Stammann, Amrei; Heiß, Florian; McFadden, Daniel

Conference Paper

Estimating Fixed Effects Logit Models with Large Panel Data

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel - Session: Microeconometrics, No. G01-V3

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Estimation of Fixed Effects Logit Models with Large Panel Data

Amrei Stammann[*], Florian Heiss[*] and Daniel McFadden[**]

[*]Heinrich-Heine University Düsseldorf
[**]University of California, Berkeley

— Preliminary Draft – Aug 24, 2016 —

Parametric panel data logit models with individual time-invariant effects can be estimated by either the conditional or the unconditional fixed effects maximum likelihood estimator. The conditional fixed effects logit estimator is consistent under the usual assumptions but it has the drawback that it does not deliver estimates of the fixed effects or partial effects. It is also computationally costly if the number of observations per individual T is large. The unconditional fixed effects logit estimator can be implemented as a standard logit estimator with a dummy variable for each observational unit. It is biased for small T due to the incidental parameters problem, but bias corrections have been suggested. Another drawback of this approach is that the computational costs can be prohibitive with a large number of individuals N. This paper revisits an approach suggested by Chamberlain (1980) and Greene (2004) that makes use of the sparseness of the Hessian matrix to relieve the computational burden imposed by brute-force dummy variable regression. We show that in the context of logit models, the approach is equivalent to an intuitive pseudo-demeaning algorithm. We combine the pseudo-demeaning algorithm with a bias-correction proposed by Hahn and Newey (2004) to deal with the incidental parameter bias. Extensive Monte-Carlo simulations show that the bias-corrected parameter estimator has similar properties as the conditional logit estimator. Its computational burden is much lower, especially with relatively large T, and we can directly estimate partial effects. We offer this algorithm as an implementation in the R-package `bife`.

1. Introduction

The fixed effects logit model is a popular specification for panel data analyses of binary variables. It allows for unobserved time-invariant individual heterogeneity like the variation in tastes with an arbitrary distribution. While the technical implementation of the fixed effects estimator is rather simple in the linear case, the within transformation based on individual demeaning does not carry over to nonlinear models like the logit model.

A possible approach for nonlinear models is maximum likelihood estimation with a dummy variable for each cross-sectional unit. To stress the difference to the conditional logit estimator, we call this the unconditional logit (UCL) estimator. It can become computationally challenging when the number of fixed effects N is large since it requires the computation and inversion of a large Hessian. Apart from the computational challenge, the parameters of usual fixed effects models with a small number of observations per fixed effect T suffer from the incidental parameters problem, first noted by Neyman and Scott (1948). The estimators are inconsistent as N increases and T is held constant. Even increasing T does not necessary solve the incidental parameter bias because fixed effects estimators are asymptotically biased even if T grows at the same rate as N (Hahn and Newey 2004).

If the incidental parameter bias is of concern, often the conditional logit estimator (CL) is proposed as an alternative. It has been derived by Andersen (1970) and later generalized by Chamberlain (1980) as a solution to the incidental parameters problem. However, the CL estimator has the drawback that it does not deliver estimates of the fixed effects. Also partial or marginal effects cannot be easily and consistently estimated. The CL estimator is computationally very costly if the number of observations per individual T is large. Even if using a more efficient recursion method proposed by Gail, Lubin and Rubinstein (1981), the computational burden increases roughly quadratically with the number of individual observations T.

For these reasons, the UCL estimator is still of interest in many relevant cases. We discuss and tackle its remaining problems: the computational burden with large N panels and the bias with small T panels. The Hessian for the fixed effects has a specific sparse structure which can be exploited using tools for the partitioned inverse to dramatically decrease the computational costs, see Hall (1978), Prentice and Gloeckler (1978), Chamberlain (1980) and Greene (2004). We show that for fixed effects logit models, we can rewrite the problem in an intuitive way using an iterative pseudo-demeaning algorithm. Importantly, the computational burden increases only linearly with the number of individual observations T.

This paper combines this pseudo-demeaning algorithm with a bias-correction to deal with the incidental parameters problem for small T and call the resulting es-

timator bias-corrected logit (BCL) estimator. There is a branch of literature that deals with bias-corrections to reduce the incidental parameter bias of structural parameters and/or partial effects in non-linear fixed-effects models, see for example Hahn and Newey (2004), Arellano and Hahn (2006), Fernández-Val (2009), Hospido (2012), and Dhaene and Jochmans (2015). We implement the approach of Hahn and Newey (2004) because is allows the correction of both the parameter estimates and the partial effects, is computationally less demanding than for example jack-knife approaches and performs well.

The resulting parameter estimator has desirable properties comparable to the CL estimator and directly deliver estimates of the fixed effects and the individual and average partial effects. As we show in theory and simulations, it can also be computationally more efficient than the common estimators by orders of magnitude for larger panel dimensions N and/or T.

In an extensive simulation study we examine the UCL, BCL and CL estimators with respect to the bias of structural parameters, average partial effects and estimated standard errors, rejection frequencies and their computational complexity. We can confirm the findings of Greene (2004) of large biases in the UCL estimator for small T, but can also show that the BCL estimator improves these biases substantially.

Our Monte-Carlo experiments suggest that the BCL estimator has similar desirable properties as the CL estimator in terms of the distribution of the parameter estimates. Additionally, the BCL estimator improves the rejection frequencies and reduces the variance of the estimator. Besides, we find, that for small T average partial effects computed from the BCL estimates are less biased compared to the ones computed from UCL estimates. We confirm that the computational burden of the UCL and BCL estimators increase linearly with T, whereas the burden of CL increases quadratically which makes a dramatic difference for large T.

In order to make our pseudo-demeaning algorithm accessible for applied work, we offer it in the R-package `bife`.[1] This implementation allows for fast estimation of structural parameters and average partial effects of fixed effects logit and probit models with pseudo-demeaning and bias-corrected pseudo-demeaning.

The paper is organized as follows. Section 2 presents a short recap of the fixed effects logit models along with the UCL and CL estimators. Section 3 introduces the pseudo-demeaning approach and the full algorithm with bias-correction. It follows a discussion of how average partial effects can be computed for the UCL, BCL and CL estimator in Section 4. In Section 5, the design and results of a series of Monte Carlo simulations are presented before Section 6 concludes.

[1] See `https://cran.r-project.org/web/packages/bife/`.

2. The fixed effects logit model

For the sake of notational simplicity, assume we have a balanced panel of $i = 1, \ldots, N^*$ individuals for $t = 1, \ldots, T$ time periods. The same type of model applies to situations where we include fixed effects for N^* groups of size T each. Suppose we observe the discrete dependent variable y_{it}. It usually is a binary choice variable, such that $y_{it} = 1$ if an event occurs (case) and $y_{it} = 0$ if it does not occur (control). Define $N = \sum_{i=1}^{N^*} 1[0 < \sum_{t=1}^{T} y_{it} < T]$ as the number of cross-sectional units for which y_{it} varies over time. The $N^* - N$ individuals without varying y_{it} do not contribute to the identification or estimation of the fixed effects logit model and can be dropped from the analysis without affecting the estimator of the structural parameters.

The fixed effects logit model is defined by the logistic probability of y_{it}

$$f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i) = p_{it}^{y_{it}}(1 - p_{it})^{1-y_{it}} \tag{2.1}$$

with

$$p_{it} = \Pr(y_{it} = 1|\mathbf{x}_{it}, \alpha_i, \boldsymbol{\beta}) = \frac{1}{1 + e^{-\alpha_i - \mathbf{x}_{it}\boldsymbol{\beta}}}$$

$$y_{it} = 1[\alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + \epsilon_{it} > 0]$$

where $\boldsymbol{\beta}$ is the $(M \times 1)$ parameter vector of the M regressors $\mathbf{x}_{it}$ and ϵ_{it} is the logistically distributed error term, $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The parameter α_i is called a fixed effect if $E(\epsilon_{it}|\mathbf{x}_i, \alpha_i) = 0$, $\mathbf{x}_i = \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}$ and arbitrary correlations between the fixed effects and the regressors are allowed.

