Individual and time effects in nonlinear panel models with large N, T

Fernández-Val, Iván

North-Holland

I Fernández-Val, M Weidner. 2016. "Individual and time effects in nonlinear panel models with large N, T." Journal of Econometrics, Volume 192, Issue 1, pp. 291 - 312. https://doi.org/10.1016/j.jeconom.2015.12.014
https://hdl.handle.net/2144/30895

Boston University
Individual and time effects in nonlinear panel models with large $N, T$

Iván Fernández-Val$^{a,*}$, Martin Weidner$^{b,c}$

$^a$Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215–1403, USA
$^b$Department of Economics, University College London, Gower Street, London WC1E 6BT, UK
$^c$Centre for Microdata Methods and Practice, Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE, UK

**A R T I C L E I N F O**

Article history:
Received 15 July 2014
Received in revised form 30 November 2015
Accepted 3 December 2015
Available online 16 January 2016

JEL classification:
C13
C23

Keywords:
Panel data
Nonlinear model
Dynamic model
Asymptotic bias correction
Fixed effects
Time effects

**A B S T R A C T**

We derive fixed effects estimators of parameters and average partial effects in (possibly dynamic) nonlinear panel data models with individual and time effects. They cover logit, probit, ordered probit, Poisson and Tobit models that are important for many empirical applications in micro and macroeconomics. Our estimators use analytical and jackknife bias corrections to deal with the incidental parameter problem, and are asymptotically unbiased under asymptotic sequences where $N/T$ converges to a constant. We develop inference methods and show that they perform well in numerical examples.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Fixed effects estimators of nonlinear panel data models can be severely biased because of the incidental parameter problem (Neyman and Scott, 1948). A growing literature, surveyed in Arellano and Hahn (2007), shows that the leading term of an asymptotic expansion of the bias as both the cross-sectional dimension $N$ and time series dimension $T$ of the panel grow, can be characterized and corrected. In models with individual effects, the leading bias term is of order $1/T$ and comes from the estimation of the individual effects. This result, however, does not apply to models with individual and time effects, where both of these effects are treated as parameters to be estimated. In this paper we show that the estimation of the time effects causes an additional incidental parameter bias of order $1/N$. Thus, if $N$ and $T$ are similarly large, the bias produced by the estimation of the time effects is of similar order of magnitude to the bias produced by the estimation of the individual effects, and both biases need to be corrected. We provide the corresponding analytical and jackknife bias corrections.

The asymptotic approximation to the fixed effects estimators that lets the two dimensions of the panel to grow with the sample size is motivated by the recent availability of long panels and other large pseudo-panel data structures where the indices might not correspond to individuals and time periods. Examples of these data sets include traditional microeconomic panel surveys with a long history of data such as the PSID and NLSY, international cross-country panels such as the Penn World Table, U.S. state level panels over time such as the CPS, and square pseudo-panels of trade flows across countries such as the Feenstra’s World Trade Flows and CEPII, where the indices correspond to the same countries indexed as importers and exporters.
We focus on semi-parametric models with log-likelihood functions that are concave in all parameters, and where each individual effect $\alpha_i$ and time effect $\gamma_t$ enter the log-likelihood for observation $(i, t)$ additively as $\alpha_i + \gamma_t$. This is the most common specification for the individual and time effects in linear models and is also a natural specification in the nonlinear models that we consider. Imposing concavity of the log-likelihood function greatly facilitates showing consistency in our setting where the dimension of the parameter space grows with the sample size. The most popular limited dependent variable models, including logit, probit, ordered probit, Tobit and Poisson models have concave log-likelihood functions, possibly after reparametrization (Olsen, 1978, and Pratt, 1981). We note here that the general expansion that we derive in Appendix B do not impose additivity and concavity, but we use these restrictions to apply the expansion to fixed effects estimators. The models that we consider are semi-parametric because the joint distribution of the explanatory variables and the unobserved effects is left unspecified. The explanatory variables can be either strictly exogenous or predetermined.

We derive bias expansions and corrections for fixed effects estimators of common parameters $\beta$ and average partial effects (APEs). The vector $\beta$ includes all the unknown parameters that enter the log-likelihood function other than the individual and time effects, such as index coefficients in a probit model. The APEs are functions of the data, the common parameters, and the individual and time effects in nonlinear models. We find that the properties of the fixed effects estimators of $\beta$ and the APEs are different. For $\beta$, the order of the bias is $1/T + 1/N$, which is of the same as the rate of convergence $1/\sqrt{NT}$ under sequences where $N/T$ converges to a constant. For the APEs, we uncover that the incidental parameter problem is negligible asymptotically because the order of the bias, $1/N + 1/T$, is smaller than the rate of convergence, which is $1/\sqrt{N} + 1/\sqrt{T}$, slower than for model parameters. To the best of our knowledge, this rate result is new for fixed effects estimators of average partial effects in nonlinear panel models with individual and time effects. In numerical examples we find that the bias corrections, while not necessary to center the asymptotic distribution of APE estimators, do improve their finite-sample properties, specially in dynamic models.

The bias correction eliminates the bias terms of orders $1/T$ and $1/N$ from the fixed effects estimators. We consider two methods to implement the correction: an analytical bias correction similar to Hahn and Newey (2004) and Hahn and Kuersteiner (2011), and a suitable modification of the split panel jackknife of Dhaene and Jochmans (2015). However, the theory of the previous papers does not cover the models that we consider, because, in addition to not allowing for time effects, it assumes either identical distribution or stationarity over time for the processes of the observed variables, conditional on the unobserved effects. These assumptions are violated in our models due to the presence of the time effects, so we need to adjust the asymptotic theory accordingly. The individual and time effects introduce strong correlation in both dimensions of the panel. Conditional on the unobserved effects, we impose cross-sectional independence and weak time-serial dependence, and we allow for heterogeneity in both dimensions.

Simulation evidence indicates that our corrections improve the estimation and inference performance of the fixed effects estimators of parameters and average effects. The analytical corrections dominate the jackknife corrections in a probit model for sample sizes that are relevant for empirical practice. In the online supplement (see Appendix E), we illustrate the corrections with an empirical application on the relationship between competition and innovation using a panel of U.K. industries, following Aghion et al. (2005). We find that the inverted-U pattern relationship found by Aghion et al. is robust to relaxing the strict exogeneity assumption of competition with respect to the innovation process and to the inclusion of innovation dynamics. We also uncover substantial positive state dependence in the innovation process.

**Literature review.** The Neyman and Scott incidental parameter problem has been extensively discussed in the econometric literature; see, for example, Heckman (1981), Lancaster (2000), and Greene (2004). There is also a vast literature that shows how to tackle the problem in specific models under asymptotic sequences where $T$ is fixed and $N$ grows to infinity. However, there are results, e.g. from Honoré and Tamer (2006), Chamberlain (2010), and Chernozhukov et al. (2013), showing that model parameters and APEs are not point identified in important nonlinear panel data models under fixed-$T$ asymptotic sequences, implying that no fixed-$T$ consistent point estimators exist in these models.

A recent response to the incidental parameter problem is to adopt an alternative asymptotic approximation where both $N$ and $T$ grow with the sample size. Under these large-$T$ sequences, the fixed effects estimator is consistent but has bias in the asymptotic distribution. This asymptotic bias is the large-$T$ version of the incidental parameter problem and has motivated the development of bias corrections. Examples of papers that use this approximation include Phillips and Moon (1999), Hahn and Kuersteiner (2002), Lancaster (2002), Woutersen (2002), Alvarez and Arellano (2003), Hahn and Newey (2004), Carro (2007), Arellano and Bonhomme (2009), Fernández-Val (2009), Hahn and Kuersteiner (2011), Fernández-Val and Vella (2011) and Kato et al. (2012). This previous work, however, does not cover models with time effects. Our contribution to this literature is to extend the large-$T$ bias corrections to models with two-way unobserved effects such as the individual and time effects commonly included in linear models.

The large-$T$ panel literature on models with both individual and time effects is sparse. Pesaran (2006), Bai (2009) and Moon and Weidner (forthcoming, 2015b) study linear regression models with interactive individual and time fixed effects. The fixed effects estimators in these models also have asymptotic bias of order $1/T + 1/N$, but the methods used to derive this bias rely on linearity and therefore cannot be applied to the nonlinear models that we consider. Hahn and Moon (2006) consider bias corrected fixed effects estimators in panel linear autoregressive models with additive individual and time effects. Regarding non-linear models, there is independent and contemporaneous work by Charbonneau (2012, 2014), which extends the conditional fixed effects estimators to logit and Poisson models with individual and time effects. She differences out the individual and time effects by conditioning on sufficient statistics. The conditional approach completely eliminates the asymptotic bias coming from the estimation of the incidental parameters, but it does not permit estimation of average partial effects and has not been developed for models with predetermined regressors. We instead consider estimators of model parameters and average partial effects in nonlinear models with predetermined regressors. The two approaches can therefore be considered as complementary.

---

1 Galvao and Kato (2014) also found slow rates of convergence for fixed effects estimators in linear models with individual effects under misspecification. Fernández-Val and Lee (2013) pointed out this issue in nonlinear models with only individual effects.

2 A similar split panel jackknife bias correction method was outlined in Hu (2002).

3 An exception is Woutersen (2002), which considers a special type of grouped time effects whose number is fixed with $T$. We instead consider an unrestricted set of $T$ time effects, one for each time period.
Outline of the paper. The rest of the paper is organized as follows. Section 2 introduces the model and fixed effects estimators. Section 3 describes the bias corrections to deal with the incidental parameter problem and illustrates how the bias corrections work through an example. Section 4 provides the asymptotic theory. Section 5 presents Monte Carlo results. The Appendix collects the proofs of the main results, and an online supplement to the paper contains additional technical derivations, numerical examples, and an empirical application (Fernández-Val and Weidner, 2015b) (see Appendix E).

2. Model and estimators

2.1. Model

The data consist of $N \times T$ observations $\{(Y_{it}, X_{it})' : 1 \leq i \leq N, 1 \leq t \leq T\}$, for a scalar outcome variable of interest $Y_{it}$ and a vector of explanatory variables $X_{it}$. We assume that the outcome for individual $i$ at time $t$ is generated by the sequential process:

$$Y_{it} | X_{it}, \alpha, \beta, \gamma \sim f_{Y}(· | X_{it}, \alpha, \gamma, \beta),$$

$$\{i = 1, \ldots, N; t = 1, \ldots, T\},$$

where $X_{it} = (X_{1it}, \ldots, X_{Nit})$, $\alpha = (\alpha_1, \ldots, \alpha_N)$, $\gamma = (\gamma_1, \ldots, \gamma_T)$, $f_{Y}$ is a known probability function, and $\beta$ is a finite dimensional parameter vector. The variables $\alpha_i$ and $\gamma_i$ are unobserved individual and time effects that in economic applications capture individual heterogeneity and aggregate shocks, respectively. The model is semiparametric because we do not specify the distribution of these effects nor their relationship with the explanatory variables. The conditional distribution $f_{Y}$ represents the parametric part of the model. The vector $X_{it}$ contains predetermined variables with respect to $Y_{it}$. Note that $X_{it}$ can include lags of $Y_{it}$ to accommodate dynamic models.

We consider two running examples throughout the analysis:

Example 1 (Binary Response Model). Let $Y_{it}$ be a binary outcome and $F$ be a cumulative distribution function, e.g. the standard normal or standard logistic distribution. We can model the conditional distribution of $Y_{it}$ using the single-index specification with individual and time effects

$$f_{Y}(y | X_{it}, \alpha_i, \gamma_i, \beta) = F\left(X_{it}^{\prime} \beta + \alpha_i + \gamma_i y\right)^y \times \left[1 - F\left(X_{it}^{\prime} \beta + \alpha_i + \gamma_i y\right)^{1-y}\right], \quad y \in [0, 1].$$

In a labor economics application, $Y$ can be an indicator for female labor force participation and $X$ can include fertility indicators and other socio-economic characteristics.

Example 2 (Poisson Model). Let $Y_{it}$ be a non-negative integer-valued outcome, and $f(·; \lambda)$ be the probability mass function of a Poisson random variable with mean $\lambda > 0$. We can model the conditional distribution of $Y_{it}$ using the single index specification with individual and time effects

$$f_{Y}(y | X_{it}, \alpha_i, \gamma_i, \beta) = f(y; \exp(X_{it}^{\prime} \beta + \alpha_i + \gamma_i)), \quad y \in \{0, 1, 2, \ldots\}.$$

In an industrial organization application, $Y$ can be the number of patents that a firm produces and $X$ can include investment in R&D and other firm characteristics.

For estimation, we adopt a fixed effects approach, treating the realization of the unobserved individual and time effects as parameters to be estimated. We collect all these effects in the vector $\phi_{NT} = (\alpha_1, \ldots, \alpha_N, \gamma_1, \ldots, \gamma_T)^\prime$. The model parameter $\beta$ usually includes regression coefficients of interest, while the vector $\phi_{NT}$ is treated as a nuisance parameter. The true values of the parameters, denoted by $\beta^0$ and $\phi_{NT}^0 = (\alpha_1^0, \ldots, \alpha_N^0, \gamma_1^0, \ldots, \gamma_T^0)$, are the solution to the population conditional maximum likelihood problem

$$\max_{(\beta, \phi_{NT}) \in \mathbb{R}^\dim \beta \times \mathbb{R}^\dim \phi_{NT}} \mathbb{E}_{\phi} [\mathcal{L}_{NT}(\beta, \phi_{NT})],$$

$$\mathcal{L}_{NT}(\beta, \phi_{NT}) := (NT)^{-1/2} \left\{ \sum_{i,t} \log f_{Y}(Y_{it} | X_{it}, \alpha_i, \gamma_i, \beta) \right. \left. - b(v_{NT}^0 \phi_{NT}^0)^2 / 2 \right\},$$

(2.1)

for every $N, T$, where $\mathbb{E}_{\phi}$ denotes the expectation with respect to the distribution of the data conditional on the unobserved effects and initial conditions including strictly exogenous variables, $b > 0$ is an arbitrary constant, $v_{NT} = (Y_{it}, -1)^\prime$, and $1_N$ and $1_T$ denote vectors of ones with dimensions $N$ and $T$. Existence and uniqueness of the solution to the population problem will be guaranteed by our assumptions in Section 4, including concavity of the objective function in all parameters. The second term of $\mathcal{L}_{NT}$ is a penalty that imposes a normalization needed to identify $\phi_{NT}$ in models with scalar individual and time effects that enter additively into the log-likelihood function as $\alpha_i + \gamma_i$. In this case, adding a constant to all $\alpha_i$ while subtracting it from all $\gamma_i$ does not change $\alpha_i + \gamma_i$. To eliminate this ambiguity, we normalize $\phi_{NT}^0$ to satisfy $v_{NT}^0 \phi_{NT}^0 = 0$, i.e. $\sum_i \alpha_i^0 = \sum_i \gamma_i^0$. The penalty produces a maximizer of $\mathcal{L}_{NT}$ that is automatically normalized. We could equivalently impose $v_{NT}^0 \phi_{NT}^0 = 0$ as a constraint, but for technical reasons we prefer to work with an unconstrained optimization problem. There are other possible normalizations for $\phi_{NT}$, such as $\alpha_i = 0$. The model parameter $\beta$ is invariant to the choice of normalization, that is, our asymptotic results on the estimator for $\beta$ are independent of this choice of normalization. Our choice is convenient for certain intermediate results that involve the incidental parameter $\phi_{NT}$, its score vector and its Hessian matrix. The pre-factor $(NT)^{-1/2}$ in $\mathcal{L}_{NT}(\beta, \phi_{NT})$ is just a rescaling.

Other quantities of interest involve averages over the data and unobserved effects

$$\delta_{NT}^0 := \mathbb{E} [\Delta_{NT}(\beta^0, \phi_{NT}^0)],$$

$$\Delta_{NT}(\beta, \phi_{NT}) := (NT)^{-1} \sum_{i,t} \Delta(X_{it}, \beta, \alpha_i, \gamma_i),$$

(2.2)

where $\mathbb{E}$ denotes the expectation with respect to the joint distribution of the data and the unobserved effects, provided that the expectation exists. $\delta_{NT}$ is indexed by $N$ and $T$ because the marginal distribution of $(X_{it}, \alpha_i, \gamma_i) : 1 \leq i \leq N, 1 \leq t \leq T$ can be heterogeneous across $i$ and/or $t$; see Section 4.2. These averages include average partial effects (APEs), which are often the ultimate quantities of interest in nonlinear models. The APEs are invariant to the choice of normalization for $\phi_{NT}$ if $\alpha_i$ and $\gamma_i$ enter $\Delta(X_{it}, \beta, \alpha_i, \gamma_i)$ as $\alpha_i + \gamma_i$. Some examples of partial effects that satisfy this condition are the following:

Example 1 (Binary Response Model). If $X_{it,k}$, the $k$th element of $X_{it}$, is binary, its partial effect on the conditional probability of $Y_{it}$ is

$$\Delta(X_{it}, \beta, \alpha_i, \gamma_i) = F(\beta_k + X_{it,k} \beta_{-k} + \alpha_i + \gamma_i).$$

4 In Appendix B we derive asymptotic expansions that apply to general models with multiple unobserved effects. In order to use these expansions to obtain the asymptotic distribution of the panel fixed effects estimators, we need to derive the properties of the expected Hessian of the incidental parameters, a matrix with increasing dimension, and to show the consistency of the estimator of the incidental parameter vector. The additive specification $\alpha_i + \gamma_i$ is useful to characterize the Hessian and we impose strict concavity of the objective function to show the consistency.
\[-F(X'_{it} \beta_k + \alpha_1 + \gamma_1), \tag{2.3}\]

where \( \beta_k \) is the \( k \)th element of \( \beta \), and \( X_{it} \) and \( \beta_k \) include all elements of \( X_{it} \) and \( \beta \) except for the \( k \)th element. If \( X_{it} \) is continuous and \( D \) is differentiable, the partial effect of \( X_{it} \) on the conditional probability of \( Y_{it} \) is

\[\Delta(X_{it}, \beta, \alpha_1, \gamma_1) = \beta_k \partial F(X'_{it} \beta + \alpha_1 + \gamma_1), \tag{2.4}\]

where \( \partial F \) is the derivative of \( F \).

