in\Theta_1}\mathbb{P}_\theta\left(\max_{1\leq i\leq n}\left\|\frac{1}{m}\sum_{t=1}^{m}(\hat{V}_{it}^{11} - V_{it}^{11})\right\|_2 > \frac{\varepsilon}{2}\right) = o(1)$$

follows. Next, again by Assumptions 2 and 3, an application of (a uniform version of) Lemma 3 of Hahn and Kuersteiner (2011) (see the supplement) gives

$$\sup_{\theta\in\Theta_1}\mathbb{P}_\theta\left(\max_{1\leq i\leq n}\left\|\frac{1}{m}\sum_{t=1}^{m}(V_{it}^{11} - \mathbb{E}_\theta(V_{it}^{11}))\right\|_2 > \frac{\varepsilon}{2}\right) = o(1).$$

Hence,

$$\sup_{\theta\in\Theta_1}\mathbb{P}_\theta\left(\max_{1\leq i\leq n}\left\|\frac{1}{m}\sum_{t=1}^{m}(\hat{V}_{it}^{11} - \mathbb{E}_\theta(V_{it}^{11}))\right\|_2 > \varepsilon\right) = o(1),$$

and the proof is complete. □

References

Andrews, D. W. K. (2005). Higher-order improvements of the parametric bootstrap for Markov processes. In D. W. K. Andrews and J. H. Stock (Eds.), *Identification and Inference for Econometric Models*, Chapter 9, pp. 171–215. Cambridge University Press.

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

Arellano, M. and J. Hahn (2007). Understanding bias in nonlinear panel models: Some recent developments. In R. Blundell, W. K. Newey, and T. Persson (Eds.), *Advances In Economics and Econometrics*, Volume III. Econometric Society: Cambridge University Press.

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.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap methods and their application.* Cambridge University Press.

Dhaene, G. and K. Jochmans (2015a). Profile-score adjustments for incidental-parameter problems. Mimeo.

Dhaene, G. and K. Jochmans (2015b). Split-panel jackknife estimation of fixed-effect models. *Review of Economic Studies 82*, 991–1030.

Dhaene, G. and K. Jochmans (2016). Likelihood inference in an autoregression with fixed effects. *Econometric Theory 32*, 1178–1215.

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 (2016). Individual and time effects in nonlinear panel models with large N, T. *Journal of Econometrics 192*, 291–312.

Gonçalves, S. and M. Kaffo (2015). Bootstrap inference for linear dynamic panel data models with individual fixed effects. *Journal of Econometrics 186*, 407–426.

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 (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.

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

Kim, M. S. and Y. Sun (2016). Bootstrap and k-step bootstrap bias corrections for the fixed effects estimator in nonlinear panel data models. *Econometric Theory 32*, 1523–1568.

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.

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

Prentice, R. and L. Gloeckler (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics 34*, 57–67.