We restrict ourselves to the parametric estimation techniques of the conditional and the unconditional fixed effects logit estimator. The following two subsections depict the advantages and drawbacks of both estimators before we combine the best properties of both to a straightforward estimation procedure in section 3.

2.1. The conditional logit estimator

The idea of the conditional logit estimator is to condition the fixed effects out of the objective function such that the inconsistency of the fixed effects in case of a unconditional maximum likelihood estimation does not contaminate the estimates of the structural parameters $\boldsymbol{\beta}$. Thus, the CL estimator is consistent as $N \to \infty$ and T is held fix. Secondly, the CL estimator facilitates the computational burden, if N is large because it abstains from the estimation of possible thousands of fixed effects. However, this computational advantage in N can become also a drawback; the CL estimator does not deliver estimates of the fixed effects, although these become

important if the researcher is for instance interested in partial effects. Additionally, the CL estimator cannot retain its computational advantage in large N if T is large. In fact, the CL estimator becomes computationally costly if T is large even if one uses a clever recursion algorithm proposed by Gail et al. (1981).

Computational issues

Next, we demonstrate the computational complexity of the brute-force and recursive algorithm of the CL estimator. We consider the general case of unbalanced panel data. Suppose we observe individual i for T_i periods of which t_{1i} are cases ($y_{it} = 1$) and t_{0i} are controls ($y_{it} = 0$) such that $T_i = t_{0i} + t_{1i}$.

The objective function of the conditional logit estimator is derived by conditioning the density of y_{it} on the sufficient statistic $t_{i1} = \sum_{t=1}^{T_i} y_{it}$. Thereby, the fixed effects α_i are eliminated from the log-likelihood (Chamberlain 1980):

$$\max_{\beta} \quad L_c(\boldsymbol{\beta}) = \sum_{i=1}^{N} \ln \frac{\exp\left(\sum_{t=1}^{T_i} \mathbf{x}_{it} y_{it} \boldsymbol{\beta}\right)}{\sum_{d \in B_i} \exp\left(\sum_{t=1}^{T_i} \mathbf{x}_{it} d_t \boldsymbol{\beta}\right)} \tag{2.2}$$

with

$$B_i = \{d = (d_1, ..., d_T) | d_t = 0 \text{ or } 1 \text{ and } \sum_{t=1}^{T} d_t = \sum_{t=1}^{T_i} y_{it}\}$$

B_i defines the alternative set for individual i that consists of c_i possible selections of $1, \ldots, t_{1i}$ case status from $1, \ldots, T_i$ subjects:

$$c_i = \binom{T_i}{t_{1i}} = \frac{T_i!}{t_{1i}!(T_i - t_{1i})!} \tag{2.3}$$

To keep things simple let us consider the contribution L_i of group i to the conditional likelihood:

$$\exp(L_i) = \frac{\prod_{k=1}^{t_{1i}} \exp(\mathbf{x}_k \boldsymbol{\beta})}{\sum_{h=1}^{c_i} \prod_{k_h=1}^{t_{1i}} \exp(\mathbf{x}_{hk_h} \boldsymbol{\beta})} \tag{2.4}$$

The index k denotes the observed data and the index k_h the h-th possible assignment.

A direct evaluation of the denominator in (2.4) requires the summation of c_i terms and becomes prohibitive if T_i increases (Gail et al. 1981). For instance, the computation of the denominator for a single individual with $T_i = 100$ and $t_{i1} = 5$ requires $\binom{100}{5} \approx 7.53 \cdot 10^6$ evaluations and for an individual with $T_i = 100$ and $t_{i1} = 50$ even $\binom{100}{50} \approx 1.01 \cdot 10^{29}$ evaluations.

However, the calculation of the conditional likelihood (2.2) can be accelerated by the implementation of a recursive calculation proposed by Gail et al. (1981) without loosing the exactness of the brute force approach in (2.2). Therefore, consider again

the denominator in (2.4)

$$f_i(t_{1i}, T_i) = \sum_{h=1}^{c_i} \prod_{k_h=1}^{t_{1i}} \exp(\mathbf{x}_{hk_h}\boldsymbol{\beta}) = \sum_{h=1}^{c_i} \prod_{k_h=1}^{t_{1i}} U_{k_h} \tag{2.5}$$

The recursion is defined by

$$f_i(t_{1i}, T_i) = f_i(t_{1i}, T_i - 1) + U_{T_i} f_i(t_{1i} - 1, T_i - 1) \tag{2.6}$$

with $f_i(0, T_i) = 1$ for $T_i \geq 0$ and $f_i(t_{1i}, T_i) = 0$ for $t_{1i} > T_i$.

The recursion reduces the number of arithmetic operations per individual from $(T_i!/(T_i - t_{1i})!(t_{1i} - 1)!) - 1$ to $2t_{1i}(T_i - t_{1i} + 1)$ operations (Gail et al. 1981). Finally, the conditional log-likelihood in (2.2) can be rewritten to

$$L = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} \left(\sum_{t=1}^{T_i} y_{it}\mathbf{x}_{it}\boldsymbol{\beta} - \ln f_i(t_{1i}, T_i) \right) \tag{2.7}$$

The maximization of the conditional log-likelihood (2.7) has in general no explicit solution. It is solved iteratively with gradient based maximization techniques. Let $\mathbf{H_c}$ define the Hessian and $\mathbf{g_c}$ the gradient vector. Each entry of $\mathbf{H_c}$ and $\mathbf{g_c}$ requires $O(\sum_{i=1}^{N} t_{1i}(T_i - t_{1i}))$ operations[2] (Reid and Tibshirani 2014). The computation of the Hessian is the most demanding part since it requires the computation of M^2 entries. Altogether, the computational complexity of the recursive algorithm requires $O(M^2 \sum_{i=1}^{N} t_{1i}(T_i - t_{1i})))$ if $N \gg T \gg M$.[3] This already suggests an upper bound of the computational complexity if $t_{1i}/T_i = 0.5 \, \forall i$. The computation complexity is linear in N and roughly quadratic in T_i since t_{1i} itself is a linear function of T_i.

2.2. Unconditional logit estimation via dummy variables

The UCL estimator is only T-consistent, but it has the advantage over the CL estimator, that it produces estimates of the fixed effects which are necessary to compute partial effects. Additionally, we demonstrate that unlike the CL estimator, the computational burden of the UCL estimator is linear in T if we use the pseudo-demeaning proposed in section 3. Since the incidental parameter problem vanishes as T increases, the UCL estimator might be preferable to the CL estimator in applications with large T.

[2] There are $O(t_{1i}(T_i - t_{1i}))$ operations for each of the N individuals.

[3] The inversion of the $(M \times M)$ Hessian costs $O(M^3)$. $O(M^2 \sum_{i=1}^{N} t_{1i}(T_i - t_{1i}))) > O(M^3)$ assuming $N \gg T \gg M$.

Computational issues

The unconditional logit estimator maximizes the log-likehood function

$$\begin{aligned} \max_{\alpha,\beta} \qquad L(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \sum_{i=1}^{N} \sum_{t=1}^{T} \ln f(y_{it} | \mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i) \\ &= \sum_{i=1}^{N} \sum_{t=1}^{T} y_{it} \ln(p_{it}) + (1 - y_{it}) \ln(1 - p_{it}) \end{aligned} \tag{2.8}$$

with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$. The maximization problem (2.8) of the UCL estimator has in general no explicit solution. Thus, it is solved iteratively with gradient based maximization techniques.