**Example 2 (Poisson Model)** If \( X_0 \) includes \( Z_0 \) and some known transformation \( H(Z_0) \) with coefficients \( \beta_1 \) and \( \beta_2 \), the partial effect of \( Z_0 \) on the conditional expectation of \( Y_0 \) is

\[\Delta(X_0, \beta, \alpha_1, \gamma_1) = [\beta_1 + \beta_2 H(Z_0)] \exp(X'_{0} \beta + \alpha_1 + \gamma_1). \tag{2.5}\]

### 2.2. Fixed effects estimators

We estimate the parameters by solving the sample analog of problem (2.1), i.e.

\[
\text{max}_{(\beta, \phi) \in \text{dim} \beta + \text{dim} \phi_{NT}} \mathcal{L}_{NT} (\beta, \phi_{NT}). \tag{2.6}\]

As in the population case, we shall impose conditions guaranteeing that the solution to this maximization problem exists and is unique with probability approaching one as \( N \) and \( T \) become large. For computational purposes, we note that the solution to the program (2.6) for \( \beta \) is the same as the solution to the program that imposes \( \nu_{NT} = 0 \) directly as a constraint in the optimization, and is invariant to the normalization. In our numerical examples we impose either \( \alpha_1 = 0 \) or \( \gamma_1 = 0 \) directly by dropping the first individual or time effect. This constrained program has good computational properties because its objective function is concave and smooth in all the parameters. We have developed the commands `probit` and `logit` in Stata to implement the methods of the paper for probit and logit models (Cruz-González et al., 2015). \(^5\) When \( N \) and \( T \) are large, e.g., \( N > 2,000 \) and \( T > 50 \), we recommend the use of optimization routines that exploit the sparsity of the design matrix of the model to speed up computation such as the package Speedglm in R (Enea, 2012). For a probit model with \( N = 2,000 \) and \( T = 52 \), Speedglm computes the fixed effects estimator in less than 2 min with a \( 2 \times 2.66 \) GHz 6-Core Intel Xeon processor, more than 7.5 times faster than our Stata command `probit`.\(^6\)

To analyze the statistical properties of the estimator of \( \beta \) it is convenient to first concentrate out the nuisance parameter \( \phi_{NT} \). For given \( \beta \), we define the optimal \( \phi_{NT} \) as

\[
\hat{\phi}_{NT} \beta = \text{argmax}_{\phi_{NT} \in \text{dim} \phi_{NT}} \mathcal{L}_{NT} (\beta, \phi_{NT}). \tag{2.7}\]

The fixed effects estimators of \( \beta \) and \( \phi_{NT} \) are

\[
\hat{\beta}_{NT} = \text{argmax}_{\beta \in \text{dim} \beta} \mathcal{L}_{NT} (\beta, \hat{\phi}_{NT} (\beta)), \quad \hat{\phi}_{NT} = \hat{\phi}_{NT} (\hat{\beta}). \tag{2.8}\]

Estimators of APEs can be formed by plugging-in the estimators of the model parameters in the sample version of (2.2), i.e.

\[
\hat{\gamma}_{NT} = \Delta_{NT} (\hat{\beta}, \hat{\phi}_{NT}). \tag{2.9}\]

Again, \( \hat{\gamma}_{NT} \) is invariant to the normalization chosen for \( \phi_{NT} \) if \( \alpha_1 \) and \( \gamma_1 \) enter \( \Delta(X_{it}, \beta, \alpha_1, \gamma_1) \) as \( \alpha_1 + \gamma_1 \).

---

\(^5\) We refer to this companion work for computational details.

\(^6\) Additional comparisons of computational times are available from the authors upon request.

### 3. Incidental parameter problem and bias corrections

In this section we give a heuristic discussion of the main results, leaving the technical details to Section 4. We illustrate the analysis with numerical calculations based on a variation of the classical Neyman and Scott (1948) variance example.

#### 3.1. Incidental parameter problem

Fixed effects estimators in nonlinear models suffer from the incidental parameter problem (Neyman and Scott, 1948). The source of the problem is that the dimension of the nuisance parameter \( \phi_{NT} \) increases with the sample size under asymptotic approximations where either \( N \) or \( T \) passes to infinity. To describe the problem let

\[
\bar{\beta}_{NT} := \text{argmax}_{\beta \in \text{dim} \beta} \mathbb{E}_\phi \left[ \mathcal{L}_{NT} (\beta, \phi_{NT} (\beta)) \right]. \tag{3.1}\]

The fixed effects estimator is inconsistent under the traditional Neyman and Scott asymptotic sequences where \( N \to \infty \) and \( T \) is fixed, i.e., \( \lim_{N \to \infty} \bar{\beta}_{NT} \neq \beta^0 \). Similarly, the fixed effects estimator is inconsistent under asymptotic sequences where \( T \to \infty \) and \( N \) is fixed, i.e., \( \lim_{T \to \infty} \bar{\beta}_{NT} \neq \beta^0 \). Note that \( \bar{\beta}_{NT} \) is \( \beta^0 \) if \( \phi_{NT} (\beta) \) is replaced by \( \hat{\phi}_{NT} \beta \) directly as a constraint in the optimization, and is invariant to the normalization. In our numerical examples we impose either \( \alpha_1 = 0 \) or \( \gamma_1 = 0 \) directly by dropping the first individual or time effect. This constrained program has good computational properties because its objective function is concave and smooth in all the parameters. We have developed the commands `probit` and `logit` in Stata to implement the methods of the paper for probit and logit models (Cruz-González et al., 2015). \(^5\) When \( N \) and \( T \) are large, e.g., \( N > 2,000 \) and \( T > 50 \), we recommend the use of optimization routines that exploit the sparsity of the design matrix of the model to speed up computation such as the package Speedglm in R (Enea, 2012). For a probit model with \( N = 2,000 \) and \( T = 52 \), Speedglm computes the fixed effects estimator in less than 2 min with a \( 2 \times 2.66 \) GHz 6-Core Intel Xeon processor, more than 7.5 times faster than our Stata command `probit`.\(^6\)

To analyze the statistical properties of the estimator of \( \beta \) it is convenient to first concentrate out the nuisance parameter \( \phi_{NT} \). For given \( \beta \), we define the optimal \( \phi_{NT} \) as

\[
\hat{\phi}_{NT} \beta = \text{argmax}_{\phi_{NT} \in \text{dim} \phi_{NT}} \mathcal{L}_{NT} (\beta, \phi_{NT}). \tag{2.7}\]

The fixed effects estimators of \( \beta \) and \( \phi_{NT} \) are

\[
\hat{\beta}_{NT} = \text{argmax}_{\beta \in \text{dim} \beta} \mathcal{L}_{NT} (\beta, \hat{\phi}_{NT} (\beta)), \quad \hat{\phi}_{NT} = \hat{\phi}_{NT} (\hat{\beta}). \tag{2.8}\]

Estimators of APEs can be formed by plugging-in the estimators of the model parameters in the sample version of (2.2), i.e.

\[
\hat{\gamma}_{NT} = \Delta_{NT} (\hat{\beta}, \hat{\phi}_{NT}). \tag{2.9}\]

Again, \( \hat{\gamma}_{NT} \) is invariant to the normalization chosen for \( \phi_{NT} \) if \( \alpha_1 \) and \( \gamma_1 \) enter \( \Delta(X_{it}, \beta, \alpha_1, \gamma_1) \) as \( \alpha_1 + \gamma_1 \).

---

\(^5\) We refer to this companion work for computational details.

\(^6\) Additional comparisons of computational times are available from the authors upon request.
effects estimator is asymptotically biased because
\[
\sqrt{N}\left(\hat{\beta}_{\text{NJ}} - \beta^0\right) = \sqrt{N}\left(\hat{\beta}_{\text{NJ}} - \hat{\beta}_{\text{NJ}}\right) \\
+ \sqrt{N}\left(\hat{B}_{\text{NI}}/T + \hat{B}_{\text{NI}}/N + o_p(T^{-1} \vee N^{-1})\right) \\
\rightarrow_d \mathcal{N}(\kappa \hat{B}_{\infty} + \kappa^{-1}\hat{B}_{\infty}, \nabla_{\infty}). \tag{3.3}
\]

Relative to fixed effects estimators with only individual effects, the presence of time effects introduces additional asymptotic bias through \(\hat{B}_{\infty}\). This asymptotic result predicts that the fixed effects estimator can have significant bias relative to its dispersion. Moreover, confidence intervals constructed around the fixed effects estimator can severely underestimate the true value of the parameter even in large samples. We show that these predictions provide a good approximations to the finite sample behavior of the fixed effects estimator through analytical and simulation examples in Sections 3.3 and 5.

The analytical bias correction consists of subtracting estimators of the leading terms of the bias from the fixed effect estimator of \(\beta\). Let \(\hat{B}_{\text{NI}}^0\) and \(\hat{B}_{\text{NI}}^0\) be estimators of \(\hat{B}_{\text{NI}}^0\) and \(\hat{B}_{\text{NI}}^0\) as defined in (4.6). The bias corrected estimator can be formed as
\[
\hat{\beta}_{\text{NJ}}^0 = \hat{\beta}_{\text{NJ}} - \hat{B}_{\text{NI}}^0/T - \hat{B}_{\text{NI}}^0/N.
\]

If \(N/T \rightarrow k^2\), \(\hat{B}_{\text{NI}}^0 \rightarrow k\rho_{\infty}\), and \(\hat{D}_{\text{NI}}^0 \rightarrow k\hat{D}_{\infty}\), then
\[
\sqrt{N}(\hat{\beta}_{\text{NJ}}^0 - \beta^0) \rightarrow_d \mathcal{N}(0, \nabla_{\infty}).
\]

The analytical correction therefore centers the asymptotic distribution at the true value of the parameter, without increasing asymptotic variance. This asymptotic result predicts that in large samples the corrected estimator has smaller bias relative to dispersion, the correction does not increase dispersion, and the confidence intervals constructed around the corrected estimator have coverage probabilities close to the nominal levels. We show that these predictions provide a good approximations to the behavior of the corrections in Sections 3.3 and 5 even in small panels with \(N < 60\) and \(T < 15\).

We also consider a jackknife bias correction method that does not require explicit estimation of the bias. This method is based on the split panel jackknife (SPJ) of Dhaene and Jochmans (2015) applied to the time and cross-section dimension of the panel. Alternative jackknife corrections based on the leave-one-observation-out panel jackknife (PJ) of Hahn and Newey (2004) and combinations of PJ and SPJ are also possible. We do not consider corrections based on PJ because they are theoretically justified by second-order expansions of \(\hat{B}_{\text{NJ}}\) that are beyond the scope of this paper.

To describe our generalization of the SPJ, define the fixed effects estimator of \(\beta\) in the subpanel with cross sectional indices \(A\) and time series indices \(B\) as
\[
\hat{\beta}_{A,B} = \arg\max_{\beta \in \mathbb{R}} \max_{\alpha(\beta) \in \mathbb{R}} \sum_{\gamma(B) \in \mathbb{R}} \sum_{i \in A, t \in B} \log f(Y_{it} | X_{it}, \alpha_i, \gamma_t, \beta),
\]

where \(\alpha(A) = (\alpha_i : i \in A)\) and \(\gamma(B) = (\gamma_t : t \in B)\). Let \(\hat{\beta}_{A,B,2}\) be the average of the 2 split jackknife estimators in the subpanels with \(A = \{1, 2, \ldots, N\}\) and \(B = \{T/2 + 1, T/2 + 2, \ldots, T\}\), i.e., including all the individuals and leaving out the first and second halves of the time periods. Let \(\hat{\beta}_{A,B,2}\) be the average of the 2 split jackknife estimators in the subpanels with \(B = \{1, 2, \ldots, T\}\) and \(A = \{1, 2, \ldots, N/2\}\) or \(A = \{N/2 + 1, N/2 + 2, \ldots, N\}\), i.e., including all the time periods and leaving out half of the individuals of the panel. In choosing the cross sectional indexing of the panel, one might want to take into account individual clustering structures and other dependences to preserve them in the SPJ. For example, all the individuals belonging to the same cluster should be indexed such that they remain in the same subpanel after the cross sectional split. If there are no cross sectional dependences, the indexing of the individuals is unrestricted. We recommend to construct \(\hat{\beta}_{N/2,T,2}\) as the average of the estimators obtained from all possible partitions of \(N/2\) individuals to avoid ambiguity and arbitrariness in the choice of the division.\(^9\)

The bias corrected estimator is
\[
\hat{\beta}_{N/2,T,2} = 3\hat{\beta}_{N/2,T} - \hat{\beta}_{N/2,T/2} - \hat{\beta}_{N/2,T}. \tag{3.4}
\]

To give some intuition about how the corrections work, note that
\[
\hat{\beta}_{N/2,T} - \beta_0 = (\hat{\beta}_{N/2,T} - \beta_0) - (\hat{\beta}_{N/2,T} - \hat{\beta}_{N/2,T}) - (\hat{\beta}_{N/2,T} - \beta_0),
\]

where \(\hat{\beta}_{N/2,T} - \hat{\beta}_{N/2,T} = \hat{B}_{\infty}/T + o_p(T^{-1} \vee N^{-1})\) and \(\hat{\beta}_{N/2,T} - \hat{\beta}_{N/2,T} = \hat{D}_{\infty}/N + o_p(N^{-1} \vee N^{-1})\). Relative to \(\hat{\beta}_{N/2,T}, \hat{\beta}_{N/2,T} \) has double the bias coming from the estimation of the individual effects because it is based on subpanels with half of the time periods, and \(\hat{\beta}_{N/2,T} \) has double the bias coming from the estimation of the time effects because it is based on subpanels with half of the individuals. The time series split removes the bias term \(\hat{B}_{\infty}\) and the cross sectional split removes the bias term \(\hat{D}_{\infty}\).

### 3.3. Illustrative example

To illustrate how the bias corrections work in finite samples, we consider a simple model where the solution to the population program (3.1) has closed form. This model corresponds to a variation of the classical Neyman and Scott (1948) variance example that includes both individual and time effects, \(Y_{it} \sim N(\alpha_i + \gamma_t, \beta)\). It is well-known that in this case
\[
\hat{\beta}_{NJ} = (NT)^{-1} \sum_{i,t} (Y_{it} - \bar{Y}_t - \bar{Y}_i + \bar{Y})^2,
\]

where \(\bar{Y}_i = T^{-1} \sum_{t} Y_{it}, \bar{Y}_t = N^{-1} \sum_{i} Y_{it}\), and \(\bar{Y} = (NT)^{-1} \sum_{i,t} Y_{it}\). Moreover, from the well-known results on the degrees of freedom adjustment of the estimated variance
\[
\hat{\beta}_{NJ} = \text{E}_f[\hat{\beta}_{NJ}] = \beta^0 \left(\frac{N-1}{NT} - 1\right),
\]

so that \(\hat{B}_{NJ} = -\beta^0 \hat{D}_{NJ} = -\beta^0\).\(^7\)

To form the analytical bias correction we can set \(\hat{B}_{NJ} = -\hat{\beta}_{NJ}\) and \(\hat{D}_{NJ} = -\hat{\beta}_{NJ}\). This yields \(\hat{\beta}_{NJ} = \hat{\beta}_{NJ}(1 + 1/T + 1/N)\) with
\[
\hat{\beta}_{NJ} = \text{E}_f[\hat{\beta}_{NJ}] = \beta^0 \left(1 - \frac{1}{T^2} - \frac{1}{N^2} - \frac{1}{NT} + \frac{1}{NT^2} + \frac{1}{N^2T}\right).
\]

This correction reduces the order of the bias from \((T^{-1} \vee N^{-1})\) to \((T^{-2} \vee N^{-2})\), and introduces additional higher order terms.

---

\(^7\) When \(T\) is odd we define \(\hat{\beta}_{N/2,T/2}\) as the average of the 2 split jackknife estimators that use overlapping subpanels with \(B = \{1, 2, \ldots, T/2\}\) and \(B = \{T/2 + 1, T/2 + 2, \ldots, T\}\). We define \(\hat{\beta}_{N/2,T/2}\) similarly when \(N\) is odd.

\(^8\) There are \(P = \binom{N}{N/2}\) different cross sectional partitions with \(N/2\) individuals.

\(^9\) Okui (2013) derived the bias of fixed effects estimators of autocovariances and autocorrelations in this model.
The analytical correction increases finite-sample variance because the factor \((1 + 1/T + 1/N) > 1\). We compare the biases and standard deviations of the fixed effects estimator and the corrected estimator in a numerical example below.

For the jackknife correction, straightforward calculations give

\[
\overline{\beta}_{NT} = \frac{1}{2} \left[ \overline{\beta}_{NT} + \overline{\beta}_{NT} \right],
\]

The correction therefore reduces the order of the bias from \((T^{-1} \vee N^{-1})\) to \((TN)^{-1}\).