The brute-force approach to estimate a non-linear fixed effects model with unconditional maximum likelihood is to put a dummy for each fixed effect. This delivers a $(NT \times (N + M))$ regressor matrix $\mathbf{Z} = (\mathbf{D}, \mathbf{X})$, where $\mathbf{D}$ denotes the $(NT \times N)$ dummy variable matrix and $\mathbf{X}$ the $(NT \times M)$ matrix of the remaining regressors.[4] Following a similar notation as Greene (2004), we define the gradient and Hessian of the log-likelihood given in (2.8) and use Newton Raphson for its maximization.

The $((N + M) \times 1)$ gradient:

$$\mathbf{g} = \begin{pmatrix} \mathbf{g}_\alpha \\ \mathbf{g}_\beta \end{pmatrix} \tag{2.9}$$

with

$$\mathbf{g}_\beta = \frac{\partial L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{x}_{it}(y_{it} - p_{it})$$

$$\mathbf{g}_{\alpha_i} = \frac{\partial L}{\partial \alpha_i} = \sum_{t=1}^{T} (y_{it} - p_{it})$$

The $(N + M) \times (N + M)$ Hessian:

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{\beta\beta} & \mathbf{h}_{\beta\alpha_1} & \mathbf{h}_{\beta\alpha_2} & \cdots & \mathbf{h}_{\beta\alpha_N} \\ \mathbf{h}_{\alpha_1\beta} & h_{\alpha_1\alpha_1} & 0 & \cdots & 0 \\ \mathbf{h}_{\alpha_2\beta} & 0 & h_{\alpha_2\alpha_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_{\alpha_N\beta} & 0 & 0 & \cdots & h_{\alpha_N\alpha_N} \end{pmatrix} = \begin{pmatrix} \mathbf{H}_{\beta\beta} & \mathbf{H}_{\beta\alpha} \\ \mathbf{H}_{\alpha\beta} & \mathbf{H}_{\alpha\alpha} \end{pmatrix} \tag{2.10}$$

[4] Please note, that some software routines, such as `glm` in R, include N^* dummies instead of N and computation becomes even more costly. Remember, $N = \sum_{i=1}^{N^*} 1[0 < \sum_{t=1}^{T} y_{it} < T]$ where N^* denotes the total number of individuals in the dataset.

with

$$\mathbf{H}_{\beta\beta} = \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{\partial^2 \ln L}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = (-1)\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbf{x}'_{it}\mathbf{x}_{it}p_{it}(1-p_{it})$$

$$\mathbf{h}_{\beta\alpha_i} = \sum_{t=1}^{T} \frac{\partial^2 \ln L}{\partial\boldsymbol{\beta}\partial\alpha_i} = (-1)\sum_{t=1}^{T} \mathbf{x}_{it}p_{it}(1-p_{it})$$

$$h_{\alpha_i\alpha_i} = \sum_{t=1}^{T} \frac{\partial^2 \ln L}{\partial^2\alpha_i} = (-1)\sum_{t=1}^{T} p_{it}(1-p_{it})$$

Further, define the $((N+M)\times 1)$ parameter vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})'$. The naive approach uses Newton's method with the gradient and Hessian defined in (2.9) and (2.10). The k-th iteration step is

$$\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1} = -\mathbf{H}_{k-1}^{-1}\mathbf{g}_{k-1} \tag{2.11}$$

The Hessian is a $((N+M)\times(N+M))$ matrix. If the number of fixed effects N is large, its computation and inversion becomes computationally costly. It can be easily seen, that the dummy variable approach in (2.11) would require a computational complexity of $O(N^3T^2)$ for a balanced panel. To see this, we reformulate (2.11) as an iteratively re-weighted least squares (IRWLS) problem. Therefore, we define the $(NT\times(N+M))$ regressor matrix $\mathbf{Z} = (\mathbf{D}, \mathbf{X})$, where $\mathbf{D}$ denotes the $(NT\times N)$ dummy variable matrix. The k-th IRWLS-iteration step in (2.11) can be rewritten into

$$\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1} = -\mathbf{H}_{k-1}^{-1}\mathbf{g}_{k-1} = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y}-\mathbf{p}) \tag{2.12}$$

where $\mathbf{W}$ serves as a $(NT\times NT)$ weighting matrix. $\mathbf{W}$ is a diagonal-matrix $\mathrm{diag}(w_{it})$ with weights $w_{it} = p_{it}(1-p_{it})$.

The most demanding part is the computation of the $((N+M)\times(N+M))$ Hessian. Especially, the multiplication of the $((N+M)\times NT)$ matrix $\mathbf{Z}'$ with the $(NT\times NT)$ matrix $\mathbf{W}$ is computationally extensive. It is well know, that the multiplication of a general $(R\times S)$ matrix with a general $(S\times S)$ matrix requires $O(RS^2)$ time. Hence, the multiplication $\mathbf{Z}'\mathbf{W}$ costs $O((N+M)N^2T^2)$ what is asymptotically approximately $O(N^3T^2)$, assuming $N \gg T \gg M$.[5]

[5] Even a more elegant implementation that abstracts from the usage of the matrix $\mathbf{W}$ would require $O(N^3T)$ time. This implementation uses a transformed matrix $\mathbf{Z_w}$ instead, where each element of $\mathbf{Z}$ has been multiplied with its corresponding weight w_{it}, such that $\mathbf{Z}'\mathbf{W}\mathbf{Z} = \mathbf{Z_w}'\mathbf{Z}$. It is well known that matrix multiplication $\mathbf{Z_w}'\mathbf{Z}$ costs $O((N+M)^2NT) \approx O(N^3T)$, matrix multiplication $\mathbf{Z}'\mathbf{Y}$ costs $O((N+M)NT) \approx O(N^2T)$, matrix inversion $(\mathbf{Z_w}'\mathbf{Z})^{-1}$ costs $O((N+M)^3) \approx O(N^3)$ and finally the product of the Hessian and the gradient costs $O((N+M)^2) \approx O(N^2)$, assuming $N \gg T \gg M$. Further, $O(N^3T) > O(N^2T) > O(N^2)$ for $T > 1$. The symbol $\approx$ means "asymptotically approximately".

3. Computationally efficient unconditional logit estimation

Greene (2004) and Chamberlain (1980), among others, propose an algorithm which avoids the inversion of the large Hessian in (2.11). Their method utilizes the partitioned inverse approach and exploits the sparsity of the Hessian. A detailed description of that algorithm can be found in the Appendix A.1.

We propose a different approach to reduce the dimensionality of the maximization problem in (2.8) that produces the identical parameter estimates as the dummy variable and partitioned inverse approaches. The basic idea is to reformulate the maximization problem as an iteratively reweighted least squares problem and to eliminate the fixed effects from the resulting estimation equation similar to linear fixed effects models.

The reformulation of a regression-like problem even allows to apply the well-known Frisch-Waugh-Lovell theorem for demeaning (see (Frisch and Waugh 1933), (Lovell 1963)). The final estimation formulas for the parameters are closely related to (Chamberlain 1980) who derived them with the partitioned inverse. However, in contrast to Greene (2004) and Chamberlain (1980), our final estimation equations directly translates into an intuitive pseudo-demeaning. The same methodology applies also to other non-linear models like probit and poisson models.

3.1. Pseudo-Demeaning

We use the naive dummy variable approach in section (2.2) as a starting point for further computational simplifications. Consider again the k-th IRWLS-iteration step as in (2.12)

$$\begin{aligned}\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1} &= (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y}-\mathbf{p}) \\ &= (\mathbf{Z}'\sqrt{\mathbf{W}}'\sqrt{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}'\sqrt{\mathbf{W}}'\sqrt{\mathbf{W}}^{-\mathbf{1}}(\mathbf{y}-\mathbf{p}) \end{aligned} \tag{3.1}$$

Remember that $\mathbf{W}$ is a diagonal-matrix with weights $w_{it} = p_{it}(1-p_{it})$ on the main diagonal. So (3.1) is equivalent to a regression of the dependent variable $\frac{y_{it}-p_{it}}{\sqrt{(p_{it}(1-p_{it}))}}$ on the independent variables $\sqrt{(p_{it}(1-p_{it}))}\mathbf{z}_{it}$. Reformulated as a regression model for individual i, (3.1) becomes

$$\tilde{y}_{it} = \tilde{w}_{it}(\alpha_i^k - \alpha_i^{k-1}) + \tilde{\mathbf{x}}_{it}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) \tag{3.2}$$

with

$$\begin{aligned}\tilde{y}_{it} &= \frac{y_{it} - p_{it}}{\sqrt{p_{it}(1-p_{it})}} \\ \tilde{w}_{it} &= \sqrt{w_{it}} = \sqrt{p_{it}(1-p_{it})} \\ \tilde{\mathbf{x}}_{it} &= \tilde{w}_{it}\mathbf{x}_{it} = \sqrt{p_{it}(1-p_{it})}\mathbf{x}_{it}\end{aligned}$$

The following results are derived in detail in Appendix A.4. The update of the fixed effects $(\alpha_i^k - \alpha_i^{k-1})$ can be removed from the estimation equation by inserting the following expression into (3.2)

$$(\alpha_i^k - \alpha_i^{k-1}) = \frac{\sum_{t=1}^{T} \tilde{w}_{it}\tilde{y}_{it}}{\sum_{t=1}^{T} \tilde{w}_{it}^2} - \frac{\sum_{t=1}^{T} \tilde{w}_{it}\tilde{\mathbf{x}}_{it}}{\sum_{t=1}^{T} \tilde{w}_{it}^2}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) \tag{3.3}$$

We get the pseudo-demeaned estimation equation

$$\underbrace{\tilde{y}_{it} - \tilde{w}_{it}\frac{\sum_{t=1}^{T} \tilde{w}_{it}\tilde{y}_{it}}{\sum_{t=1}^{T} \tilde{w}_{it}^2}}_{\bar{y}_{it}} = \underbrace{\left(\tilde{\mathbf{x}}_{it} - \tilde{w}_{it}\frac{\sum_{t=1}^{T} \tilde{w}_{it}\tilde{\mathbf{x}}_{it}}{\sum_{t=1}^{T} \tilde{w}_{it}^2}\right)}_{\bar{\mathbf{x}}_{it}}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) \tag{3.4}$$

and finally, the formula for the $\boldsymbol{\beta}$ updates becomes

$$(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \bar{\mathbf{x}}_{it}'\bar{\mathbf{x}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} \bar{\mathbf{x}}_{it}'\bar{y}_{it}\right) \tag{3.5}$$

Lets denote the converged $\boldsymbol{\beta}^k$ in (3.5) and α_i^k in (3.3) with $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}_i$.

The formula (3.5) with pseudo-demeaning directly translates in an intuitive approach that reminds of demeaning in a linear regression model. In fact, the pseudo-demeaning (3.5) is equivalent to a regression of a weighted demeaned dependent variable $\bar{\mathbf{y}}$ on weighted demeaned regressors $\bar{\mathbf{x}}$.

Computational complexity

In contrast to the CL estimator, the computational complexity of the pseudo-demeaning is linear in T_i and N. It can be expressed by $O(NT)$ if the panel is balanced $T_i = T \quad \forall i$ and $O(\sum_{i=1}^{n} T_i)$ if it is unbalanced, neglecting the number of structural parameters M. This can be easily derived from the computationally most intensive operation which is the computation of $\boldsymbol{\beta}$ in the formula (3.5).

Consider the balanced panel case and define $\bar{\mathbf{X}}$ as the regressor matrix in 3.5 and $\bar{\mathbf{y}}$ as the dependent variable vector in (3.5). The computation of $\bar{\mathbf{X}}$ and $\bar{\mathbf{y}}$ requires asymptotically approximate $O(NT)$ time, the matrix multiplication $\bar{\mathbf{X}}'\bar{\mathbf{X}}$ costs asymptotically approximate $O(M^2NT)$, matrix multiplication $\bar{\mathbf{X}}'\bar{\mathbf{Y}}$ costs asymptotically approximate $O(MNT)$, matrix inversion $(\bar{\mathbf{X}}'\bar{\mathbf{X}})^{-1}$ costs asymptotically approximate $O(M^3)$ and finally the product of the Hessian and the gradient costs asymptotically approximate $O(M^2)$, assuming $N \gg T \gg M$. $O(M^2NT) > O(MNT) > O(M^3)$ if $N \gg T \gg M$. Altogether, the computation time of the pseudo-demeaning is linear in T and N.

3.2. The complete algorithm

The algorithm presented above can be concisely summarized in the following pseudo code:

1. Get starting values for $\boldsymbol{\beta}^k$ and α_i^k, $i = 1, ..., N$, $k = 0$,[6]
2. For $k = 1, ..., k^{max}$:
3. Given $\boldsymbol{\beta}^{k-1}$ and α_i^{k-1} compute p_{it}^{k-1} as in formula (2.1),
4. Given p_{it}^{k-1} compute $(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})$ with the pseudo-demeaning formula (3.5),
5. Given $(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})$ and p_{it}^{k-1} compute $(\alpha_i^k - \alpha_i^{k-1})$ with formula (3.3),
6. If $||\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}|| < \epsilon$, stop.

If T is small, the UCL estimator suffers from the incidental parameters bias. Appendix A.2 discusses the algorithm of Hahn and Newey (2004) for correction of the parameter estimates. The combination with a bias-correction makes the pseudo-demeaning approach attractive even in situations where N is large and T is small especially if the CL estimator is no alternative since partial effects are of interest.

To include the Hahn and Newey (2004) bias correction in the pseudo code, let $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}_i$ denote the converged (but biased) estimates of the Newton-Raphson algorithm described in steps 1 to 6 and add the following two steps:

7. Given $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}_i$ compute the bias-corrected coefficient $\tilde{\boldsymbol{\beta}}$ with formula (A.12),
8. Given $\tilde{\boldsymbol{\beta}}$ compute $\tilde{\alpha}_i$ using the Newton-Raphson algorithm (holding $\tilde{\boldsymbol{\beta}}$ fixed) as described in (A.14).

This is the algorithm implemented in the R-package `bife`.

The computational complexity of the bias-correction is asymptotically approximate equivalent to the one of the pseudo-demeaning. Hence, the bias-corrected

[6] Choose $\boldsymbol{\beta}^{k=0} = \alpha_i^{k=0} = 0$ or use the starting values of the model estimated as a linear probability model combined with the linear fixed effects estimator.

estimator costs $O(NT)$ if the panel is balanced $T_i = T \quad \forall i$ and $O(\sum_{i=1}^{n} T_i)$ if it is unbalanced, neglecting the number of structural parameters M.

3.3. Standard errors

For applied work, the full covariance matrix of all parameters including the fixed effects will be of minor interest. If N is large, this computation becomes prohibitive. Instead of computing the standard errors of the parameters as the inverse of the full Hessian in (2.10), we can calculate the standard errors of the structural parameters $\boldsymbol{\beta}$ can be easily obtained as the square-root of the diagonal of the concentrated inverse Hessian

$$V(\boldsymbol{\beta}) = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \bar{x}'_{it} \bar{x}_{it} \right)^{-1} \tag{3.6}$$

If this is of interest, the standard error of the fixed effect α_i can be obtained as the square root of the variance given by

$$V(\alpha_i) = \frac{1}{\sum_{t=1}^{T} \tilde{w}_{it}^2} + \frac{\sum_{t=1}^{T} \tilde{w}_{it} \tilde{x}_{it}}{\sum_{t=1}^{T} \tilde{w}_{it}^2} V(\boldsymbol{\beta}) \left(\frac{\sum_{t=1}^{T} \tilde{w}_{it} \tilde{x}_{it}}{\sum_{t=1}^{T} \tilde{w}_{it}^2} \right)' \tag{3.7}$$

4. Average Partial Effects

Average partial effects are often of major interest for applied work, since the estimates of the structural parameters obtained with (3.5) are not directly interpretable.[7] In this section we demonstrate how to obtain estimated average partial effects for UCL, BCL and CL estimator. For the CL estimator the computation of average partial effects is not straightforward. We discuss a widely used but inconsistent approach and propose a different strategy to obtain average partial effects for the CL estimator.