Table 1 presents numerical results for the bias and standard deviations of the fixed effects and bias corrected estimators in finite samples. We consider panels with \(N, T \in \{10, 25, 50\}\), and only report the results for \(T \leq N\) since all the expressions are symmetric in \(N\) and \(T\). All the numbers in the table are in percentage of the true parameter value, so we do not need to specify the value of \(\beta^0\). We find that the analytical and jackknife corrections offer substantial improvements over the fixed effects estimator in terms of bias. The first and fourth row of the table show that the bias of the fixed effects estimator is of the same order of magnitude as the standard deviation, where \(\overline{\beta}_{NT} = \text{Var}[\overline{\beta}_{NT}] = 2(N - 1)(T - 1)(\beta^0)^2/(NT)^2\) under independence of \(Y_{it}\) over \(i\) and \(t\) conditional on the unobserved effects. The fifth row shows the increase in standard deviation due to analytical bias correction is small compared to the bias reduction, where \(\overline{\beta}_{NT} = \text{Var}[\overline{\beta}_{NT}^0] = (1 + N + 1/T)^2 \overline{\beta}_{NT}\). The last row shows that the jackknife yields less precise estimates than the analytical correction when \(T = 10\).

Table 2 illustrates the effect of the bias on the inference based on the asymptotic distribution. It shows the coverage probabilities of 95% asymptotic confidence intervals for \(\beta^0\) constructed in the usual way as

\[
\text{Cl}_{95}(\hat{\beta}) = \hat{\beta} \pm 1.96 \sqrt{\text{Var}(\hat{\beta})} = \hat{\beta}(1 \pm 1.96 \sqrt{2/(NT)}),
\]

where \(\hat{\beta} = (\hat{\beta}_{NT}, \hat{\beta}_{NT}, \hat{\beta}_{NT})\) and \(\overline{\beta}_{NT} = 2\hat{\beta}^2/(NT)\) is an estimator of the asymptotic variance \(\overline{\beta}_{NT}/\overline{\beta}_{NT} = 2\hat{\beta}^2/(NT)\). To find the coverage probabilities, we use that \(\text{Var}[\hat{\beta}_{NT}/\beta^0] \sim X^2_{N-1}(1-1)\), and \(\hat{\beta}_{NT} = (1 + N + 1/T)^2 \hat{\beta}_{NT}\). These probabilities do not depend on the value of \(\beta^0\) because the limits of the intervals are proportional to \(\beta\). For the Jackknife we compute the probabilities numerically by simulation with \(\beta^0 = 1\). As a benchmark of comparison, we also consider confidence intervals constructed from the unbiased estimator \(\beta_{NT} = NT \beta_{NT} / (N - 1)(T - 1)\). Here we find that the confidence intervals based on the fixed effect estimator display

Table 2

| N = 10 | N = 25 | N = 50 |
|--------|--------|--------|
| N = 10 | N = 25 | N = 50 |
| \(\text{Cl}_{95}(\hat{\beta}_{NT})\) | 1.56 | 1.55 | 1.55 |
| \(\text{CI}_{95}(\hat{\beta}_{NT}^0)\) | 0.89 | 0.92 | 0.92 |
| \(\text{CI}_{95}(\hat{\beta}_{NT})\) | 0.89 | 0.91 | 0.93 |
| \(\text{CI}_{95}(\hat{\beta}_{NT})\) | 0.91 | 0.93 | 0.94 |

Notes: Nominal coverage probability is .95. \(\text{CI}_{95}(\hat{\beta}_{NT}^0)\) obtained by 50,000 simulations with \(\beta^0 = 1\).

4. Asymptotic theory for bias corrections

In nonlinear panel data models the population problem (3.1) generally does not have closed form solution, so we need to rely on asymptotic arguments to characterize the terms in the expansion of the bias (3.2) and to justify the validity of the corrections.

4.1. Asymptotic distribution of model parameters

We consider panel models with scalar individual and time effects that enter the likelihood function additively through \(\alpha_i + \gamma_t\). In these models the dimension of the incidental parameters is \(\dim \phi_{NT} = N + T\). The leading cases are single index models, where the dependence of the likelihood function on the parameters is through an index \(X_{it}\alpha_i + \gamma_t\). These models cover the probit and Poisson specifications of Examples 1 and 2. The additive structure only applies to the unobserved effects, so we can allow for scale parameters to cover the Tobit and negative binomial models. We focus on these additive models for computational and analytical tractability, because we can establish the consistency of the fixed effects estimators under a concavity assumption in the log-likelihood function with respect to all the parameters.

The parametric part of our panel models takes the form

\[
\text{log} f(Y_{it} | \alpha_i, \gamma_t, \beta) = \ell_{\alpha} (\beta, \pi_{\alpha}).
\]

We denote the derivatives of the log-likelihood function \(\ell_{\alpha} \) by \(\partial_{\alpha} \ell_{\alpha} (\beta, \pi_{\alpha}) := \partial_{\alpha} \ell_{\alpha} (\beta, \pi_{\alpha}) : = \partial_{\alpha} ^2 \ell_{\alpha} (\beta, \pi_{\alpha}) / \partial \beta / \partial \beta\) and \(\partial_{\alpha} \ell_{\alpha} (\beta, \pi_{\alpha}) : = \partial_{\alpha} ^2 \ell_{\alpha} (\beta, \pi_{\alpha}) / \partial \beta / \partial \pi_{\alpha}\), q = 1, 2, 3, etc. We drop the arguments \(\beta\) and \(\pi\) when the derivatives are evaluated at the true parameter values \(\beta^0\) and \(\pi_{\alpha}^0\). We drop the dependence on \(NT\) from all the sequences of functions and parameters, e.g. we use \(\mathcal{L}\) for \(\mathcal{L}_{NT}\) and \(\phi\) for \(\phi_{NT}\).

We make the following assumptions:

In this example it is possible to develop higher-order jackknife corrections that completely eliminate the bias because we know the entire expansion of \(\overline{\beta}_{NT}\). For example, \(E_{\alpha}(4\overline{\beta}_{NT} - 2\overline{\beta}_{NT/2} + \overline{\beta}_{NT/4}) = \beta^0\), where \(\overline{\beta}_{NT/2}\) is the average of the four split jackknife estimators that leave out half of the individuals and the first or the second halves of the time periods. See Dhaene and Jochmans (2015) for a discussion on higher-order biases of panel fixed effects estimators.
Assumption 4.1 (Panel Models). Let $v > 0$ and $\mu > 4(8 + v)/v$. Let $\varepsilon > 0$ and let $B^2_\varepsilon$ be a subset of $\mathbb{R}^{\dim(\beta + 1)}$ that contains an $\varepsilon$-neighborhood of $(\beta^0, \sigma^0)$ for all $i, t, N, T$.\footnote{For example, $B^2_\varepsilon$ can be chosen to be the Cartesian product of the $\varepsilon$-ball around $\beta^0$ and the interval $[\pi_{\min}, \pi_{\max}]$, with $\pi_{\min} \leq \pi_{\max} \leq \pi_{\max} + \varepsilon$ for all $i, t, N, T$. We can have $\pi_{\min} = -\infty$ and $\pi_{\max} = \infty$, as long as this is compatible with Assumption 4.1(iv) and (v).}

(i) Asymptotics: we consider limits of sequences where $N/T \to k^2$, $0 < k < \infty$, as $N, T \to \infty$.

(ii) Sampling: conditional on $\phi_i$, $\{(Y^*_i, X^*_i) : 1 \leq i \leq N\}$ is independent across $i$ and, for each $i$, $\{(Y_i, X_i) : 1 \leq t \leq T\}$ is $\alpha$-mixing with mixing coefficients satisfying sup $a_i(m) = \Theta(m^{-r})$ as $m \to \infty$, where

$$a_i(m) := \sup_{T} |P(A \cap B) - P(A)P(B)|,$$

and for $Z_i = (Y_i, X_i)$, $A^i$ is the sigma field generated by $(Z_i, Z_{i-1}, \ldots)$, and $B^2_\varepsilon$ is the sigma field generated by $(Z_i, Z_{i+1}, \ldots)$.

(iii) Model: for $X^*_i = (X_i : s = 1, \ldots, T)$, we assume that for all $i, t, N, T$,

$$Y_i | X^*_i, \phi, \beta \sim \exp\{\ell_i(\beta, \alpha_i, \gamma_i)\}.$$

The realizations of the parameters and unobserved effects that generate the observed data are denoted by $\beta^0$ and $\phi^0$.

(iv) Smoothness and moments: We assume that $\ell_i(\beta, \phi)$ is four times continuously differentiable over $B^2_\varepsilon$ a.s. The partial derivatives of $\ell_i(\beta, \phi)$ with respect to the elements of $\ell_i(\beta, \pi)$ up to fourth order are bounded in absolute value uniformly over $B^2_\varepsilon$ by a function $M(Z_i) > 0$ a.s., and $\max_{i,ij} \phi_i[M(Z_i)^{\beta_{ij}}]$ is a.s. uniformly bounded over $N, T$.

(v) Concavity: For all $N, T, (\beta, \phi) \mapsto \ell_i(\beta, \phi)$ is concave in $(\beta, \alpha_i, \gamma_i)$.

Remark 1 (Assumption 4.1). Assumption 4.1(i) defines the large-$T$ asymptotic framework and is the same as in Hahn and Kuersteiner (2011). The relative rate of $N$ and $T$ exactly balances the order of the bias and variance producing a non-degenerate asymptotic distribution.

Assumption 4.1(ii) does not impose identical distribution nor stationarity over the time series dimension, conditional on the unobserved effects, unlike most of the large-$T$ panel literature, e.g., Hahn and Newey (2004) and Hahn and Kuersteiner (2011). These assumptions are violated by the presence of the time effects, because they are treated as parameters. The mixing condition is used to bound covariances and moments in the application of laws of large numbers and central limit theorems—it could replace by other conditions that guarantee the applicability of these results.

Assumption 4.1(iii) is the parametric part of the panel model. We rely on this assumption to guarantee that $\partial \ell_i/\partial \alpha_i$ and $\partial \ell_i/\partial \gamma_i$ have martingale difference properties. Moreover, we use certain Bartlett identities implied by this assumption to simplify some expressions, but those simplifications are not crucial for our results. We provide expressions for the asymptotic bias and variance that do not apply these simplifications in Remark 3.

Assumption 4.1(iv) imposes smoothness and moment conditions in the log-likehood function and its derivatives. These conditions guarantee that the higher-order stochastic expansions of the fixed effect estimator that we use to characterize the asymptotic bias are well-defined, and that the remainder terms of these expansions are bounded.

The most commonly used nonlinear models in applied economics such as logit, probit, ordered probit, Poisson, and Tobit models have smooth log-likelihoods functions that satisfy the concavity condition of Assumption 4.1(v), provided that all the elements of $X_i$ have cross sectional and time series variation. Assumption 4.1(v) guarantees that $\beta^0$ and $\phi^0$ are the unique solution to the population problem (2.1), that is all the parameters are point identified.

To describe the asymptotic distribution of the fixed effects estimator $\hat{\beta}$, it is convenient to introduce some additional notation. Let $\mathcal{F}$ be the $(N + T) \times (N + T)$ expected Hessian matrix of the log-likelihood with respect to the nuisance parameters evaluated at the true parameters, i.e.,

$$\mathcal{F} = \mathbb{E}_\phi[-\partial \phi \partial \ell \mathbf{L}^\top] = \left( \begin{array}{ll} \mathcal{F}^{(\alpha \alpha)}_{\phi} & \mathcal{F}^{(\alpha \gamma)}_{\phi} \\ \mathcal{F}^{(\gamma \alpha)}_{\phi} & \mathcal{F}^{(\gamma \gamma)}_{\phi} \end{array} \right) + \frac{b}{\sqrt{NT}} \mathbf{v} \mathbf{v}^\top,$$

where $\mathcal{F}^{(\alpha \alpha)}_{\phi} = \mathbb{E}_\phi[-\partial \partial_\alpha \ell_\alpha]/\sqrt{\mathcal{F}^{\gamma \gamma}}$, $\mathcal{F}^{(\alpha \gamma)}_{\phi} = \mathbb{E}_\phi[-\partial \partial_\gamma \ell_\gamma]/\sqrt{\mathcal{F}^{\gamma \gamma}}$, and $\mathcal{F}^{(\gamma \gamma)}_{\phi} = \mathbb{E}_\phi[-\partial \partial_\gamma \ell_\gamma]/\sqrt{\mathcal{F}^{\gamma \gamma}}$. Furthermore, let $\mathcal{F}^{-1}_{\phi \alpha \alpha}, \mathcal{F}^{-1}_{\phi \alpha \gamma}, \mathcal{F}^{-1}_{\phi \gamma \alpha}$, and $\mathcal{F}^{-1}_{\phi \gamma \gamma}$ denote the $N \times N, N \times T, T \times N$ and $T \times T$ blocks of the inverse $\mathcal{F}^{-1}_{\phi}$ of $\mathcal{F}$. We define the dim $\beta$-vector $\mathcal{S}_\beta$ and the operator $D_{\beta \alpha}$ as

$$\mathcal{S}_\beta := -\frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \sum_{i=1}^{N} \left( \mathcal{F}^{(\alpha \alpha)}_{\phi} \partial_\alpha \ell_i + \mathcal{F}^{(\alpha \gamma)}_{\phi} \partial_\gamma \ell_i \right) \mathbb{E}_\phi[-\partial \partial_\beta \ell_\beta] \mathbb{E}_\phi[-\partial \partial_\alpha \ell_\alpha]$$

\begin{equation}
(\alpha_k^*, \gamma_k^*) = \arg \min_{\alpha_k, \gamma_k} \sum_{i,t} \mathbb{E}_\phi[-\partial \partial_\alpha \ell_i] \mathbb{E}_\phi[-\partial \partial_\beta \ell_i] \mathbb{E}_\phi[-\partial \partial_\alpha \ell_i] - \alpha_k - \gamma_k^2
\end{equation}

The operator $D_{\beta \alpha}$ partials out individual and time effects in nonlinear models. It corresponds to individual and time differentiating when the model is linear. To see this, consider the normal linear model $Y_i | X_i, \alpha_i, \gamma_i \sim N(X_i^\top \alpha_i + \gamma_i, 1)$. Then, $\mathcal{S}_\beta = T^{-1} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{E}_\phi[\partial \partial_\beta \ell_i] \mathbb{E}_\phi[-\partial \partial_\alpha \ell_i]$. $D_{\beta \alpha} \mathcal{S}_\beta = \sum_{i=1}^{N} \mathbb{E}_\phi[-\partial \partial_\alpha \ell_i]$. $D_{\beta \alpha} \mathcal{S}_\beta = \sum_{i=1}^{N} \mathbb{E}_\phi[-\partial \partial_\gamma \ell_i]$. $D_{\beta \alpha} \mathcal{S}_\beta = 0$, where $\mathcal{S}_\beta = Y_i - X_i^{\top} \alpha_i - \gamma_i$ and $\mathcal{S}_\beta = \mathcal{S}_\beta - \mathbb{E}_\phi[-\partial \partial_\beta \ell_i] = \mathbb{E}_\phi[-\partial \partial_\gamma \ell_i]$. This is the vector of individual and time demeaned explanatory variables.

The following theorem establishes the asymptotic distribution of the fixed effects estimator $\hat{\beta}$.

**Theorem 4.1 (Asymptotic Distribution of $\hat{\beta}$).** Suppose Assumption 4.1 holds, then the following limits exist

\[ B_{\infty} = \lim_{N,T \to \infty} \left( T \sum_{i=1}^{T} \sum_{i=1}^{N} \mathbb{E}_\phi[-\partial \partial_\alpha \ell_i] D_{\beta \alpha} \mathcal{S}_\beta + \frac{1}{2} \sum_{i=1}^{T} \mathbb{E}_\phi[-\partial \partial_\beta \ell_i] \mathbb{E}_\phi[-\partial \partial_\alpha \ell_i] \right) \]
\[ \bar{B}_\infty = \lim_{n,T \to \infty} \left[ -\frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} \varepsilon_{i} \left( \partial_{\beta} \ell_{i} \beta_{i} + \frac{1}{2} \partial_{\beta} \beta_{i} \ell_{i} \right) \right], \]

\[ \bar{W}_\infty = \lim_{n,T \to \infty} \left[ -\frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \varepsilon_{i} \left( \partial_{\beta \beta} \ell_{i} \beta_{i} \right) - \frac{1}{2} \partial_{\beta \beta} \ell_{i} \varepsilon_{i} \right], \]

and that \( \bar{W}_\infty > 0. \) Then,

\[ \sqrt{NT} (\beta - \beta^0) \to_d \bar{W}_\infty^{-1/2} \mathcal{N}(\kappa \bar{B}_\infty + \kappa^{-1} \bar{D}_\infty, \bar{W}_\infty), \]

so that \( \bar{B}_\infty = \bar{W}_\infty^{-1} \bar{B}_\infty, \bar{D}_\infty ^d = \bar{W}_\infty^{-1} \bar{D}_\infty, \) and \( \bar{W}_\infty = \bar{W}_\infty^{-1} \ln (3.2) \) and (3.3).

Remark 2. The complete proof of Theorem 4.1 is provided in the Appendix. Here we point out why the argument for the consistency proof in models with only individual effects does not apply to our setting, give a heuristic derivation of the asymptotic distribution, and highlight where some of the assumptions are used in the proof.

(i) The consistency proof for models with only individual effects relies on partitioning the log-likelihood in the sum of individual log-likelihoods that depend on a fixed number of parameters, the model parameter \( \beta \) and the corresponding individual effect \( \alpha_i. \) The maximizers of the individual log-likelihood are then consistent estimators of all the parameters as \( T \) becomes large by standard arguments. This approach does not work in models with individual and time effects because there is no partition of the data that is only affected by a fixed number of parameters, and whose size grows with the sample size.