When computing average partial effects we distinguish between continuous and discrete regressors. Let $G(\cdot)$ denote the logistic cdf and $g(\cdot)$ denote the logistic pdf. For the continuous case we define the individual partial effect at time t of the k-th regressor on the conditional probability that $y_{it} = 1$

$$\begin{aligned} m_k(\mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i) &= \frac{\partial \Pr(y_{it} = 1 | \mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i)}{\partial x_{itk}} \\ &= \frac{\partial G(\alpha_i + \mathbf{x}_{it}\boldsymbol{\beta})}{\partial x_{itk}} \\ &= g(\alpha_i + \mathbf{x}_{it}\boldsymbol{\beta})\beta_k \end{aligned}$$

[7] Sometimes partial effects are also referred to as marginal effects.

Averaging all individual partial effects delivers the average partial effect of the k-th continuous regressor x_k

$$APE_k = \frac{1}{T}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} m_k(\mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i) \tag{4.1}$$

For the discrete case we define the individual partial effect at time t of the j-th regressor as a one-unit increase in the regressor on the conditional probability that $y_{it} = 1$

$$\begin{aligned}\tilde{m}_j(\mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i) &= \Pr(y_{it} = 1 | x_{itj} = 1, \mathbf{x}_{it\{-j\}}, \boldsymbol{\beta}, \alpha_i) - \Pr(y_{it} = 1 | x_{itj} = 0, \mathbf{x}_{it\{-j\}}, \boldsymbol{\beta}, \alpha_i) \\ &= G(\alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} | x_{itj} = 1) - G(\alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} | x_{itj} = 0)\end{aligned}$$

The average partial effect of the j-th discrete regressor x_j is defined as

$$APE_j = \frac{1}{T}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{m}_j(\mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i) \tag{4.2}$$

To get estimated average partial effects the parameters $\boldsymbol{\beta}$ and α_i in formula (4.1) and (4.2) are replaced by their estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}_i$. This is straightforward with the UCL and BCL estimators since they directly deliver estimates for all parameters.

The CL estimator delivers no estimates of the fixed effects and provides no guidance towards the computation of average partial effects. Thus, we discuss two approaches to obtain estimated average partial effects for the CL estimator. Define $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$ as the estimates of the structural parameters obtained with the CL estimator.

The first approach is most widely used. It simply replaces the structural parameters in (4.1) and (4.2) with $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$ and sets the fixed effects with zero ($\hat{\alpha}_i = 0 \, \forall i$). This method is for example implemented in Stata's post-estimation routines for `clogit` and `xtlogit`.

We propose a different approach to compute average partial effects for the CL estimator. It re-calculates the fixed effects with the first-order condition of the UCL estimator. To be more precise, in order to recover estimates of fixed effects the algorithm proposed in section A.3 can be applied. Therefore, replace the linear predictor with $\mathbf{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$.

Our simulations show that this second approach performs well whereas the simpler approach to replace the α parameters with zeros can lead to substantial biases.

5. A Simulation Study

We analyse the behavior of the unconditional logit estimator (UCL), the bias-corrected unconditional logit estimator (BCL) and the conditional logit estimator (CL) in a simulation study. The focus of our analysis are the biases of the estimators, the biases of the average partial effects (APE) and the computation time.

To justify large values of T, we abstract from the classical econometricians view of panel data. Instead of observing N individuals for T time periods, assume, that T can be also considered as group size and N as number of groups. For example, fixed effects logit models can be also useful if i represents ZIP code areas and t is an index of individuals.[8]

5.1. The setup

The simulation setup follows Greene (2004).[9] We examine the data generating process

$$y_{it} = \mathbf{1}[w_{it} + v_{it} > 0], \tag{5.1}$$

where $v_{it} = \log[u_{it}/(1-u_{it})]$, $u_{it} \sim U(0,1)$.
with the index function

$$w_{it} = \alpha_i + \beta x_{it} + \delta d_{it}, \tag{5.2}$$

where

$$\begin{aligned} \beta &= \delta = 1, \\ x_{it} &\sim N(0, 1^2), \\ d_{it} &= \mathbf{1}[x_{it} + h_{it} > 0], & h_{it} &\sim N(0, 1^2) \\ \alpha_i &= \sqrt{T}\bar{x}_i + a_i, & a_i &\sim N(0, 1^2). \end{aligned}$$

Throughout our experiments, we consider several model specifications with different group sizes T and number of groups N^*. To be more specific, we vary the number of groups ($N^* = 100, 1000$) and the group sizes ($T = 4, 8, 10, 12, 16, 20, 50, 100, 200$), such that we get 18 model specifications. Each of these specifications is fitted $1,000$ times, at which for each fit α_i, x_{it} d_{it} and y_{it} vary.

[8] In the vignette of the R-package `bife` we estimated a fixed effects logit model to analyse the labor force participation of $662,775$ married women in $N = 51$ states, where the smallest state consists of $T_{min} = 855$ women and the largest of $T_{max} = 74,752$ women. This application demonstrates the advantage of pseudo-demeaning over the brute-force dummy variable approach and over conditional logit estimation, see `https://cran.r-project.org/web/packages/bife/vignettes/bife_introduction.html` .

[9] We also did robustness-checks on different parameter values, number of iterations and another simulation setup, that was inspired by Coupé (2005).

5.2. Biases

The number of groups N^* turns out not to have any relevant effect on our findings in terms of the bias in the estimates and average partial effects. However, as expected standard errors become tighter as N^* increases.

Table 1 list the means of the empirical sampling distributions of the three estimators for group size $N^* = 100$ (table B.1 and B.3 respectively for $N = 1,000$). Irrespective of the number of groups N^*, we observe that the bias of the UCL estimator and the BCL estimator vanishes as T grows. Additionally, the bias of the BCL estimator is substantially lower compared to UCL and converges more quickly to the true parameter value.

Figure 1 depicts the convergence of the three estimators for $N^* = 100$ (left) and $N^* = 1,000$ (right) and confirms that the BCL estimator converges quickly to the unbiased CL estimator. For $N^* = 100$, the bias-correction is able to reduce the absolute bias of the UCL estimator β (δ) from 53.1% to 12.9% (from 49.3% to 7.1%) for $T = 4$ and from 19.9% to 2.9% (from 19.7% to 3.3%) for $T = 8$.

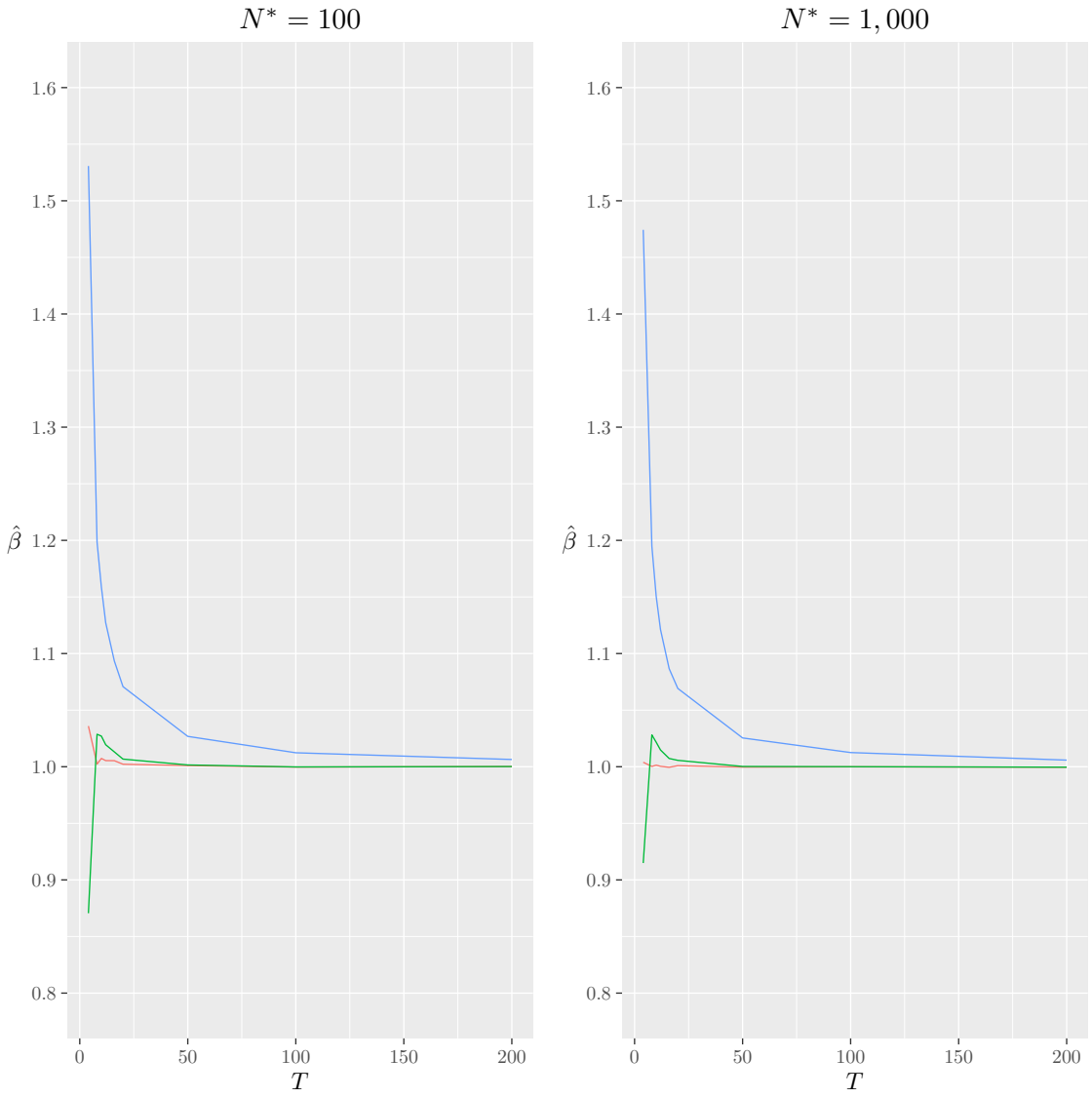
As Figure 2 and 5 shows, not only the mean estimates but the whole distribution of the structural parameters converges to the distribution of the CL estimator as T increases, irrespective of N^*. Already for $T = 16$ and almost for $T = 12$ the two-sided Kolmogorov-Smirnoff of test can't be rejected at any usual significance level for the structural parameter β when $N^* = 100$ (table 2). For $N^* = 1,000$ this convergence process lasts longer, maybe because standard errors shrink faster than the bias.

Table 1: Mean estimates of β and δ

	β			δ		
	UCL	BCL	CL	UCL	BCL	CL
$T = 4$	1.5307	0.8706	1.0359	1.4928	0.9289	1.0193
$T = 8$	1.1985	1.0287	1.0023	1.1971	1.0330	1.0112
$T = 10$	1.1588	1.0271	1.0074	1.1493	1.0225	1.0085
$T = 12$	1.1272	1.0194	1.0054	1.1101	1.0067	0.9977
$T = 16$	1.0934	1.0131	1.0054	1.0842	1.0072	1.0029
$T = 20$	1.0708	1.0067	1.0022	1.0618	1.0009	0.9986
$T = 50$	1.0268	1.0016	1.0010	1.0253	1.0016	1.0014
$T = 100$	1.0123	0.9998	0.9997	1.0132	1.0015	1.0014
$T = 200$	1.0064	1.0002	1.0002	1.0041	0.9983	0.9983

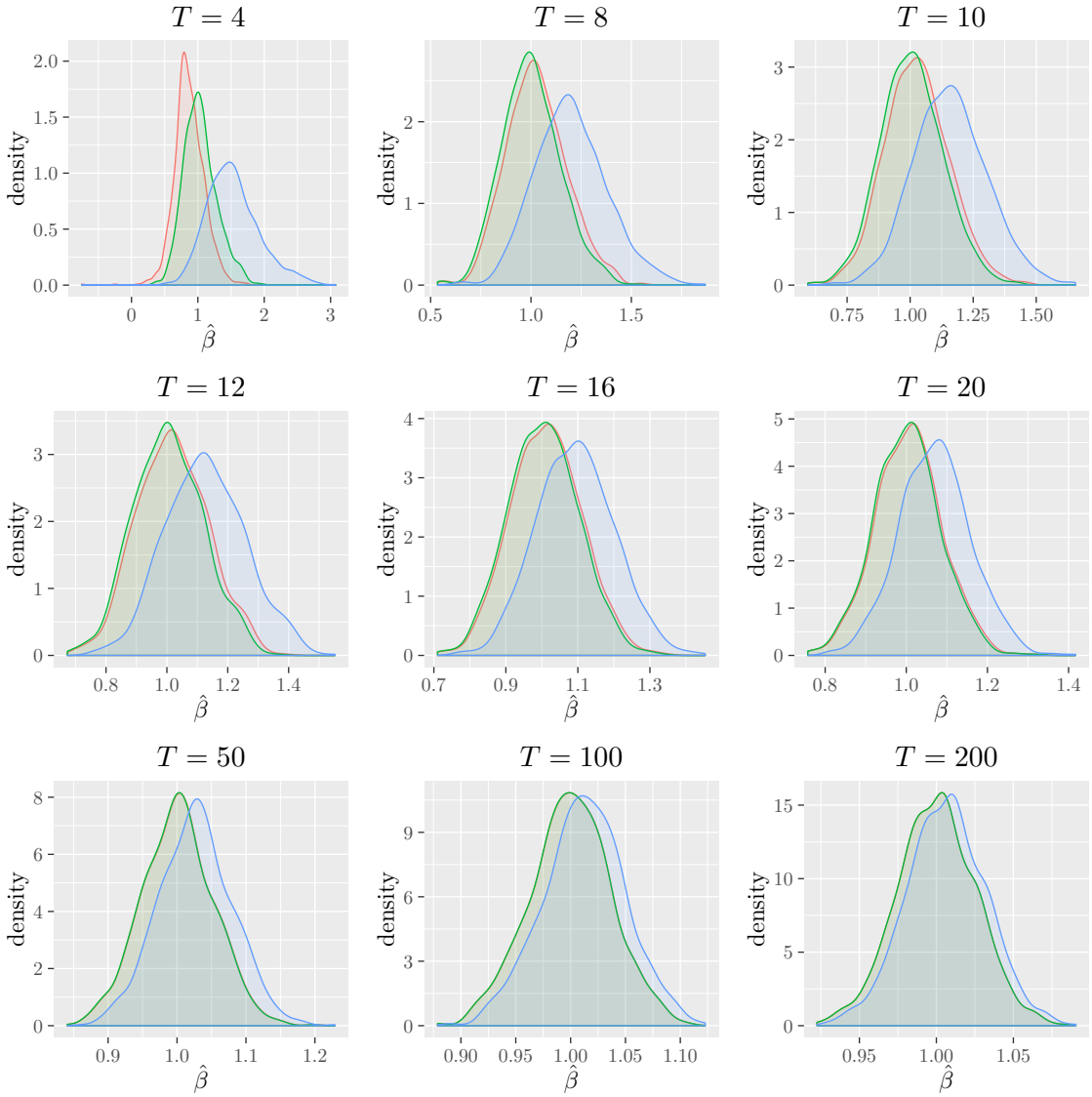
Note: 1,000 replications; $N^* = 100$; *UCL* denotes the unconditional logit estimator; *BCL* denotes the bias-corrected estimator of Hahn and Newey (2004) based on Bartlett equalities; *CL* denotes the conditional logit estimator.