(ii) In the following we give a heuristic discussion of the asymptotic distribution result for \( \beta. \) A first-order Taylor series expansion to approximate the first order conditions of (2.8) around \( \beta^0 \) gives

\[ 0 = \partial_{\beta} \ell(\beta, \hat{\varphi}(\beta)) \approx \partial_{\beta} \ell(\beta^0, \hat{\varphi}^0) \]

\[ -\frac{1}{\bar{W}_\infty} \sqrt{NT} (\beta - \beta^0), \]

where \( \hat{\varphi}^0 = \hat{\varphi}(\beta^0). \) A second-order Taylor series expansion to approximate \( \partial_{\beta} \ell(\beta^0, \hat{\varphi}^0) \) around \( \hat{\varphi}^0 \) yields

\[ \partial_{\beta} \ell(\beta^0, \hat{\varphi}^0) \approx \partial_{\beta} \ell(\beta^0, \hat{\varphi}^0) + \partial_{\beta \beta} \ell(\beta^0, \hat{\varphi}^0)[\hat{\varphi}^0 - \varphi^0] \]

\[ + \sum_{g=1}^{\dim \hat{\varphi}} \partial_{\beta \hat{\varphi}^g} \ell(\beta^0, \hat{\varphi}^0)[\hat{\varphi}^0 - \varphi^0][\hat{\varphi}^g - \varphi^g]/2, \]

where the first term has zero mean and determines the asymptotic variance, and the second and third term determine the asymptotic bias. Thus, by the central limit theorem and the information equality,

\[ \partial_{\beta} \ell(\beta^0, \varphi^0) \to_d \mathcal{N}(0, \bar{W}_\infty). \]

The second and third terms satisfy

\[ \partial_{\beta \varphi} \ell(\beta^0, \varphi^0)[\hat{\varphi}^0 - \varphi^0] \]

\[ + \sum_{g=1}^{\dim \hat{\varphi}} \partial_{\beta \hat{\varphi}^g} \ell(\beta^0, \varphi^0)[\hat{\varphi}^0 - \varphi^0][\hat{\varphi}^g - \varphi^g]/2 \]

\[ \approx \sqrt{NT}(\bar{B}_\infty + \bar{D}_\infty)/N, \]

where \( \bar{B}_\infty \) and \( \bar{D}_\infty \) are characterized from a second-order Taylor series expansion to approximate \( \hat{\varphi}^0 \) around \( \varphi^0. \) We refer to the Appendix for details of this derivation. There we show that \( \bar{B}_\infty \) and \( \bar{D}_\infty \) originate from the elements of \( \hat{\varphi}^0 \) corresponding to the individual effects and time effects, respectively. Plugging those results into (4.4), and solving for \( \sqrt{NT}(\beta - \beta^0) \) yields

\[ \sqrt{NT}(\beta - \beta^0) \approx \bar{W}_\infty^{-1/2} \partial_{\beta} \ell(\beta^0, \varphi^0) + \bar{D}_\infty \sqrt{NT}/T \]

\[ + \bar{B}_\infty \sqrt{T}/N \to_d \bar{W}_\infty^{-1/2} \mathcal{N}(\kappa \bar{B}_\infty + \kappa^{-1} \bar{D}_\infty, \bar{W}_\infty). \]

This derivation shows that the source of the bias is that the score \( \partial_{\beta} \ell(\beta, \varphi(\beta)) \) is not centered at zero when \( \beta = \beta^0. \) This problem arises from the substitution of the incidental parameter \( \varphi \) by the sample analog \( \hat{\varphi}^0 \) that has a rate of convergence slower than \( \sqrt{NT}. \) Thus, \( \bar{B}_\infty \) originates from the estimators of the individual effects in \( \varphi, \) which have rate of convergence \( \sqrt{T}; \) whereas \( \bar{D}_\infty \) originates from the estimators of the time effects in \( \varphi, \) which have convergence rate \( \sqrt{N}. \)

(iii) The two key assumptions in the derivation of the asymptotic distribution are the additive separability of \( \alpha_i \) and \( \gamma_i \) in Assumption 4.1(iii) and the concavity in Assumption 4.1(v). We resort to concavity to prove consistency of \( \beta \) and to bound the remainder terms in all the expansions. Additive separability is convenient to characterize the order of the inverse average Hessian, \( \bar{H}, \) defined in (4.2). This inverse Hessian features prominently in the second-order Taylor series expansion of \( \varphi^0 \) around \( \varphi^0 \) used to characterize \( \bar{B}_\infty \) and \( \bar{D}_\infty. \)

It is instructive to evaluate the expressions of the bias in our running examples.

Example 1 (Binary Response Model). In this case

\[ \ell_i(\beta, \pi) = Y_i \log F(X_i \beta + \pi) + (1 - Y_i) \log[1 - F(X_i \beta + \pi)], \]

so that \( \partial_{\beta} \ell_i = H_i \partial_{\beta} X_i, \partial_{\beta} \pi_i = \partial_{\beta} \varphi(X_i), X_i \partial_{\beta} \pi_i = -H_i \partial_{\beta} F_i + \partial_{\beta} H_i, \partial_{\beta} \varphi(X_i), X_i \partial_{\beta} \varphi(X_i), \partial_{\beta \beta} \ell_i = \partial_{\beta \beta} \varphi(X_i), X_i \partial_{\beta \beta} \varphi(X_i), \}

where \( H_i = \partial_{\beta} F_i / \partial_{\beta} F_i \) and \( \partial_{\beta \beta} \varphi(X_i) \partial_{\beta \beta} \varphi(X_i), \}

for any function \( G \) and \( j = 0, 1, 2. \) Substituting these values in the expressions of the bias of Theorem 4.1 yields

\[ \bar{B}_\infty = \lim_{n,T \to \infty} \left[ -\frac{1}{2N} \sum_{i=1}^{N} \sum_{t=1}^{T} \varepsilon_{i} \left[ H_i \partial_{\beta} X_i \right] \right], \]

\[ \bar{D}_\infty = \lim_{n,T \to \infty} \left[ -\frac{1}{2T} \sum_{t=1}^{T} \sum_{i=1}^{N} \varepsilon_{i} \left[ H_i \partial_{\beta} X_i \right] \right], \]

\[ \bar{W}_\infty = \lim_{n,T \to \infty} \left[ -\frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \varepsilon_{i} \left[ \partial_{\beta} \varphi(X_i) \right] \right], \]

where \( \partial_{\beta} \varphi = H_i \partial_{\beta} F_i \) and \( \tilde{X}_i \) is the residual of the population projection of \( X_i \) on the space spanned by the incidental parameters under a metric weighted by \( E_{i} \left( \partial_{\beta} \varphi \right). \) For the probit model where all the components of \( X_i \) are strictly exogenous,
\[ D_\infty = \lim_{N,T \to \infty} \left[ \frac{1}{2T} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{E}_\phi \left[ (\omega_i \hat{X}_D \hat{X}_i' \beta) \right] \right] \beta^0. \]

The asymptotic bias is therefore a positive definite matrix weighted average of the true parameter value as in the case of the probit model with only individual effects \citep{Fernandez-Val2009}.

**Example 2 (Poisson Model).** In this case
\[ \ell_{it}(\beta, \pi) = (X_{it}'\beta + \pi)Y_{it} - \exp(X_{it}'\beta + \pi) - \log Y_{it}! , \]
so that \( \frac{\partial \ell_{it}}{\partial \beta} = Y_{it} - \exp(X_{it}'\beta + \pi) \) and \( \frac{\partial \ell_{it}}{\partial \pi} = Y_{it} - \exp(X_{it}'\beta + \pi) \). Substituting these values in the expressions of the bias of Theorem 4.1 yields
\[ D_\infty = \lim_{N,T \to \infty} \left[ \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{E}_\phi \left[ (Y_{it} - \exp(X_{it}'\beta + \pi))^2 \hat{X}_i \hat{X}_i' \beta \right] \right] , \]
and \( \tilde{D}_\infty = 0 \), where \( \hat{X}_i \) is the residual of the population projection of \( X_i \) on the space spanned by the incidental parameters under a metric weighted by \( \mathbb{E}_\phi (\omega_i) \). If in addition all the components of \( X_{it} \) are strictly exogenous, then we get the no asymptotic bias result \( \tilde{D}_\infty = \tilde{D}_\infty = 0 \).

**Remark 3 (Bias and Variance Expressions for Conditional Moment Models).** In the derivation of the asymptotic distribution, we apply Bartlett identities implied by Assumption 4.1(iii) to simplify the expressions. The following expressions of the asymptotic bias and variance do not make use of these identities and therefore remain valid in conditional moment models that do not specify the entire conditional distribution of \( Y_{it} \):
\[ B_\infty = \lim_{N,T \to \infty} \left[ \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}_\phi \left[ (\frac{\partial \ell_{it}}{\partial \beta} D_{\beta \beta} \ell_{it}) \right] \right] , \]
and \( \tilde{D}_\infty = 0 \), where \( \hat{X}_i \) is the residual of the population projection of \( X_i \) on the space spanned by the incidental parameters under a metric weighted by \( \mathbb{E}_\phi (\omega_i) \). If in addition all the components of \( X_{it} \) are strictly exogenous, then we get the no asymptotic bias result \( \tilde{D}_\infty = \tilde{D}_\infty = 0 \).

For example, consider the Poisson fixed effects estimator in the conditional mean model \( \mathbb{E}[Y_{it} | X_{it}, \phi, \beta] = \omega_i t = \exp(X_{it}'\beta + \alpha_i + \gamma_i) \). Applying the previous expressions to \( \ell_{it}(\beta, \pi) = (X_{it}'\beta + \pi)Y_{it} - \exp(X_{it}'\beta + \pi) - \log Y_{it}! \) yields the same expressions for \( \tilde{D}_\infty, \tilde{D}_\infty, \tilde{W}_\infty \) as in Example 2, and
\[ \tilde{P}_\infty = \lim_{N,T \to \infty} \left[ \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{E}_\phi \left[ (Y_{it} - \exp(X_{it}'\beta + \pi))^2 \hat{X}_i \hat{X}_i' \beta \right] \right] , \]
where \( \hat{X}_i \) is defined as in Example 2. If all the components of \( X_{it} \) are strictly exogenous, then we get again the no asymptotic bias result \( \tilde{D}_\infty = \tilde{D}_\infty = 0 \).

### 4.2. Asymptotic distribution of APEs

In nonlinear models we are often interested in APEs, in addition to model parameters. These effects are averages of the data, parameters and unobserved effects; see expression (2.2). For the panel models of Assumption 4.1 we specify the partial effects as \( \Delta(X_{it}, \beta, \alpha, \gamma) = \Delta(t, \beta, \pi_t) \). The restriction that the partial effects depend on \( \alpha_i \) through \( \pi_{it} \) is natural in our panel models since
\[ \mathbb{E}[Y_{it} | X_{it}, \phi, \beta, \alpha, \gamma] = \int y \exp(\ell_{it}(\beta, \pi_t)) dy , \]
and the partial effects are usually defined as differences or derivatives of this conditional expectation with respect to the components of \( X_{it} \). For example, the partial effects for the binary response and Poisson models described in Section 2 satisfy this restriction.

The distribution of the unobserved individual and time effects is not ancillary for the APEs, unlike for model parameters. We therefore need to make assumptions on this distribution to define and interpret the APEs, and to derive the asymptotic distribution of their estimators. We control the heterogeneity of the partial effects assuming that the individual effects and explanatory variables are identically distributed cross sectionally and/or stationary over time. If \( (X_{it}, \alpha_i, \gamma_i) \) is identically distributed over \( i \) but can be heterogeneously distributed over \( t \), then \( \mathbb{E}[\Delta_{at}] = \delta_{at}^0 \), and \( \delta_{at}^0 = \sum_{i=1}^{N} \mathbb{E}_\phi (\delta_{it}) \). Finally, if \( (X_{it}, \alpha_i, \gamma_i) \) is identically distributed over \( i \) and stationary over \( t \), then \( \mathbb{E}[\Delta_{at}] = \delta_{at}^0 \), and \( \delta_{at}^0 = \delta_{at}^0 \) does not change with \( N \) and \( T \). We also impose smoothness and moment conditions on the function \( \Delta \) that defines the partial effects. We use these conditions to derive higher-order stochastic expansions for the fixed effect estimator of the APEs and to bound the remainder terms in these expansions. Let \( (\alpha_i)_{Nt} := (\alpha_i : 1 \leq i \leq N) \), \( (\gamma_i)_{Nt} := (\gamma_i : 1 \leq t \leq T) \), and \( (X_{it}, \alpha_i, \gamma_i)_{Nt} := ((X_{it}, \alpha_i, \gamma_i) : 1 \leq i \leq N, 1 \leq t \leq T) \).

**Assumption 4.2 (Partial Effects).** Let \( \nu > 0, \epsilon > 0 \), and \( \delta_{at}^0 \) all be as in Assumption 4.1.

(i) **Sampling:** for all \( N, T, (X_{it}, \alpha_i, \gamma_i)_{Nt} \) is identically distributed across \( i \) and/or stationary across \( t \).\footnote{In the working paper version, Fernández-Val and Weidner (2015a), we also consider inference conditional on the unobserved effects by assuming that \( \{\alpha_i\}_N \) and \( \{\gamma_i\}_T \) are deterministic sequences.}

(ii) **Model:** for all \( i, t, N, T \), the partial effects depend on \( \alpha_i \) and \( \gamma_i \) through \( \alpha_i + \gamma_i \):
\[ \Delta(X_{it}, \beta, \alpha_i, \gamma_i) = \Delta(t, \beta, \alpha_i + \gamma_i) , \]
and the realizations of the partial effects are denoted by \( \Delta_{it} := \Delta(t, \beta, \alpha_i + \gamma_i) \).

The realizations of the partial effects are denoted by \( \Delta_{it} := \Delta(t, \beta, \alpha_i + \gamma_i) \).
(iii) Smoothness and moments: The function $(\beta, \pi) \mapsto \Delta_u(\beta, \pi)$ is four times continuously differentiable over $\mathcal{B}_0^2$ a.s. The partial derivatives of $\Delta_u(\beta, \pi)$ with respect to the elements of $(\beta, \pi)$ up to fourth order are bounded in absolute value uniformly over $(\beta, \pi) \in \mathcal{B}_0^2$ by a function $M(Z_u) > 0$ a.s., and $\max_i, j E[|M(Z_u)|^{r+1}]$ is a.s. uniformly bounded over $N, T$.

(iv) Non-degeneracy and moments: $0 < \min_i, (E[\Delta_u^2]) - E[\Delta_u^2] \leq \max_i, j (E[\Delta_u^2]) - E[\Delta_u^2] < \infty$, uniformly over $N, T$.

Analogous to $\mathcal{E}_u$ in Eq. (4.3) we define

$$\Psi_u = -\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \left( \mathcal{P}^{-1}_{(u)} + \mathcal{P}^{-1}_{(u)} \right)_{ij}$$

$$+ \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \left( \mathcal{P}^{-1}_{(u)} + \mathcal{P}^{-1}_{(u)} \right)_{ij} \partial_u \Delta_u,$$

which is the population projection of $\partial_x \Delta_u / \mathbb{E}[\partial_x \ell_u]$. The second term comes from the spatially spanned by the metric given by $\mathbb{E}[\partial_x \ell_u]$ on the space of the derivatives with respect to $\beta$ and higher order derivatives with respect to $\pi$.

Let $\delta_u$ be the APE and its fixed effects estimator, defined as in Eqs. (2.2) and (2.9) with $\Delta(X_u, \beta, \alpha_i, \gamma_i) = \Delta_u(\beta, \alpha_i + \gamma_i)$. The following theorem establishes the asymptotic distribution of $\delta_u$.

**Theorem 4.2 (Asymptotic Distribution of $\delta_u$).** Suppose that the assumptions of Theorems 4.1 and 4.2 hold, and that the following limits exist:

$$\lim_{N,T \to \infty} \left( \partial_u \Delta_u \right)_i = \mathcal{P}^{-1}_{(u)} \sum_{i=1}^N \bigg( \mathbb{E}[\partial_x \ell_u] \bigg)_{ij} + \mathcal{P}^{-1}_{(u)} \mathbb{E}[\partial_x \ell_u]$$

$$\mathbb{E}_u = (\partial_u \Delta_u)^{\mathcal{P}^{-1}}_{(u)} \mathbb{E}_u$$

$$\mathbb{E}_u = (\partial_u \Delta_u)^{\mathcal{P}^{-1}}_{(u)} \mathbb{E}_u$$

$$\lim_{N,T \to \infty} \left( \partial_u \Delta_u \right)_i = \mathcal{P}^{-1}_{(u)} \sum_{i=1}^N \bigg( \mathbb{E}[\partial_x \ell_u] \bigg)_{ij} + \mathcal{P}^{-1}_{(u)} \mathbb{E}[\partial_x \ell_u]$$

for some deterministic sequence $r_{NT} \to \infty$ such that $r_{NT} = O(\sqrt{NT})$ and $\mathbb{V}_{NT}^2 > 0$, where $\Delta_u = \Delta_u - \mathbb{E}(\Delta_u)$ and $\Gamma_u = \mathbb{E}[\partial_u \Delta_u \partial_u \Delta_u^T]$. Then,

$$r_{NT}(\delta - \delta_{NT}) \to_{d} N(0, \mathbb{V}_{NT}^2).$$

**Remark 4 (Convergence Rate, Bias and Variance).** To understand the asymptotic distribution of $\delta$ is useful to decompose

$$\delta - \delta_{NT} = [\delta - \delta] + [\delta - \delta_{NT}],$$

where $\delta := (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T \Delta_u$. In this decomposition the first term captures variation due to parameter estimation, whereas the second term captures variation due to estimation of a population mean by a sample mean. Under Assumption 4.2(iv) the convergence rate $r_{NT}$ is determined by the convergence rate of $\delta - \delta_{NT}$, which depends on the sampling properties of the unbiased effects. For example, if $[\alpha_i]_n$ and $[\gamma_i]_i$ are independent sequences, and $\alpha_i$ and $\gamma_i$ are independent for all $i, t$, then $r_{NT} = \sqrt{NT/(N + T - 1)}$, and

$$\mathbb{V}_{NT}^2 = \lim_{N,T \to \infty} \left\{ \frac{1}{N^2 T^2} \sum_{i=1}^N \sum_{t=1}^T \mathbb{E}[\partial_u \Delta_u \partial_u \Delta_u^T] + \sum_{j \neq i} \sum_{t=1}^T \mathbb{E}[\partial_u \Delta_u \partial_u \Delta_u^T] \right\}.$$

In the expression of $\mathbb{V}_{NT}^2$, the first two terms come from $\delta - \delta_{NT}$, whereas the last term comes from $\delta - \delta$. The first two terms of $\mathbb{V}_{NT}^2$ are of order $NT(N + T - 1)^2$, whereas the last term of $\mathbb{V}_{NT}^2$ is of order $NT/(NT)^2 = O(1)$ by construction.