Figure 1: Convergence



Note: 1,000 replications; blue refers to the unconditional logit estimator; green refers to the bias-corrected estimator of Hahn and Newey (2004) based on Bartlett equalities; red refers the conditional logit estimator.

Figure 2: Density



Note: 1,000 replications; $N^* = 100$; blue refers to the unconditional logit estimator; green refers to the bias-corrected estimator of Hahn and Newey (2004) based on Bartlett equalities; red refers the conditional logit estimator.

Table 2: Kolmogorov-Smirnov test

		UCL vs. BCL	UCL vs. CL	BCL vs. CL
β	$T = 4$	0.0000	0.0000	0.0000
	$T = 8$	0.0000	0.0000	0.0015
	$T = 10$	0.0000	0.0000	0.0053
	$T = 12$	0.0000	0.0000	0.0484
	$T = 16$	0.0000	0.0000	0.4324
	$T = 20$	0.0000	0.0000	0.6476
	$T = 50$	0.0000	0.0000	1.0000
	$T = 100$	0.0000	0.0000	1.0000
	$T = 200$	0.0000	0.0000	1.0000
δ	$T = 4$	0.0000	0.0000	0.0001
	$T = 8$	0.0000	0.0000	0.1338
	$T = 10$	0.0000	0.0000	0.3410
	$T = 12$	0.0000	0.0000	0.6852
	$T = 16$	0.0000	0.0000	0.9937
	$T = 20$	0.0000	0.0000	1.0000
	$T = 50$	0.0000	0.0000	1.0000
	$T = 100$	0.0028	0.0033	1.0000
	$T = 200$	0.0171	0.0171	1.0000

Note: 1,000 replications; $N^* = 100$; reported p-values of two-sided Kolmogorov-Smirnov test; *UCL* denotes the unconditional logit estimator; *BCL* denotes the bias-corrected estimator of Hahn and Newey (2004) based on Bartlett equalities; *CL* denotes the conditional logit estimator.

Following Greene (2004) and Hahn and Newey (2004), we examine how the inconsistency of the fixed effects affects the estimated standard errors of the structural parameters. Therefore, we report the average of the $1,000$ estimated standard errors (SE) and the ratio between SE and the sample standard deviation (SD). The idea is that the empirical standard deviation SD should deliver a more reliable measure compared to the contaminated SE. In fact, it turns out, that the SE of the UCL estimator is considerably distorted for $T = 4$ to 16 for $N^* = 100$ and even up to $T = 20$ for $N^* = 1,000$. In contrast to that, the distortions in the estimated standard errors of BCL and CL are negligible for $T \geq 8$.

Similarly, we consider how the incidental parameter problem carries over to rejection frequencies. If we do not control for the distortion in the estimated standard errors we reject a significant bias of the structural parameters mostly too rarely. Table 4 and B.4 report how often the null-hypothesis of no significant bias in the structural parameters is rejected, presuming a nominal value of $p = 0.05$. The results emphasize two things. First, the UCL estimator of the bias in structural parameters, rejects the null too often. Secondly, the BCL estimator takes the rejection frequencies much closer to the nominal value.

According to the tables B.3 and B.4, the number of groups N^* has no effect on the mean estimates but on the standard-errors and thus on rejection frequencies.

Table 3: SE and SE/SD

		SE			SE/SD		
		UCL	BCL	CL	UCL	BCL	CL
β	$T = 4$	0.3044	0.2341	0.2437	0.7723	1.0272	0.9734
	$T = 8$	0.1604	0.1493	0.1445	0.8954	0.9958	0.9992
	$T = 10$	0.1368	0.1294	0.1260	0.9367	1.0133	1.0160
	$T = 12$	0.1207	0.1153	0.1128	0.9256	0.9886	0.9925
	$T = 16$	0.1003	0.0969	0.0955	0.9390	0.9874	0.9863
	$T = 20$	0.0876	0.0852	0.0843	0.9923	1.0343	1.0330
	$T = 50$	0.0523	0.0518	0.0516	0.9812	0.9984	0.9961
	$T = 100$	0.0363	0.0361	0.0360	0.9720	0.9804	0.9793
	$T = 200$	0.0254	0.0253	0.0253	0.9915	0.9953	0.9949
δ	$T = 4$	0.4725	0.4131	0.3877	0.8023	1.2004	0.9862
	$T = 8$	0.2598	0.2511	0.2379	0.9222	1.0376	1.0089
	$T = 10$	0.2224	0.2168	0.2078	0.9253	1.0166	0.9911
	$T = 12$	0.1976	0.1936	0.1870	0.9389	1.0154	0.9930
	$T = 16$	0.1657	0.1632	0.1591	0.9574	1.0158	0.9972
	$T = 20$	0.1455	0.1437	0.1409	1.0088	1.0582	1.0416
	$T = 50$	0.0880	0.0876	0.0869	0.9802	0.9990	0.9923
	$T = 100$	0.0614	0.0612	0.0610	0.9698	0.9788	0.9757
	$T = 200$	0.0431	0.0430	0.0430	0.9764	0.9807	0.9793

Note: 1,000 replications; $N^* = 100$; *SE* denotes the average standard error of the estimator; *SE/SD* denotes the ratio of the average standard error and the standard deviation of the estimator; *UCL* denotes the unconditional logit estimator; *BCL* denotes the bias-corrected estimator of Hahn and Newey (2004) based on Bartlett equalities; *CL* denotes the conditional logit estimator.

Table 4: Rejection frequencies

		SE			*SD*		
		UCL	BCL	CL	UCL	BCL	CL
β	$T = 4$	0.380	0.076	0.043	0.245	0.073	0.056
	$T = 8$	0.232	0.047	0.043	0.186	0.058	0.051
	$T = 10$	0.210	0.050	0.046	0.193	0.057	0.055
	$T = 12$	0.202	0.057	0.046	0.168	0.062	0.054
	$T = 16$	0.163	0.047	0.046	0.146	0.049	0.049
	$T = 20$	0.120	0.044	0.042	0.129	0.054	0.052
	$T = 50$	0.080	0.056	0.057	0.078	0.052	0.052
	$T = 100$	0.073	0.060	0.060	0.066	0.057	0.056
	$T = 200$	0.063	0.050	0.049	0.064	0.050	0.050
δ	$T = 4$	0.209	0.018	0.041	0.128	0.043	0.045
	$T = 8$	0.127	0.045	0.050	0.099	0.054	0.054
	$T = 10$	0.111	0.043	0.062	0.090	0.054	0.057
	$T = 12$	0.098	0.039	0.050	0.081	0.044	0.045
	$T = 16$	0.088	0.041	0.043	0.076	0.047	0.049
	$T = 20$	0.065	0.029	0.031	0.075	0.038	0.037
	$T = 50$	0.072	0.051	0.052	0.066	0.050	0.050
	$T = 100$	0.064	0.058	0.058	0.057	0.054	0.054
	$T = 200$	0.059	0.053	0.054	0.053	0.049	0.049

Note: 1,000 replications; $N^* = 100$; Rejection frequencies of two-sided t-test $H_0 : \mu = 1$ with the nominal value $p = 0.05$ based on the standard errors *SE* and based on the standard deviation *SD*; *UCL* denotes the unconditional logit estimator; *BCL* denotes the bias-corrected estimator of Hahn and Newey (2004) based on Bartlett equalities; *CL* denotes the conditional logit estimator.

Table 5 reports the average relative bias of estimated average partial effects to the truth for $N^* = 100$, table B.5 respectively for $N^* = 1,000$. The bias of the unconditional logit estimator passes over to the APEs. We computed the estimated average partial effects on the four different ways described in section 4.