**Remark 5 (Average Effects from Bias Corrected Estimators).** The first term in the expressions of the biases $\mathbb{E}_{NT}^* \Delta_u$ and $\mathbb{E}_{NT}^* \Delta_u$ comes from the bias of the estimator of $\beta$. It drops out when the APEs are constructed from asymptotically unbiased or bias corrected estimators of the parameter $\beta$, i.e.

$$\delta = \Delta(\beta, \hat{\phi}(\beta)),$$

where $\hat{\phi}$ is such that $\sqrt{NT}(\beta - \beta^0) \to_d N(0, \mathbb{V}_{NT}^2)$. The asymptotic variance of $\delta$ is the same as in Theorem 4.2.

In the following examples we assume that the APEs are constructed from asymptotically unbiased estimators of the model parameters.

**Example 1 (Binary Response Model).** Consider the partial effects defined in (2.3) and (2.4) with

$$\Delta_u(\beta, \pi) = F(\beta_u + X_{u-k} + \pi) - F(X_{u-k} + \pi)$$

and

$$\Delta_u(\beta, \pi) = \beta_u F(X_{u} + \beta_u + \pi).$$
Using the notation previously introduced for this example, the components of the asymptotic bias of $\delta$ are

$$
P_{\infty}^{\delta} = \lim_{N,T \to \infty} \left[ \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{t' = t+1}^{T} \mathbb{E}_\phi \left( H_\alpha (Y_{it} - F_{it}) \omega_t \psi_{it} \right) + \frac{1}{2N} \sum_{i=1}^{N} \sum_{t=1}^{T} H_\alpha \frac{\partial^2 F_{it}}{\omega_t} \mathbb{E}_\phi (\alpha_t \Delta_{it}) \right],$$

where $\psi_{it}$ is the residual of the population regression of $-\partial_t \Delta_{it} / \mathbb{E}_\phi [\omega_{it}]$ on the space spanned by the incidental parameters under the metric given by $\mathbb{E}_\phi [\omega_{it}]$. If all the components of $X_{it}$ are strictly exogenous, the first term of $P_{\infty}^{\delta}$ is zero.

**Example 2 (Poisson Model).** Consider the partial effect $\Delta_\beta (\beta, \pi) = \gamma_0 (\beta) \exp(X_{it}^\prime \beta + \pi),$

where $\gamma_0$ does not depend on $\pi$. For example, $\gamma_0 (\beta) = \beta_0 + \beta_1 h(Z_t)$ in (2.5). Using the notation previously introduced for this example, the components of the asymptotic bias are

$$
P_{\infty}^{\delta} = \lim_{N,T \to \infty} \left[ \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{t' = t+1}^{T} \mathbb{E}_\phi \left( Y_{it} - \omega_t \psi_{it} \right) \right],$$

and $D_{\infty}^{\delta} = 0$, where $\bar{g}_0$ is the residual of the population projection of $g_0$ on the space spanned by the incidental parameters under a metric weighted by $\mathbb{E}_\phi [\omega_{it}]$. The asymptotic bias is zero if all the components of $X_{it}$ are strictly exogenous or $g_0 (\beta)$ is constant. The latter arises in the leading case of the partial effect of the kth component of $X_{it}$ since $g_0 (\beta) = \beta_k$. This asymptotic bias result applies to any type of regressor, strictly exogenous or predetermined.

### 4.3. Bias corrected estimators

The results of the previous sections show that the asymptotic distributions of the fixed effects estimators of the model parameters and APEs can have biases of the same order as the variances under sequences where $T$ grows at the same rate as $N$. This is the large-$T$ version of the incidental parameter problem that invalidates any inference based on the fixed effect estimators even in large samples. In this section we describe how to construct analytical and jackknife bias corrections for the fixed effect estimators and give conditions for the asymptotic validity of these corrections.

The jackknife correction for the model parameter $\beta$ in Eq. (3.4) is generic and applies to the panel model. For the APEs, the jackknife correction is formed similarly as

$$
\hat{\delta}_{NT} = \hat{\delta}_{NT} - \hat{\delta}_{N,T/2} - \hat{\delta}_{N/2,T},
$$

where $\hat{\delta}_{N,T/2}$ is the average of the 2 split jackknife estimators of the APE that use all the individuals and leave out the first and second halves of the time periods, and $\hat{\delta}_{N/2,T}$ is the average of the 2 split jackknife estimators of the APE that use all the time periods and leave out half of the individuals.

The analytical corrections are constructed using sample analogs of the expressions in Theorems 4.1 and 4.2, replacing the true values of $\beta$ and $\phi$ by the fixed effects estimators. To describe these corrections, we introduce some additional notation. For any function of the data, unobserved effects and parameters $g_0 (\beta, \alpha_i + \gamma_i, \alpha_i + \gamma_{i-1})$ with $0 \leq j < t$, let $\tilde{g}_0 = g_0 (\tilde{\beta}, \tilde{\alpha}_i + \tilde{\gamma}_i, \tilde{\alpha}_i + \tilde{\gamma}_{i-1})$ denote the fixed effects estimator, e.g., $\mathbb{E}_\phi [\bar{g}_0 (\omega_{it})] = \mathbb{E}_\phi [\bar{g}_0 (\omega_{it})]$ denotes the fixed effects estimator of $\mathbb{E}_\phi [\bar{g}_0 (\omega_{it})]$. Let $\tilde{\mathcal{H}}^{\delta^{-1}}, \tilde{\mathcal{H}}_i^{(o),\delta^{-1}}, \tilde{\mathcal{H}}^{(y),\delta^{-1}},$ and $\tilde{\mathcal{H}}^{(y),\delta^{-1}}$ denote the blocks of the matrix $\tilde{\mathcal{H}}^{\delta^{-1}}$, where

$$
\tilde{\mathcal{H}}^{\delta^{-1}} = \left( \begin{array}{c} \tilde{\mathcal{H}}_i^{(o),\delta^{-1}} \\ \tilde{\mathcal{H}}_i^{(y),\delta^{-1}} \end{array} \right) + \frac{b}{\sqrt{NT}} u^\prime v,
$$

$$
\tilde{\mathcal{H}}_i^{(o),\delta^{-1}} = \mathbb{E}_\phi \left[ \left( \begin{array}{c} \tilde{\mathcal{H}}_i^{(o),\delta^{-1}} \\ \tilde{\mathcal{H}}_i^{(y),\delta^{-1}} \end{array} \right) \right] = \mathbb{E}_\phi \left[ \left( \begin{array}{c} \tilde{\mathcal{H}}_i^{(o),\delta^{-1}} \\ \tilde{\mathcal{H}}_i^{(y),\delta^{-1}} \end{array} \right) \right].
$$

The kth component of $\tilde{\mathcal{H}}^{\delta^{-1}}$ corresponds to a least squares regression of $\mathbb{E}_\phi [\tilde{\partial}_{it} (\omega_{it})] / \mathbb{E}_\phi [\tilde{g}_0 (\omega_{it})]$ on the space spanned by the incidental parameters weighted by $\mathbb{E}_\phi [\bar{g}_0 (\omega_{it})]$. The analytical bias corrected estimator of $\beta^0$ is

$$
\tilde{\beta}^0 = \beta - \tilde{\beta}_{NT} / T - \tilde{\beta}_{NT} / N,
$$

where

$$
\tilde{\beta}_{NT} = \mathbb{E}_\phi \left[ \tfrac{\partial^2 F_{it}}{\omega_t} \right] - \mathbb{E}_\phi \left[ \tfrac{\partial^2 F_{it}}{\omega_t} \right] = \mathbb{E}_\phi \left[ \tfrac{\partial^2 F_{it}}{\omega_t} \right] - \mathbb{E}_\phi \left[ \tfrac{\partial^2 F_{it}}{\omega_t} \right].
$$

(4.6)

and $L$ is a trimming parameter for estimation of spectral expectations such that $L \to \infty$ and $L/T \to 0$ (Hahn and Kuersteiner, 2011). Here we use truncation instead of kernel smoothing in the estimation of spectral expectations following Hahn and Kuersteiner (2007). Note that, unlike for variance estimation, a kernel is not needed to ensure that the bias estimator be positive. Instead of choosing a value of $L$, our recommendation for practice is to conduct a sensitivity analysis by reporting estimates for multiple values of $L$ starting from $L = 0$. From our experience based on extensive Monte Carlo simulations, we do not recommend values of $L$ greater than 4, because the finite-sample dispersion of the estimator quickly increases with $L$. We refer to Section 5 for an example of sensitivity analysis with respect to $L$. The factor $T/(T - 1)$ is a degrees of freedom adjustment that rescales the time series averages $T^{-1} \sum_{t'=t+1}^{T} \mathbb{E}_\phi [\bar{g}_0 (\omega_{it})]$, by the number of observations instead of by $T$. Similar corrections for conditional mean models can be formed using the sample analogs of the
expressions of \( \widetilde{B}_\infty \) and \( \overline{B}_\infty \) in Remark 3. We do not spell out these estimators for the sake of brevity. Asymptotic \((1-p)\)-confidence intervals for the components of \( \beta \) can be formed as
\[
\hat{\beta}^k_{\kappa} = z_{1-p}\sqrt{\overline{W}_{kk}/(NT)}, \quad k = \{1, \ldots, \dim \beta\},
\]
where \( z_{1-p} \) is the \((1-p)\)-quantile of the standard normal distribution, and \( \overline{W}_{kk} \) is the \((k,k)\)-element of the matrix \( \overline{W}^{-1} \). In conditional moment models we replace \( \overline{W}_{kk} \) by the \((k,k)\)-element of the matrix \( W^{-1}\overline{2}W^{-1} \), where
\[
\hat{\Omega} = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^T \sum_{k=1}^T \mathbb{E}_\phi \left[ D_{\beta}(\hat{\theta}_i \ell_j)(D_\phi \ell_{it})' \right].
\]

We have implemented the analytical correction at the level of the estimator. Alternatively, we can implement the correction at the level of the score or first order conditions by solving
\[
(NT)^{-1/2} \partial_\beta \mathcal{L}(\beta, \hat{\phi}(\beta)) = \hat{B}/T + \hat{D}/N,
\]
for \( \beta \). Global concavity of the objective function guarantees that the solution to (4.8) is unique. Other possible extensions such as continuously updated score corrections where \( \overline{B}_\infty \) and \( \overline{B}_\infty \) are estimated together with \( \beta \), corrections at the level of the objective function, or iterative corrections are left to future research.

The analytical bias corrected estimator of \( \delta_{NT}^i \) is
\[
\hat{\delta}_i = \hat{\delta} - \hat{B}^i/T - \hat{D}^i/N,
\]
where \( \hat{\delta} \) is the APE constructed from a bias corrected estimator of \( \delta_{NT}^i \). Let
\[
\hat{\Psi}_{it} = \frac{1}{\sqrt{NT}} \sum_{j=1}^N \sum_{t=1}^T \left( \hat{H}_{(\alpha,\gamma)}^{-1} + \hat{H}_{(y,\gamma)}^{-1} \right) \partial_\beta \Delta_{it}.
\]
The fixed effects estimator of the components of the asymptotic biases are
\[
\hat{\beta} = \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \left[ \frac{1}{(T-(j-1))} \sum_{j=1}^T \mathbb{E}_\phi \left( \partial_\beta \delta_{i-1,j} \ell_{j,t} \ell_{j,t} \psi_{it} \right) \right. \\
\left. - \frac{1}{2N} \sum_{t=1}^T \mathbb{E}_\phi \left( \partial_\beta \delta_{i,j} \ell_{j,t} \right) \right] \\
\hat{\delta} = \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \mathbb{E}_\phi \left( \partial_\beta \delta_{i,j} \ell_{j,t} \right).
\]
The estimator of the asymptotic variance depends on the sampling properties of the unobserved effects. Under the independence assumption of Remark 4,
\[
\hat{\nu}_i = \frac{r_{NT}}{NT^2} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^T \sum_{t=1}^T \hat{\Delta}_{it} \hat{\Delta}_{jt} + \sum_{j=1}^T \sum_{t=1}^T \hat{\Delta}_{it} \hat{\Delta}_{jt} \\
+ \sum_{t=1}^T \sum_{j=1}^T \mathbb{E}_\phi \left( \partial_\beta \Delta_t \partial_\beta \Delta_{jt} \right),
\]
where \( \hat{\Delta}_{it} = \hat{\Delta}_{it} - N^{-1} \sum_{t=1}^T \hat{\Delta}_{it} \) under identical distribution over \( i \), \( \hat{\Delta}_{it} = \hat{\Delta}_{it} - T^{-1} \sum_{t=1}^T \Delta_{it} \) under stationarity over \( t \), and \( \hat{\Delta}_{it} = \hat{\Delta}_{it} - \hat{\delta} \) under both. Note that we do not need to specify the convergence rate \( r_{NT} \) to make inference because the standard errors \( \sqrt{\hat{\nu}_i} \) do not depend on \( r_{NT} \). Bias corrected estimators and confidence intervals can be constructed in the same fashion as for the model parameter.

We use the following homogeneity assumption to show the validity of the jackknife corrections for the model parameters and APEs. It implies that \( \hat{\beta}_{NT,1/2} - \hat{\beta}_{NT} = \overline{B}_\beta/T + o_p(T^{-1} \vee N^{-1}) \) and \( \hat{\beta}_{NT,2/2} - \hat{\beta}_{NT} = \overline{B}_\beta/N + o_p(T^{-1} \vee N^{-1}) \), which are weaker but higher level sufficient conditions for the validity of the jackknife for the model parameter. For APEs, Assumption 4.3 also ensures that these effects do not change with \( T \) and \( N \), i.e. \( \delta_{NT}^0 = \delta \). The analytical corrections do not require this assumption.

**Assumption 4.3 (Unconditional Homogeneity).** The sequence \( \{(Y_{it}, X_{it}, \alpha_t, \gamma_t) : 1 \leq i \leq N, 1 \leq t \leq T\} \) is identically distributed across \( i \) and strictly stationary across \( t \) for each \( N, T \).

This assumption might seem restrictive for dynamic models where \( X_{it} \) includes lags of the dependent variable because in this case it restricts the unconditional distribution of the initial conditions of \( Y_{it} \). Note, however, that Assumption 4.3 allows the initial conditions to depend on the unobserved effects. In other words, it does not impose that the initial conditions are generated from the stationary distribution of \( Y_{it} \) conditional on \( X_{it} \) and \( \phi \). Assumption 4.3 rules out time trends and structural breaks in the processes for the unobserved effects and observed variables.

**Remark 6 (Test of Homogeneity).** Assumption 4.3 is a sufficient condition for the validity of the jackknife corrections. It has the testable implications that the probability limits of the fixed effects estimator are the same in all the partitions of the panel. For example, it implies that \( \hat{\beta}_{NT,1/2} = \hat{\beta}_{NT,2/2} \), where \( \hat{\beta}_{NT,1/2} \) and \( \hat{\beta}_{NT,2/2} \) are the probability limits of the fixed effects estimators of \( \beta \) in the subpanels that include all the individuals and the first and second halves of the time periods, respectively. These implications can be tested using variations of the Chow-type test proposed in Dhaene and Jochmans (2015). We provide an example of the application of these tests to our setting in Section S.5.1 of the supplemental material.

The following theorems are the main results of this section. They show that the analytical and jackknife bias corrections eliminate the bias from the asymptotic distribution of the fixed effects estimators of the model parameters and APEs without increasing variance, and that the estimators of the asymptotic variances are consistent.

**Theorem 4.3 (Bias Corrections for \( \hat{\beta} \)).** Under the conditions of Theorem 4.1,
\[
\hat{W} \rightarrow_p \overline{W}_\infty,
\]
and, if \( L \rightarrow \infty \) and \( L/T \rightarrow 0 \),
\[
\sqrt{NT} (\hat{\beta} - \beta^0) \rightarrow_d \mathcal{N}(0, \overline{W}_\infty).
\]
Under the conditions of Theorems 4.1 and 4.3,
\[
\sqrt{NT} (\hat{\beta} - \beta^0) \rightarrow_d \mathcal{N}(0, \overline{W}_\infty).
\]

**Theorem 4.4 (Bias Corrections for \( \hat{\delta} \)).** Under the conditions of Theorems 4.1 and 4.2,
\[
\hat{\nu}_i \rightarrow_p \overline{\nu}_\infty,
\]
and, if \( L \rightarrow \infty \) and \( L/T \rightarrow 0 \),
\[
r_{NT} (\hat{\delta}_i - \delta_{NT}^0) \rightarrow_d \mathcal{N}(0, \overline{\nu}_\infty).
\]
Under the conditions of Theorems 4.1 and 4.2, and Assumption 4.3, 
$r_{NT}(\hat{\theta} - \theta_0) \to_d N(0, \tilde{V}_i).$

Remark 7 (Rate of Convergence). The rate of convergence $r_{NT}$ depends on the properties of the sampling process for the explanatory variables and unobserved effects (see Remark 4).

5. Monte Carlo experiments

This section reports evidence on the finite sample behavior of fixed effects estimators of model parameters and APEs in static models with strictly exogenous regressors and dynamic models with predetermined regressors such as lags of the dependent variable. We analyze the performance of uncorrected and bias-corrected fixed effects estimators in terms of bias and inference accuracy of their asymptotic distribution. In particular we compute the biases, standard deviations, and root mean squared errors of the estimators, the ratio of average standard errors to the simulation standard deviations (SE/SD); and the empirical coverages of confidence intervals with 95% nominal value (p . 95). Overall, we find that the analytically corrected estimators dominate the uncorrected and jackknife corrected estimators. A possible explanation for the better finite-sample performance of the analytical over the jackknife corrections is that the jackknife increases dispersion because the components of the bias are estimated from subsamples that include half of the observations of the panel. We observe this variance increase in all our numerical examples, specially in short panels. The jackknife corrections are also more sensitive than the analytical corrections to Assumption 4.3. All the results are based on 500 replications. The designs correspond to static and dynamic probit models. As in the analytical example of Section 3.3, we find that our large T asymptotic approximations capture well the behavior of the fixed effects estimator and the bias corrections in moderately long panels with $N = 56$ and $T = 14$.

5.1. Static probit model

The data generating process is

$$Y_{it} = \left\{ \begin{array}{ll} X_{it} \beta_0 + \alpha_i + \gamma_t > \varepsilon_{it}, & i = 1, \ldots, N; \ t = 1, \ldots, T, \\ \end{array} \right.$$ 

where $\alpha_i \sim \mathcal{N}(0, 1/16)$, $\gamma_t \sim \mathcal{N}(0, 1/16)$, $\varepsilon_{it} \sim \mathcal{N}(0, 1)$, and $\beta = 1$. We consider two alternative designs for $X_{it}$: autoregressive process and linear trend process both with individual and time effects.

In the first design, $X_{it} = X_{i,t-1}/2 + \alpha_i + \gamma_t + \upsilon_{it}, \upsilon_t \sim \mathcal{N}(0, 1/2),$ and $\beta = 1$. In the second design, $X_{it} = 2t/T + \alpha_i + \gamma_t + \upsilon_{it}, \upsilon_t \sim \mathcal{N}(0, 3/4),$ which violates Assumption 4.3. In both designs $X_0$ is strictly exogenous with respect to $e_{it}$ conditional on the individual and time effects. The variables $\alpha_i, \gamma_t, \xi_t, \upsilon_{it},$ and $\upsilon_t$ are independent and i.i.d. across individuals and time periods. We generate panel data sets with $N =$ 56 individuals and three different numbers of time periods $T$: 14, 28 and 56.

Table 3 reports the results for the probit coefficient $\beta_0$ and the APE of $X_{it}$. We compute the APE using (2.4). Throughout the table, MLE-FEFE corresponds to the probit maximum likelihood estimator with individual and time fixed effects. Analytical is the bias corrected estimator that uses the analytical correction, and Jackknife is the bias corrected estimator that uses SJ in both the individual and time dimensions. The cross-sectional division in the jackknife follows the order of the observations. All the results are reported in percentage of the true parameter value.

We find that the bias is of the same order of magnitude as the standard deviation for the uncorrected estimator of the probit coefficient causing severe undercoverage of the confidence intervals. This result holds for both designs and all the sample sizes considered. The bias corrections, specially Analytical, remove the bias without increasing dispersion, and produce substantial improvements in rmse and coverage probabilities. For example, Analytical reduces rmse by 50% and increases coverage by 26% in the first design with $T = 14$. As in Hahn and Newey (2004) and Fernández-Val (2009), we find very little bias in the uncorrected estimates of the APE, despite the large bias in the probit coefficients. Jackknife performs relatively worse in the second design that does not satisfy Assumption 4.3.

5.2. Dynamic probit model

The data generating process is

$$Y_{it} = \left\{ \begin{array}{ll} Y_{i,t-1} \beta_0 + Z_{it} \beta_2 + \alpha_i + \gamma_t > \varepsilon_{it}, & i = 1, \ldots, N; \ t = 1, \ldots, T, \\ \end{array} \right.$$ 

where $\alpha_i \sim \mathcal{N}(0, 1/16)$, $\gamma_t \sim \mathcal{N}(0, 1/16)$, $\varepsilon_{it} \sim \mathcal{N}(0, 1)$, $\beta_0 = 0.5,$ and $\beta_2 = 1$. We consider two alternative designs for $Z_{it}$: autoregressive process and linear trend process both with individual and time effects. In the first design, $Z_{it} = Z_{i,t-1}/2 + \alpha_i + \gamma_t + \upsilon_{it}, \upsilon_t \sim \mathcal{N}(0, 1/2),$ and $\beta = 1$. In the second design, $Z_{it} = 1.5t/T + \alpha_i + \gamma_t + \upsilon_{it}, \upsilon_t \sim \mathcal{N}(0, 3/4),$ which violates Assumption 4.3. The variables $\alpha_i, \gamma_t, \xi_t, \upsilon_t,$ and $Z_{it}$ are independent and i.i.d. across individuals and time periods. We generate panel data sets with $N =$ 56 individuals and three different numbers of time periods $T$: 14, 28 and 56.

Table 4 reports the simulation results for the probit coefficient $\beta_0$ and the APE of $Y_{i,t-1}$. We compute the partial effect of $Y_{i,t-1}$ using the expression in Eq. (2.3) with $X_{i,t} = Y_{i,t-1}$. This effect is commonly reported as a measure of state dependence for dynamic binary processes. Table 5 reports the simulation results for the estimators of the probit coefficient $\beta_2$ and the APE of $Z_{it}$. We compute the partial effect using (2.4) with $X_{i,t} = Z_{i,t}$. Throughout the tables, we compare the same estimators as for the static model. For the analytical correction we consider two versions, Analytical ($L = 1$) sets the trimming parameter to estimate spectral expectations $L$ to one, whereas Analytical ($L = 2$) sets $L$ to two. Again, all the results in the tables are reported in percentage of the true parameter value.

The results in Table 4 show important biases toward zero for both the probit coefficient and the APE of $Y_{i,t-1}$ in the two designs. This bias can indeed be substantially larger than the corresponding standard deviation for short panels yielding coverage probabilities below 70% for $T = 14$. The analytical corrections significantly reduce biases and rmse, bring coverage probabilities close to their nominal level, and have little sensitivity to the trimming parameter $L$. The jackknife corrections reduce bias but increase dispersion, producing less drastic improvements in rmse and coverage than the analytical corrections. The results for the APE of $Z_{it}$ in Table 5

14 The standard errors are computed using the expressions (4.7) and (4.9) with $\Delta \hat{\beta}_0 = \hat{\beta}_0 - \hat{\beta}_0$ evaluated at uncorrected estimates of the parameters. We find little difference in performance of constructing standard errors based on corrected estimates.

15 Kristensen and Salanié (2013) also found that analytical corrections dominate jackknife corrections to reduce the bias of approximate estimators.

16 Following a suggestion from an anonymous referee, we obtained results for panel data sets with $T = 56$ and $N$ in {14, 28, 56}. These results are similar to the results reported and are available from the authors upon request.

17 In results not reported for brevity, we find little difference in performance of increasing the trimming parameters to $L = 3$ and $L = 4$. These results are available from the authors upon request.
are similar to the static probit model. There are significant bias and undercoverage of confidence intervals for the coefficient $\beta_2$, which are removed by the corrections, whereas there are little bias and undercoverage in the APE. As in the static model, Jackknife performs relatively worse in the second design.

### 6. Concluding remarks

In this paper we develop analytical and jackknife corrections for fixed effects estimators of model parameters and APEs in semiparametric nonlinear panel models with additive individual and time effects. Our analysis applies to conditional maximum likelihood estimators with concave log-likelihood functions, and therefore covers logit, probit, ordered probit, ordered logit, Poisson, negative binomial, and Tobit estimators, which are the most popular nonlinear estimators in empirical economics.

We are currently developing similar corrections for nonlinear models with interactive individual and time effects (Chen et al., 2014). Another interesting avenue of future research is to derive higher-order expansions for fixed effects estimators with individual and time effects. These expansions are needed to justify theoretically the validity of alternative corrections based on the leave-one-observation-out panel jackknife method of Hahn and Newey (2004).

### Appendix A. Notation and choice of norms

We write $\text{vec}$ for the transpose of a matrix or vector $A$. We use $1_n$ for the $n \times n$ identity matrix, and $1_T$ for the column vector of length $n$ whose entries are all unity. For square $n \times n$ matrices $B$, $C$, we use $B > C$ (or $B \geq C$) to indicate that $B - C$ is positive (semi) definite.

We write wpa I for “with probability approaching one” and wtr for “with respect to”. All the limits are taken as $N, T \to \infty$ jointly.

As in the main text, we usually suppress the dependence on $NT$ of all the sequences of functions and parameters to lighten the notation, e.g. we write $\mathcal{L}$ for $\mathcal{L}_{NT}$ and $\phi$ for $\phi_{NT}$. Let

$$
\delta(\beta, \phi) = \partial_\phi \mathcal{L}(\beta, \phi), \quad \mathcal{H}(\beta, \phi) = -\partial_{\beta\beta}^\phi \mathcal{L}(\beta, \phi),
$$

where $\partial_\phi$ denotes the partial derivative of $\phi$ with respect to $\phi$, and additional subscripts denote higher-order partial derivatives. We refer to the dim $\phi$-vector $\delta(\beta, \phi)$ as the incidental parameter score, and to the dim $\phi \times \dim \phi$ matrix $\mathcal{H}(\beta, \phi)$ as the incidental parameter Hessian. We omit the arguments of the functions when they are evaluated at the true parameter values $(\beta^0, \phi^0)$, e.g. $\mathcal{H} = \mathcal{H}(\beta^0, \phi^0)$. We use a bar to indicate expectations conditional on $\phi$, e.g. $\bar{\beta}_\phi \mathcal{L} = \mathbb{E}_{\phi} [\partial_\phi \mathcal{L}]$, and a tilde to denote variables in deviations with respect to expectations, e.g. $\tilde{\beta}_\phi \mathcal{L} = \beta_\phi \mathcal{L} - \bar{\beta}_\phi \mathcal{L}$.

We use the Euclidean norm $\| \cdot \|$ for vectors of dimension $\dim \beta$ and we use the norm induced by the Euclidean norm for the corresponding matrices and tensors, which we also denote by $\| \cdot \|$. For matrices of dimension $\dim \beta \times \dim \beta$ this induced norm is the spectral norm. The generalization of the spectral norm to higher order tensors is straightforward, e.g. the induced norm of the dim $\beta \times \dim \beta \times \dim \beta$ tensor of third partial derivatives of $\mathcal{L}(\beta, \phi)$ wtr $\beta$ is given by

$$
\| \bar{\partial}_{\beta\beta\beta} \mathcal{L}(\beta, \phi) \| = \max_{u \in \mathbb{R}^{\dim \beta}, \| u \|_1 = 1} \left\{ \sum_{k=1}^{\dim \beta} u_k \bar{\partial}_{\beta u_k \beta} \mathcal{L}(\beta, \phi) \right\}.
$$

This choice of norm is immaterial for the asymptotic analysis because dim $\beta$ is fixed with the sample size.
In contrast, it is important what norms we choose for vectors of dimension $\dim \phi$, and their corresponding matrices and tensors, because $\dim \phi$ is increasing with the sample size. For vectors of dimension $\dim \phi$, we use the $\ell_q$-norm

$$
\|\phi\|_q = \left( \sum_{g=1}^{\dim \phi} |\phi_g|^q \right)^{1/q},
$$

where $2 \leq q \leq \infty$.\(^{18}\) The particular value $q = 8$ will be chosen later.\(^{19}\) We use the norms induced by the $\ell_q$-norm for the corresponding matrices and tensors, e.g. the induced $q$-norm of the $d \phi \times \dim \phi \times \dim \phi$ tensor of third partial derivatives of $\mathcal{L}(\beta, \phi)$ wrt $\mathbf{t}$ is

$$
\|\partial_{\phi\phi\phi} \mathcal{L}(\beta, \phi)\|_q = \max_{u, v \in \mathbb{R}^{\dim \phi}, \|u\|_q = 1, \|v\|_q = 1} \| \sum_{k,l=1}^{\dim \phi} u_k v_l \partial_{\beta_1 \beta_2 \beta_3} \mathcal{L}(\beta, \phi) \|_q.
$$

Note that in general the ordering of the indices of the tensor would matter in the definition of this norm, with the first index having a special role. However, since partial derivatives like $\partial_{\phi_1 \phi_2 \phi_3} \mathcal{L}(\beta, \phi)$ are fully symmetric in the indices $g, h, l$, the ordering is not important in their case.

For mixed partial derivatives of $\mathcal{L}(\beta, \phi)$ wrt $\beta$ and $\phi$, we use the norm that is induced by the Euclidean norm on $\dim \beta$-vectors and the $q$-norm on $\dim \phi$-indices, e.g.

$$
\|\partial_{\beta \phi \phi\phi} \mathcal{L}(\beta, \phi)\|_q = \max_{u, v \in \mathbb{R}^{\dim \beta}, \|u\|_q = 1, \|v\|_q = 1} \max_{u, x \in \mathbb{R}^{\dim \phi}, \|u\|_q = 1, \|x\|_q = 1} \| \sum_{k,l=1}^{\dim \phi} u_k v_l x_h \partial_{\beta_1 \beta_2 \beta_3} \mathcal{L}(\beta, \phi) \|_q.
$$

\(^{18}\) We use the letter $q$ instead of $p$ to avoid confusion with the use of $p$ for probability.

\(^{19}\) The main reason not to choose $q = \infty$ is the assumption $\|R^{-1}\|_q = o_q(1)$ below, which is used to guarantee that $\|R^{-1}\|_q$ is of the same order as $\|R^{-1}\|_q$. If we assume $\|R^{-1}\|_q = o_q(1)$ directly instead of $\|R^{-1}\|_q = O_q(1)$, then we can set $q = \infty$.\(^{18}\)
Thus, whenever we bound a scalar product of vectors, matrices and tensors in terms of the above norms we have to account for this additional factor $(\dim \phi)^{q-2}/q$. For example,

$$
\sum_{k=1}^{\dim \beta} \sum_{\phi} \inner{u_k}{v_1} w_1 x_1 y_1 \partial_{\beta_k} \phi(\beta, \phi) \mathcal{L}(\beta, \phi)
\leq (\dim \phi)^{q-2}/q \norm{u} \norm{v_1} \norm{w_1} \norm{x_1} \norm{y_1} \norm{\partial_{\beta_k} \phi(\beta, \phi)}_{q}.
$$

For higher-order tensors, we use the notation $\partial_{\beta_k} \phi(\beta, \phi)$ inside the $q$-norm $\| \cdot \|_q$ defined above, while we rely on standard index and matrix notation for all other expressions involving those partial derivatives, e.g. $\partial_{\beta_k} \phi(\beta, \phi)$ is a $d_2 \times d_1$ matrix for every $g = 1, \ldots, d_2 \phi$. Occasionally, e.g. in Assumption B.1(vi), we use the Euclidean norm for $d_2 \phi$-vectors, and the spectral norm for $d_2 \phi \times d_1 \phi$-matrices, denoted by $\| \cdot \|$ and defined as $\| \cdot \|_q$ with $q = 2$. Moreover, we employ the matrix infinity norm $\| A \|_{\infty} = \max_i \sum_j |A_{ij}|$, and the matrix maximum norm $\| A \|_{\max} = \max_{ij} |A_{ij}|$ to characterize the properties of the inverse of the expected Hessian of the incidental parameters in Appendix D.

For $r \geq 0$, we define the sets $\mathcal{B}(r, \beta^0) = \{ \beta : \| \beta - \beta^0 \| \leq r \}$, and $\mathcal{B}_2(r, \beta^0) = \{ \phi : \| \phi - \phi^0 \| \leq r \}$, which are closed balls of radius $r$ around the true parameter values $\beta^0$ and $\phi^0$, respectively.

**Appendix B. Asymptotic expansions**

In this section, we derive asymptotic expansions for the score of the profile objective function, $\mathcal{L}(\beta, \phi(\beta))$, and for the fixed effects estimators of the parameters and APEs, $\hat{\beta}$ and $\hat{\phi}$. We do not employ the panel structure of the model, nor the particular form of the objective function given in Section 4. Instead, we consider the estimation of an unspecified model based on a sample of size $NT$. A generic objective function $\mathcal{L}(\beta, \phi)$, which depends on the parameter of interest $\beta$ and the incidental parameter $\phi$. The estimators $\hat{\phi}(\hat{\beta})$ and $\hat{\beta}$ are defined in (2.7) and (2.8). The proof of all the results in this Section are given in the supplementary material (see Appendix E).

We make the following high-level assumptions. These assumptions might appear somewhat abstract, but will be justified by more primitive conditions in the context of panel models.

**Assumption B.1 (Regularity Conditions for Asymptotic Expansion of $\hat{\beta}$).** Let $q > 4$ and $0 \leq u < 1/8 - 1/(2q)$. Let $r_\beta = r_{\beta,NT} > 0$, $r_\phi = r_{\phi,NT} > 0$, with $r_\beta = o \left( (NT)^{-1/2q} \right)$ and $r_\phi = o \left( (NT)^{-1/2q} \right)$. Assume that

(i) $\dim \phi = o(NT^{1/2q})$,

(ii) $\beta, \phi \mapsto \mathcal{L}(\beta, \phi)$ is $q$ times continuously differentiable in $\mathcal{B}(r_\beta, \beta^0) \times \mathcal{B}_2(r_\phi, \phi^0)$, with $p \in 1$.
(iii) \( \sup_{\beta \in \mathcal{B}(r, \beta_0)} \left\| \hat{\phi}(\beta) - \phi^0 \right\|_q = o_p(r_\phi). \)

(iv) \( \mathcal{H} > 0, \) and \( \left\| \mathcal{H}^{-1} \right\|_q = o_p(1). \)

(v) For the \( q \)-norm defined in Appendix A,
\[
\| \delta \|_q = o_p \left( (NT)^{-1/4+1/2(2q)} \right), \quad \| \partial \beta^\prime \mathcal{L} \| = o_p(1), \quad \| \hat{\mathcal{H}} \|_q = o_p(1), \quad \| \partial \beta\|_q = o_p(1), \quad \| \partial \phi \|_q = o_p(1), \quad \| \partial^2 \mathcal{L} \| = o_p(NT). \]
\[
\| \partial \phi \|_q = o_p(1), \quad \| \partial \phi \|_q = o_p((NT)^{1/4}), \quad \| \partial \phi \|_q = o_p((NT)^{1/4}). \]
\[
\| \partial \phi \|_q = o_p((NT)^{1/4}). \]

and
\[
\sup_{\beta \in \mathcal{B}(r, \beta_0)} \sup_{\phi \in \mathcal{B}(r, \phi_0)} \| \partial \phi \|_q = o_p \left( \sqrt{NT} \right), \quad \| \partial \phi \|_q = o_p \left( \sqrt{NT} \right), \quad \| \partial \phi \|_q = o_p \left( \sqrt{NT} \right), \quad \| \partial \phi \|_q = o_p \left( \sqrt{NT} \right), \quad \| \partial \phi \|_q = o_p \left( \sqrt{NT} \right). \]

(vi) For the spectral norm \( \| \cdot \| \) is \( \| \cdot \|_2, \)
\[
\| \hat{\mathcal{H}} \| = o_p \left( (NT)^{-1/8} \right), \quad \| \partial \sigma \|_2 = o_p \left( \sqrt{NT} \right), \quad \| \partial \phi \|_2 = o_p \left( (NT)^{-1/8} \right), \quad \| \partial \phi \|_2 = o_p \left( (NT)^{-1/8} \right), \quad \| \partial \phi \|_2 = o_p \left( (NT)^{-1/8} \right). \]
\[
\| \partial \phi \|_2 = o_p \left( (NT)^{-1/8} \right). \]

Let \( \partial \mathcal{L}(\beta, \hat{\phi}(\beta)) \) be the score of the profile objective function.\(^{20}\) The following theorem is the main result of this Appendix.

**Theorem B.1** (Asymptotic Expansions of \( \hat{\phi}(\beta) \) and \( \partial \beta \mathcal{L}(\beta, \hat{\phi}(\beta)). \)

Let Assumption B.1 hold. Then
\[
\hat{\phi}(\beta) - \phi^0 = \mathcal{H}^{-1} s + \mathcal{H}^{-1} \left[ \partial \phi \mathcal{L}(\beta - \beta^0) \right] + \frac{1}{2} \mathcal{H}^{-1} \sum_{\phi} \left[ \partial \phi \mathcal{L}(\beta - \beta^0) \right] \mathcal{H}^{-1} s, \]
and
\[
\partial \beta \mathcal{L}(\beta, \hat{\phi}(\beta)) = U - \overline{\mathcal{W}} \sqrt{NT} (\beta - \beta^0) + R(\beta), \]
where \( U = U^{(0)} + U^{(1)}, \) and
\[
\overline{\mathcal{W}} = - \frac{1}{\sqrt{NT}} \left[ \partial \beta \mathcal{L}(\beta - \beta^0) \right] \mathcal{H}^{-1} \left[ \partial \phi \mathcal{L}(\beta - \beta^0) \right], \]
\[
U^{(0)} = \partial \beta \mathcal{L}(\beta - \beta^0) \mathcal{H}^{-1} s \quad \text{and} \quad U^{(1)} = \partial \phi \mathcal{L}(\beta - \beta^0) \mathcal{H}^{-1} s. \]
\[
+ \frac{1}{2} \sum_{\phi} \left[ \partial \phi \mathcal{L}(\beta - \beta^0) \right] \mathcal{H}^{-1} \left[ \partial \phi \mathcal{L}(\beta - \beta^0) \right] \mathcal{H}^{-1} s. \]

The remainder terms of the expansions satisfy
\[
\sup_{\beta \in \mathcal{B}(r, \beta_0)} \left( NT \right)^{1/2-1/2q} \| R(\beta) \|_q = o_p(1), \quad \sup_{\beta \in \mathcal{B}(r, \beta_0)} \left( NT \right)^{1/2-1/2q} \| R(\beta) \|_q = o_p(1). \]

**Remark 8.** The result for \( \hat{\phi}(\beta) - \phi^0 \) does not rely on Assumption B.1.(vi). Without this assumption we can also show that
\[
\partial \beta \mathcal{L}(\beta, \hat{\phi}(\beta)) = \partial \beta \mathcal{L} + \left[ \partial \beta \mathcal{L} + \left( \partial \phi \mathcal{L} \right) \mathcal{H}^{-1} (\partial \phi \mathcal{L}) \right] (\beta - \beta^0), \]
\[
+ \left( \partial \phi \mathcal{L} \right) \mathcal{H}^{-1} s + \frac{1}{2} \sum_{\phi} \left( \partial \phi \mathcal{L} \right) \mathcal{H}^{-1} s, \]
with \( R_1(\beta) \) satisfying the same bound as \( R(\beta). \) Thus, the spectral norm bounds in Assumption B.1.(vi) for \( \dim \phi \)-vectors, matrices and tensors are only used after separating expectations from deviations of expectations for certain partial derivatives. Otherwise, the derivation of the bounds is purely based on the \( q \)-norm for \( \dim \phi \)-vectors, matrices and tensors.

The proofs are given in Section S.3 of the supplementary material. Theorem B.1 characterizes asymptotic expansions for the incidental parameter estimator and the score of the profile objective function in the incidental parameter score \( s \) up to quadratic order. The theorem provides bounds on the remainder terms \( R(\beta) \) and \( R(\beta), \) which make the expansions applicable to estimators of \( \beta \) that take values within a shrinking \( r_\beta \)-neighborhood of \( \beta^0 \) \( \text{wp}^1. \) Given such an \( r_\beta \)-consistent estimator \( \hat{\beta} \) that solves the first order condition \( \partial \beta \mathcal{L}(\beta, \hat{\phi}(\beta)) = 0, \) we can use the expansion of the profile objective score to obtain an asymptotic expansion for \( \hat{\beta}. \) This gives rise to the following corollary of Theorem B.1. Let \( \mathcal{W} = \lim_{N \to \infty} \mathcal{W}. \)

**Corollary B.2** (Asymptotic Expansion of \( \hat{\beta} \)). Let Assumption B.1 be satisfied. In addition, let \( U = o_p(1), \) and \( \mathcal{W} \) exist with \( \mathcal{W} > 0, \) and let \( \| \beta - \beta^0 \| = o_p(\mathcal{W}). \) Then,
\[
\sqrt{NT} (\hat{\beta} - \beta^0) = \mathcal{W}^{-1} U + o_p(1). \]

The following theorem states that for strictly concave objective functions no separate consistency proof is required for \( \hat{\phi}(\beta) \) and for \( \hat{\beta}. \)

**Theorem B.3** (Consistency under Concavity). Let Assumption B.1(i), (ii), (iv), (v) and (vi) hold, and let \( (\beta, \phi) \mapsto L(\beta, \phi) \) be strictly concave over \( (\beta, \phi) \in \mathbb{R}^{\dim \beta + \dim \phi}, \) \( \text{wp}^1. \) Assume furthermore that \( (NT)^{-1/4+1/2(2q)} = o_p(r_\phi) \) and \( NT^{1/2} r_\beta = o_p(r_\phi). \) Then,
\[
\sup_{\beta \in \mathcal{B}(r, \beta_0)} \left( NT \right)^{1/2-1/2q} \| R(\beta) \|_q = o_p(1), \quad \sup_{\beta \in \mathcal{B}(r, \beta_0)} \left( NT \right)^{1/2-1/2q} \| R(\beta) \|_q = o_p(1). \]

i.e. Assumption B.1(iii) is satisfied. If, in addition, \( \mathcal{W} \) exists with \( \mathcal{W} > 0, \) then \( \| \beta - \beta^0 \| = o_p \left( \left( NT \right)^{-1} \right). \)

In the application of Theorem B.1 to panel models, we focus on estimators with strictly concave objective functions. By Theorem B.3, we only need to check Assumption B.1(i), (ii), (iv), (v) and (vi), as well as \( U = o_p(1) \) and \( \mathcal{W} > 0, \) when we apply Corollary B.2 to derive the limiting distribution of \( \hat{\beta}. \) We give the proofs of Corollary B.2 and Theorem B.3 in Section S.3.

\(^{20}\) Note that \( \partial \mathcal{L}(\beta, \hat{\phi}(\beta)) = \partial \beta \mathcal{L}(\beta, \hat{\phi}(\beta)) \) by the envelope theorem.
Expansion for average effects

We invoke the following high-level assumption, which is verified under more primitive conditions for panel data models in the next section.

**Assumption B.2** (Regularity Conditions for Asymptotic Expansion of $\delta$). Let $q$, $\epsilon$, $r_\beta$ and $r_\phi$ be defined as in **Assumption B.1**. We assume that

(i) $(\beta, \phi) \mapsto \Delta(\beta, \phi)$ is three times continuously differentiable in $B(r_\beta, r_\beta^2, \epsilon, r_\phi, r_\phi^3)$, wp1.

(ii) $\left\| \partial_\beta \Delta \right\|_q = O_P(1)$, $\left\| \partial_\phi \Delta \right\|_q = O_P((NT)^{1/2q-1/2})$, $\left\| \partial_{\beta\phi} \Delta \right\|_q = O_P((NT)^{1/2} \epsilon^{-1/2})$, and

$$\sup_{\beta \in B(r_\beta, r_\beta^3, \phi \in B(r_\phi, r_\phi^3)} \left\{ \left\| \partial_\beta \Delta(\beta, \phi) \right\| = O_P(1), \right.$$ 

$$\sup_{\beta \in B(r_\beta, r_\beta^3, \phi \in B(r_\phi, r_\phi^3)} \left\{ \left\| \partial_{\beta\phi} \Delta(\beta, \phi) \right\| = O_P((NT)^{1/2} \epsilon^{-1/2}), \right.$$ 

$$\sup_{\beta \in B(r_\beta, r_\beta^3, \phi \in B(r_\phi, r_\phi^3)} \left\{ \left\| \partial_{\phi\phi} \Delta(\beta, \phi) \right\| = O_P((NT)^{1/2} \epsilon^{-1/2}) \right. \}

(iii) $\left\| \partial_\phi \Delta \right\|_q = O_P(1)$, $\left\| \partial_{\phi\phi} \Delta \right\|_q = O_P((NT)^{-1/2})$, and $\left\| \partial_{\phi\phi} \Delta \right\| = \left[ \frac{1}{16} \right] \left( \frac{1}{(NT)^{1/2}} \right)^2$.

The following result gives the asymptotic expansion for the estimator, $\hat{\delta} = \hat{\Delta}(\hat{\beta}, \hat{\phi}(\hat{\beta}))$, w.r.t. $\delta = \Delta(\beta^0, \phi^0)$.

**Theorem B.4** (Asymptotic Expansion of $\hat{\delta}$). Let **Assumptions B.1 and B.2** hold and let $\| \beta - \beta^0 \| = O_P((NT)^{-1/2}) = O_P(r_\beta)$. Then

$$\delta - \delta = \left[ \partial_\phi \Delta + (\partial_\phi \Delta) \overline{\mathcal{H}}^{-1} \partial_{\phi\phi} \overline{\mathcal{H}}^{-1} \right] (\beta - \beta^0) + U_{\Delta}^{(0)} + O_P\left( \frac{1}{\sqrt{NT}} \right),$$

where

$$U_{\Delta}^{(0)} = (\partial_\phi \Delta) \overline{\mathcal{H}}^{-1} \delta,$$

$$U_{\Delta}^{(1)} = (\partial_\phi \Delta) \overline{\mathcal{H}}^{-1} \delta - (\partial_\phi \Delta) \overline{\mathcal{H}}^{-1} \mathcal{H} \overline{\mathcal{H}} \mathcal{H}^{-1} \delta - \frac{1}{2} \delta \overline{\mathcal{H}}^{-1} \left[ \partial_{\phi\phi} \Delta + \sum_{g=1}^{\dim \phi} \left[ \partial_{\phi\phi} \phi^g \overline{\mathcal{X}} \right] \left[ \overline{\mathcal{H}}^{-1} \partial_{\phi} \Delta \right] \right] \overline{\mathcal{H}}^{-1} \delta.$$

**Remark 9.** The expansion of the profile score $\partial_\beta \xi(\beta, \hat{\phi}(\beta))$ in Theorem B.1 is a special case of the expansion in Theorem B.4, for $\Delta(\beta, \phi) = \frac{1}{\sqrt{NT}} \partial_\phi \xi(\beta, \phi)$. **Assumption B.2** also matches with the corresponding subset of **Assumption B.1**.

**Appendix C. Proofs of Section 4**

**C.1. Application of general expansion to panel estimators**

We now apply the general expansion of Appendix B to the panel fixed effects estimators considered in the main text. For the objective function specified in (2.1) and (4.1), the incidental parameter score evaluated at the true parameter value is

$$\delta = \left[ \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \partial_{\beta} \xi_{it} \right]_{i=1,...,N} + \left[ \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \partial_{\phi} \xi_{it} \right]_{i=1,...,T}.$$

The penalty term in the objective function does not contribute to $\delta$, because at the true parameter value $\nu \phi^0 = 0$. The corresponding expected incidental parameter Hessian $\overline{\mathcal{H}}$ is given in (4.2), Appendix D discusses the structure of $\overline{\mathcal{H}}$ and $\overline{\mathcal{H}}^{-1}$ in more detail. Define

$$A_{it} := -\frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \sum_{j=1}^{T} \left[ \overline{\mathcal{H}}^{-1}(\nu_{it} \nu_{ij}) + \overline{\mathcal{H}}^{-1}(\nu_{ij} \nu_{it}) \right] \partial_{\nu} \xi_{jt},$$

and the operator $D_{\beta} \Delta_{it} := \partial_{\beta} \Delta_{it} - \partial_{\beta} \Delta_{it} \xi_{it}$, which are similar to $\xi_{it}$ and $D_{\phi} \xi_{it}$ in Eq. (4.3).

The following theorem shows that **Assumptions 4.1 and 4.2** for the panel model are sufficient for **Assumptions B.1 and B.2** for the general expansion, and particularizes the terms of the expansion to the panel estimators. The proof is given in the supplementary material (see **Appendix E**).

**Theorem C.1.** Consider an estimator with objective function given by (2.1) and (4.1). Let **Assumption 4.1** be satisfied and suppose that the limit $\overline{\mathcal{W}}_{\infty}$ defined in **Theorem 4.1** exists and is positive definite. Let $q = 8, \epsilon = 1/(16 + 2\nu), r_{\beta,NT} = \log(NT)(NT)^{-1/4}$ and $r_{\phi,NT} = (NT)^{-1/16}$. Then

(i) **Assumption B.1** holds and $\| \hat{\beta} - \beta^0 \| = O_P((NT)^{-1/4})$.

(ii) The approximate Hessian and the terms of the score defined in **Theorem 1** can be written as

$$W = -\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}_\phi \left( \partial_{\beta \beta} \xi_{it} - \partial_{\beta \phi} \xi_{it} \mathbb{E}_\phi \right),$$

$$U_{\Delta}^{(0)} = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} D_{\beta} \xi_{it},$$

$$U_{\Delta}^{(1)} = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[ -A_{it} \left( D_{\phi \phi} \xi_{it} - \mathbb{E}_\phi (D_{\beta \phi} \xi_{it}) \right) \right] + \frac{1}{2} A_{it}^2 \mathbb{E}_\phi (D_{\phi \phi} \xi_{it}).$$

(iii) In addition, let **Assumption 4.2** hold. Then, **Assumption B.2** is satisfied for the partial effects defined in (2.2). By **Theorem B.4**, $\sqrt{NT} (\delta - \delta) = V_{\Delta}^{(0)} + V_{\Delta}^{(1)} + O_P(1),$ where

$$V_{\Delta}^{(0)} = \left[ \frac{1}{NT} \sum_{i,t} \left( \mathbb{E}_\phi (D_{\beta \beta} \Delta_{it}) \right) \right] \overline{\mathcal{W}}_{\infty} U_{\Delta}^{(0)} - \frac{1}{\sqrt{NT}} \sum_{i,t} \mathbb{E}_\phi (D_{\phi \beta} \xi_{it}),$$

$$V_{\Delta}^{(1)} = \left[ \frac{1}{NT} \sum_{i,t} \left( \mathbb{E}_\phi (D_{\beta \beta} \Delta_{it}) \right) \right] \overline{\mathcal{W}}_{\infty} U_{\Delta}^{(1)} + \frac{1}{\sqrt{NT}} \sum_{i,t} A_{it} \left( \mathbb{E}_\phi (D_{\phi \phi} \Delta_{it}) - \mathbb{E}_\phi (D_{\phi \phi} \xi_{it}) \right).$$

**C.2. Proofs of Theorems 4.1 and 4.2**

**Proof of Theorem 4.1.** # First, we want to show that $U_{\Delta}^{(0)} \rightarrow_d \mathcal{N}(0, W_{\infty})$. In our likelihood setting, $\mathbb{E}_\phi \partial_{\beta} \xi = 0$, $\mathbb{E}_\phi \delta = 0$, and, by the Bartlett identities, $\mathbb{E}_\phi (D_{\beta \beta} \partial_{\beta} \mathcal{X}) = -\frac{1}{\sqrt{NT}} \partial_{\beta \beta} \overline{\mathcal{X}},$

$$\mathbb{E}_\phi (\partial_{\beta} \mathcal{X} \delta') = -\frac{1}{\sqrt{NT}} \partial_{\beta \beta} \overline{\mathcal{X}} \mathbb{E}_\phi (\delta') = \frac{1}{\sqrt{NT}} \left( \overline{\mathcal{H}} - \frac{\mathbb{E}_\phi (\delta')}{\mathbb{E}_\phi (\delta')} \mathcal{X} \right).$$
Furthermore, $s'v = 0$ and $\partial_{\beta\phi} \tilde{L} = 0$. Then, by definition of $\tilde{W} = -\frac{1}{\sqrt{T}} \left( \partial_{\beta\phi} \tilde{L} + [\partial_{\beta\phi} \tilde{L}] \mathcal{F}^{-1} [\partial_{\beta\phi} \tilde{L}] \right)$ and $U^{(0)} = \partial_{\beta} L + [\partial_{\beta\phi} \tilde{L}] \mathcal{F}^{-1} \delta$, $\mathbb{E}_\phi \left( U^{(0)} \right) = 0$. Var $\left( U^{(0)} \right) = \tilde{W}$, which implies that $\lim_{T \to \infty} \text{Var} \left( U^{(0)} \right) = \lim_{T \to \infty} \tilde{W} = \tilde{W}_\infty$. Moreover, part (ii) of Theorem C.1 yields $U^{(0)} \to_d \mathcal{N} \left( 0, \lim_{T \to \infty} \text{Var} \left( U^{(0)} \right) \right) \sim \mathcal{N}(0, \tilde{W}_\infty)$.

Next, we show that $U^{(1)} = \mathcal{P} \phi \tilde{B}_\infty + \kappa^{-1} \mathcal{D}_\infty$. Part (ii) of Theorem C.1 gives $U^{(1)} = U^{(1a)} + U^{(1b)}$.

$U^{(1a)} = -\frac{1}{\sqrt{T}} \sum_{i=1}^N \sum_{t=1}^T A_{it} \left[ D_{\beta\phi} \ell_{it} - \mathbb{E}_\phi(D_{\beta\phi} \ell_{it}) \right]$, $U^{(1b)} = \frac{1}{2 \sqrt{T}} \sum_{i=1}^N \sum_{t=1}^T \mathbb{E}_\phi(D_{\beta\phi} \ell_{it})$.

Plugging-in the definition of $A_{it}$, we decompose $U^{(1a)} = U^{(1a,1)} + U^{(1a,2)} + U^{(1a,3)} + U^{(1a,4)}$, where

$U^{(1a,1)} = \frac{1}{T} \sum_{ij} \mathcal{H}_{(xy)tj} \left( \sum_j \partial_{\tau} \ell_{jT} \right) \sum_i \left[ D_{\beta\phi} \ell_{it} - \mathbb{E}_\phi(D_{\beta\phi} \ell_{it}) \right]$, $U^{(1a,2)} = \frac{1}{NT} \sum_{ij} \mathcal{H}_{(xy)tj} \left( \sum_j \partial_{\tau} \ell_{jT} \right) \sum_i \left[ D_{\beta\phi} \ell_{it} - \mathbb{E}_\phi(D_{\beta\phi} \ell_{it}) \right]$, $U^{(1a,3)} = \frac{1}{NT} \sum_{ij} \mathcal{H}_{(xy)tj} \left( \sum_j \partial_{\tau} \ell_{jT} \right) \sum_i \left[ D_{\beta\phi} \ell_{it} - \mathbb{E}_\phi(D_{\beta\phi} \ell_{it}) \right]$, $U^{(1a,4)} = \frac{1}{NT} \sum_{ij} \mathcal{H}_{(xy)tj} \left( \sum_j \partial_{\tau} \ell_{jT} \right) \sum_i \left[ D_{\beta\phi} \ell_{it} - \mathbb{E}_\phi(D_{\beta\phi} \ell_{it}) \right]$.

By the Cauchy-Schwarz inequality applied to the sum over $t$ in $U^{(1a,2)}$, $\left( U^{(1a,2)} \right)^2 \leq \frac{1}{(NT)^2} \left( \sum_j \left( \sum_i [\mathcal{H}_{(xy)tj} \partial_{\tau} \ell_{jT}]^2 \right) \right) \times \left( \sum_i \left[ D_{\beta\phi} \ell_{it} - \mathbb{E}_\phi(D_{\beta\phi} \ell_{it}) \right]^2 \right)$.

By Lemma D.1, $\mathcal{H}_{(xy)tj} = \mathcal{O}_p(1/\sqrt{NT})$, uniformly over $t,j$. Using that both $\sqrt{NT} \mathcal{H}_{(xy)tj} \partial_{\tau} \ell_{jT}$ and $D_{\beta\phi} \ell_{it} - \mathbb{E}_\phi(D_{\beta\phi} \ell_{it})$ are mean zero, independent across $i$ and $j$, and S.1 in the supplementary material across $t$, we obtain $\mathbb{E}_\phi \left( \frac{1}{\sqrt{N}} \left( \sum_{j=1}^N \mathcal{H}_{(xy)tj} \partial_{\tau} \ell_{jT} \right)^2 \right) = \mathcal{O}_p(1)$, $\mathbb{E}_\phi \left( \frac{1}{\sqrt{N}} \sum_i \left[ D_{\beta\phi} \ell_{it} - \mathbb{E}_\phi(D_{\beta\phi} \ell_{it}) \right]^2 \right) = \mathcal{O}_p(1)$, uniformly over $t$. Thus, $\sum_t \left( \sum_i [\mathcal{H}_{(xy)tj} \partial_{\tau} \ell_{jT}]^2 \right) = \mathcal{O}_p(T)$ and $\sum_t \left( \sum_i [D_{\beta\phi} \ell_{it} - \mathbb{E}_\phi(D_{\beta\phi} \ell_{it})] \right)^2 = \mathcal{O}_p(NT)$. We conclude that $\left( U^{(1a,2)} \right)^2 \leq \frac{1}{(NT)^2} \mathcal{O}_p(T) \mathcal{O}_p(NT) = \mathcal{O}_p(1/N) = \mathcal{O}_p(1)$, and therefore that $U^{(1a,2)} = \mathcal{O}_p(1)$. Analogously one can show that $U^{(1a,3)} = \mathcal{O}_p(1)$.

By Lemma D.1, $\mathcal{F}_{(xy)tj} = -\text{diag} \left( \frac{1}{\sqrt{N}} \sum_{t=1}^T \mathbb{E}_\phi(\partial_{\tau} \ell_{it}) \right)$, $\mathcal{O}_p(1/\sqrt{NT})$. Analogously to the proof of $U^{(1a,2)} = \mathcal{O}_p(1)$, one can show that the $\mathcal{O}_p(1/\sqrt{NT})$ part of $\mathcal{H}_{(xy)tj}$ has an asymptotically negligible contribution to $U^{(1a,1)}$. Thus, $U^{(1a,1)} = -\frac{1}{\sqrt{T}} \sum_i \left( \sum_j \partial_{\tau} \ell_{jT} \sum_i \left[ D_{\beta\phi} \ell_{it} - \mathbb{E}_\phi(D_{\beta\phi} \ell_{it}) \right] \right) + \mathcal{O}_p(1)$.

Our assumptions guarantee that $\mathbb{E}_\phi \left( \mathcal{H}_{(xy)tj} \right) = \mathcal{O}_p(1)$, uniformly over $t$. Note that both the denominator and the numerator of $U^{(1a,1)}$ are of order $T$. For the denominator this is obvious because of the sum over $T$. For the numerator there are two sums over $T$, but both $\partial_{\tau} \ell_{it}$ and $D_{\beta\phi} \ell_{it} - \mathbb{E}_\phi(D_{\beta\phi} \ell_{it})$ are mean zero weakly correlated processes, so that their sums are of order $\sqrt{T}$. By the WLLN over $i$ (remember that we have cross-sectional independence, conditional on $\phi$, and we assume finite moments), $\sqrt{N}^{-1} \sum_i U^{(1a,1)} = N^{-1} \sum_i \mathbb{E}_\phi(U^{(1a,1)}) + \mathcal{O}_p(1)$, and therefore $\left( U^{(1a,1)} \right)^2 = \mathcal{O}_p(1)$.
\[ A_{it}^{(4)} = -\frac{1}{\sqrt{NT}} \sum_{r=1}^{T} \sum_{\tau=1}^{T} \mathcal{F}(\gamma_{tr}) \sum_{t=1}^{T} \partial_\tau \ell_{rt}. \]

This decomposition of \( A_{it} \) induces the following decomposition of \( U^{(1b)} \):

\[ U^{(1b)} = \frac{4}{\sqrt{NT}} \sum_{\rho,q=1}^{4} U^{(1b,p,q)}, \]

\[ U^{(1b,p,q)} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{T} \mathcal{A}^{(p)}_{ij} \mathcal{A}_{ij}^{(q)} \mathcal{F}(\partial_{\rho} \ell_{it}). \]

Due to the symmetry \( U^{(1b,p,q)} = U^{(1b,q,p)} \), this decomposition has 10 distinct terms. Start with \( U^{(1b,1,2)} \) noting that

\[ U^{(1b,1,2)} = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} U^{(1b,1,2)}_{i}, \]

\[ U^{(1b,1,2)} = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{j=1}^{T} \mathcal{F}(\partial_{\rho} \ell_{it}) \mathcal{F}(\partial_{\tau} \ell_{jt}) \sum_{i=1}^{T} \partial_\tau \ell_{rt} \sum_{j=1}^{T} \partial_\tau \ell_{jt}. \]

By \( \mathcal{F}(\partial_\tau \ell_{it}) \) and \( \mathcal{F}(\partial_\tau \ell_{jt}) \) are zero, for \( \mathcal{F}(\partial_\tau \ell_{it}) \) and \( \mathcal{F}(\partial_\tau \ell_{jt}) \) are independent for all \( \rho \neq \tau \), and the properties of the inverse expected Hessian from Lemma D.1, \( \mathcal{F}(U^{(1b,1,2)}_{i}) = \mathcal{O}(1/\sqrt{N}) \), uniformly over \( i \), \( \mathcal{F}(U^{(1b,1,2)}_{i}) = \mathcal{O}(1/\sqrt{N}) \), uniformly over \( i \neq j \). This implies that \( \mathcal{F}(U^{(1b,1,2)}_{i,j}) = \mathcal{O}(1/N) \), uniformly over \( i \neq j \). This implies that \( \mathcal{F}(U^{(1b,1,2)}_{i}) = \mathcal{O}(1/\sqrt{N}) \), and therefore \( U^{(1b,1,2)} = \mathcal{O}(1) \). By similar arguments one obtains \( U^{(1b,p,q)} = \mathcal{O}(1) \) for all combinations of \( p, q = 1, 2, 3, 4 \), except for \( p = q = 1 \) and \( p = q = 4 \).

For \( p = 1 \),

\[ U^{(1b,1,1)}_{i} = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} U^{(1b,1,1)}_{i}, \]

\[ U^{(1b,1,1)}_{i} = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{j=1}^{T} \mathcal{F}(\partial_{\rho} \ell_{it}) \mathcal{F}(\partial_{\tau} \ell_{jt}) \sum_{i=1}^{T} \partial_\tau \ell_{rt} \sum_{j=1}^{T} \partial_\tau \ell_{jt}. \]

Analogous to the result for \( U^{(1b,1,2)} \), \( \mathcal{F}(U^{(1b,1,1)}_{i}) = \mathcal{O}(1/\sqrt{N}) \), and therefore \( U^{(1b,1,1)} = \mathcal{F}(U^{(1b,1,1)}_{i}, \mathcal{O}(1)) \). Furthermore,

\[ \mathcal{F}(U^{(1b,1,1)}_{i}) = \frac{1}{2\sqrt{NT}} \sum_{i=1}^{N} \sum_{j=1}^{T} \mathcal{F}(\partial_{\rho} \ell_{it}) \mathcal{F}(\partial_{\tau} \ell_{jt}) \sum_{i=1}^{T} \partial_\tau \ell_{rt} \sum_{j=1}^{T} \partial_\tau \ell_{jt}. \]

Analogously,

\[ U^{(1b,4,4)} = \mathcal{F}(U^{(1b,4,4)} + \mathcal{O}(1)) = -\sqrt{\frac{T}{N}} \sum_{i=1}^{N} \sum_{j=1}^{T} \mathcal{F}(\partial_{\rho} \ell_{it}) \sum_{j=1}^{T} \partial_\tau \ell_{jt}. \]

We have thus shown that \( U^{(1b)} = \mathcal{O}(1/\sqrt{N}) \), uniformly over \( i \neq j \). Since \( \mathcal{F}(\partial_{\rho} \ell_{at}) \) and \( \mathcal{F}(\partial_{\tau} \ell_{bt}) \) are independent for all \( \rho \neq \tau \), we thus conclude \( \mathcal{F}(\partial_{\rho} \ell_{at}) \) and \( \mathcal{F}(\partial_{\tau} \ell_{bt}) \) are independent for all \( \rho \neq \tau \).

\# We have shown \( \mathcal{F}(\partial_{\rho} \ell_{at}) \) and \( \mathcal{F}(\partial_{\tau} \ell_{bt}) \) are independent for all \( \rho \neq \tau \).

\[ \mathcal{F}(\partial_{\rho} \ell_{at}) \] and \( \mathcal{F}(\partial_{\tau} \ell_{bt}) \) are independent for all \( \rho \neq \tau \), and \( \mathcal{F}(\partial_{\rho} \ell_{at}) \) and \( \mathcal{F}(\partial_{\tau} \ell_{bt}) \) are independent for all \( \rho \neq \tau \). Then, part (ii) of Theorem C.1 yields \( \mathcal{F}(\partial_{\rho} \ell_{at}) \) and \( \mathcal{F}(\partial_{\tau} \ell_{bt}) \) are independent for all \( \rho \neq \tau \).

\[ \mathcal{F}(\partial_{\rho} \ell_{at}) \] and \( \mathcal{F}(\partial_{\tau} \ell_{bt}) \) are independent for all \( \rho \neq \tau \), and \( \mathcal{F}(\partial_{\rho} \ell_{at}) \) and \( \mathcal{F}(\partial_{\tau} \ell_{bt}) \) are independent for all \( \rho \neq \tau \).
Lemma D.1. \[ V^{(2)} = \lim_{N,T \to \infty} \left\{ \frac{r_T^2}{N^2T^2} \sum_{j \leq J} \sum_{i \leq I} \mathbb{E} \left[ \Delta h \Delta h_i \right] + \sum_{j \leq J} \mathbb{E} \left[ \Delta h_i \Delta h_i \right] \right\}. \]

Note that \( r_T \to \infty \) and \( r_T = O(\sqrt{NT}) \).

# Part (3): limit of \( r_T = O(\sqrt{NT}) \).

The conclusion of the Theorem follows because \((\delta - \delta T)^2 \) is of much larger order than the number of diagonal terms is of much larger order than the number of diagonal terms.

In panel models with individual effects, it is straightforward to determine the order of magnitude of \( \mathcal{H} \) in Assumption B.1(iv), because \( \mathcal{H} \) contains only the diagonal matrix \( \mathcal{H}(\alpha_{\alpha}) \).

We have that \( \mathcal{H} \) is no longer diagonal, but it has a special structure. The diagonal terms are order 1, whereas the off-diagonal terms are of order \( (NT)^{-1/2} \). Moreover, \( \mathcal{H} - \mathcal{H}(\alpha_{\alpha}) \) is of order \( (NT)^{-1/2} \). These observations, however, are not sufficient to establish the order of \( \mathcal{H} \) because the number of non-zero off-diagonal terms is of much larger order than the number of diagonal terms; compare \( \Theta(NT) \to \Theta(N + T) \). Note also that the expected Hessian without penalty term \( \mathcal{H} \) has the same structure as \( \mathcal{H} \) itself, but is not even invertible, i.e. the observation on the relative size of diagonal vs. off-diagonal terms is certainly not sufficient to make statements about the structure of \( \mathcal{H} \). The result of the following lemma is therefore not obvious. It shows that the diagonal terms of \( \mathcal{H} \) also dominate in determining the order of \( \mathcal{H} \).

Lemma D.1. Under Assumption 4.1,
\[ \left\| \mathcal{H}^{-1} - \text{diag} \left( \mathcal{H}(\alpha_{\alpha}) \right)^{-1} \right\|_{\text{max}} = O_p \left( (NT)^{-1/2} \right) . \]

The proof of Lemma D.1 is provided in the supplementary material (see Appendix E). The lemma result establishes that \( \mathcal{H}^{-1} \) can be uniformly approximated by a diagonal matrix, which is given by the inverse of the diagonal terms of \( \mathcal{H} \) without the penalty. The diagonal elements of \( \text{diag}(\mathcal{H}(\alpha_{\alpha}))^{-1} \) are of order 1, i.e. the order of the difference established by the lemma is relatively small.

Note that the choice of penalty in the objective function is important to obtain Lemma D.1. Different penalties, corresponding to different relaxations (e.g. a penalty proportional to \( \alpha_i^2 \), corresponding to the normalization \( \alpha_i = 0 \)), would fail to deliver Lemma D.1. However, these alternative choices do not affect the estimators \( \hat{\beta} \) and \( \hat{\delta} \), i.e. which normalization is used to compute \( \hat{\beta} \) and \( \hat{\delta} \) in practice is irrelevant (up to numerical precision errors).

Appendix E. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jeconom.2015.12.014.
Pesaran, M.H., 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica 74(4), 967–1012.
Phillips, P.C.B., Moon, H., 1999. Linear regression limit theory for nonstationary panel data. Econometrica 67(5), 1057–1111.

Pratt, J.W., 1981. Concavity of the log likelihood. J. Amer. Statist. Assoc. 76 (373), 103–106.
Woutersen, T., 2002. Robustness against incidental parameters (unpublished manuscript).