Table 5 and B.5 confirm that using bias-corrected coefficients $\hat{\beta}$ and $\hat{\delta}$ and adjusted fixed effects $\tilde{\alpha}_i$ leads to a substantial improvement of the APEs. For example, when $N^* = 100$ and $T = 4$ the APE of $\hat{\beta}$ is on average 31.9% biased upwards without bias-correction and only 5.8% with bias-correction. However, when T increases, the bias in the APEs of the UCL estimator decreases. For $T = 20$, there is no advantage anymore to compute APEs with BCL instead of UCL.

We find, that using the CL1 approach, where we set all fixed effects to zero, leads to a substantial bias of the APEs, for example a 14.8% bias on average for β and $T = 4$. This bias also remains as we increase T. Thus, we advice against using this common approach. The second approach CL2, where we recovered the fixed

effects with the first-order condition of the unconditional logit estimator, suggests a substantial improvement of the bias in the APEs compared to CL1.

The APEs computed with bias-corrected coefficients lead to less biased results compared to the APEs computed without bias-correction until $T = 20$ is reached both for $N^* = 100$ as well for $N^* = 1,000$. However, we additionally notice, that if T is too small the APEs computed both with BCL and with CL2 can fail to produce approximately unbiased results. For instance, for $T = 4$ the relative bias of the APE obtained with BCL is roughly 5.8% for the discrete regressor.

To sum up, the bias-correction is able to reduce the bias in structural parameters and average partial effects as well to reduce the distortions in the standard errors if T is not too small. However, the BCL estimator becomes appealing in situations where T is medium sized and it becomes redundant if T is large. Further, we conclude, that although the CL estimator delivers approximately unbiased estimates of the structural parameters if T is small (Tables 1 and B.1) this advantage does not transfer to produce approximately unbiased APEs. When T is medium sized or large, the approach to compute average partial effects for the CL estimator based on the f.o.c. of the UCL estimator performs approximately equally compared to computing the APEs with the BCL estimates. However, as we will see in the next subsection, the CL estimator has a clear speed disadvantage compared to UCL and BCL.

Table 5: Average Partial Effects

	β				δ			
	UCL	BCL	CL1	CL2	UCL	BCL	CL1	CL2
$T = 4$	1.3185	0.9421	1.1480	1.0649	1.3206	0.9958	1.1604	1.0457
$T = 8$	1.0886	0.9981	1.1694	0.9820	1.0975	0.9994	1.2149	0.9861
$T = 10$	1.0611	0.9909	1.1841	0.9789	1.0593	0.9833	1.2231	0.9757
$T = 12$	1.0451	0.9873	1.1939	0.9787	1.0339	0.9714	1.2222	0.9668
$T = 16$	1.0246	0.9819	1.2013	0.9770	1.0194	0.9732	1.2403	0.9714
$T = 20$	1.0150	0.9809	1.2055	0.9781	1.0087	0.9723	1.2451	0.9715
$T = 50$	1.0012	0.9886	1.2064	0.9883	1.0010	0.9876	1.2557	0.9876
$T = 100$	0.9995	0.9938	1.2013	0.9937	1.0011	0.9950	1.2542	0.9950
$T = 200$	1.0006	0.9983	1.1968	0.9983	0.9986	0.9960	1.2462	0.9960

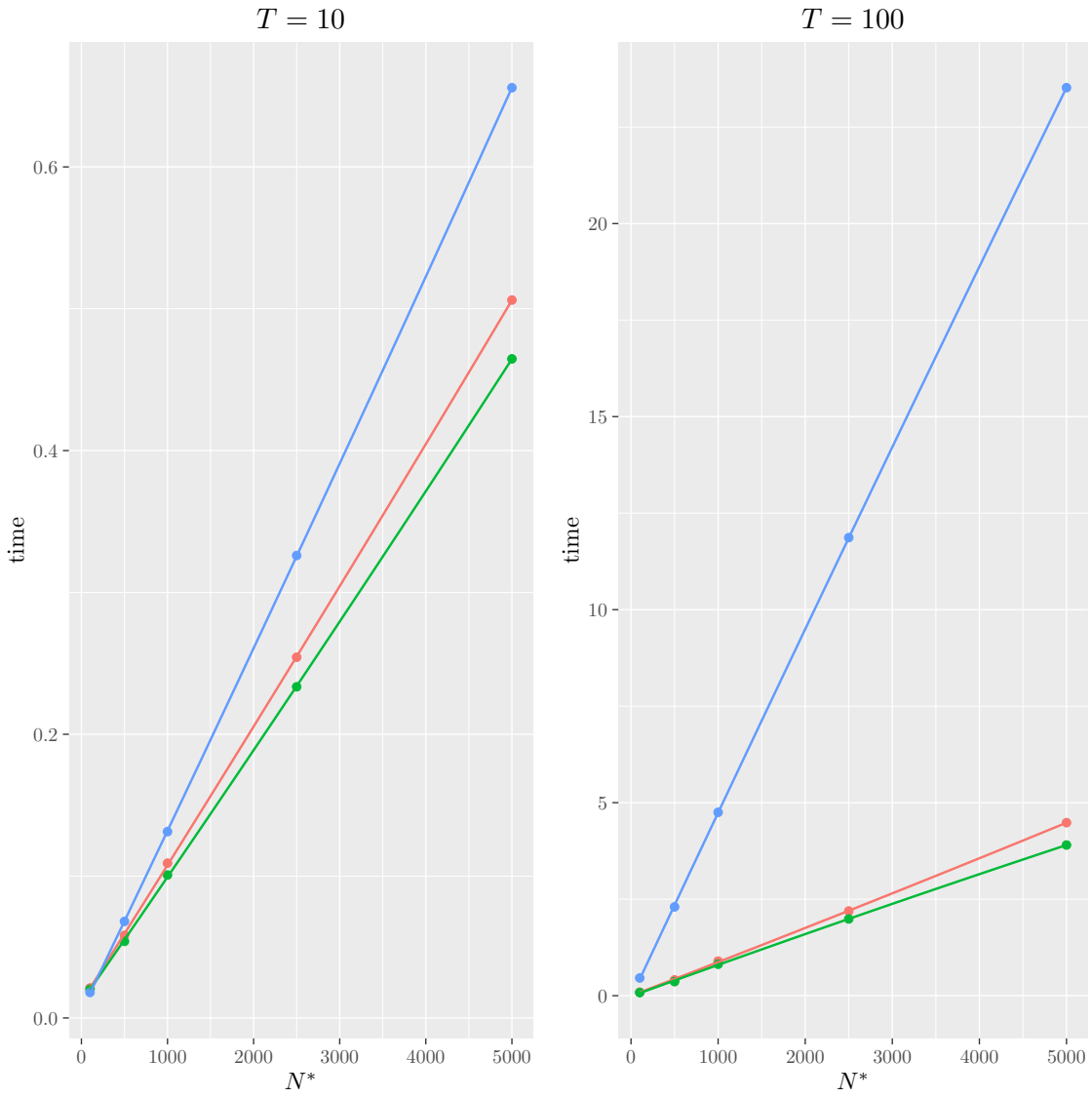
Note: 1,000 replications; $N^* = 1000$; *UCL* denotes the unconditional logit estimator; *BCL* denotes the bias-corrected estimator of Hahn and Newey (2004) based on Bartlett equalities; *CL1* denotes the conditional logit estimator with $\widehat{\alpha}_i = 0 \quad \forall i = 1, ..., N$; *CL2* denotes the conditional logit estimator with $\tilde{\alpha}_i \quad \forall i = 1, ..., N$ computed with the f.o.c. of *UCL*.

5.3. Computational costs

We test the theoretical computational complexity with real measured computation times. We compare the bias-corrected UCL algorithm described in subsection 3.2 with the recursive CL algorithm described in subsection 2.1. All calculations are done in R. We used the implementation of the CL estimator in the R-package `survival` and our own implementation of the UCL and BCL in the R-package `bife`.

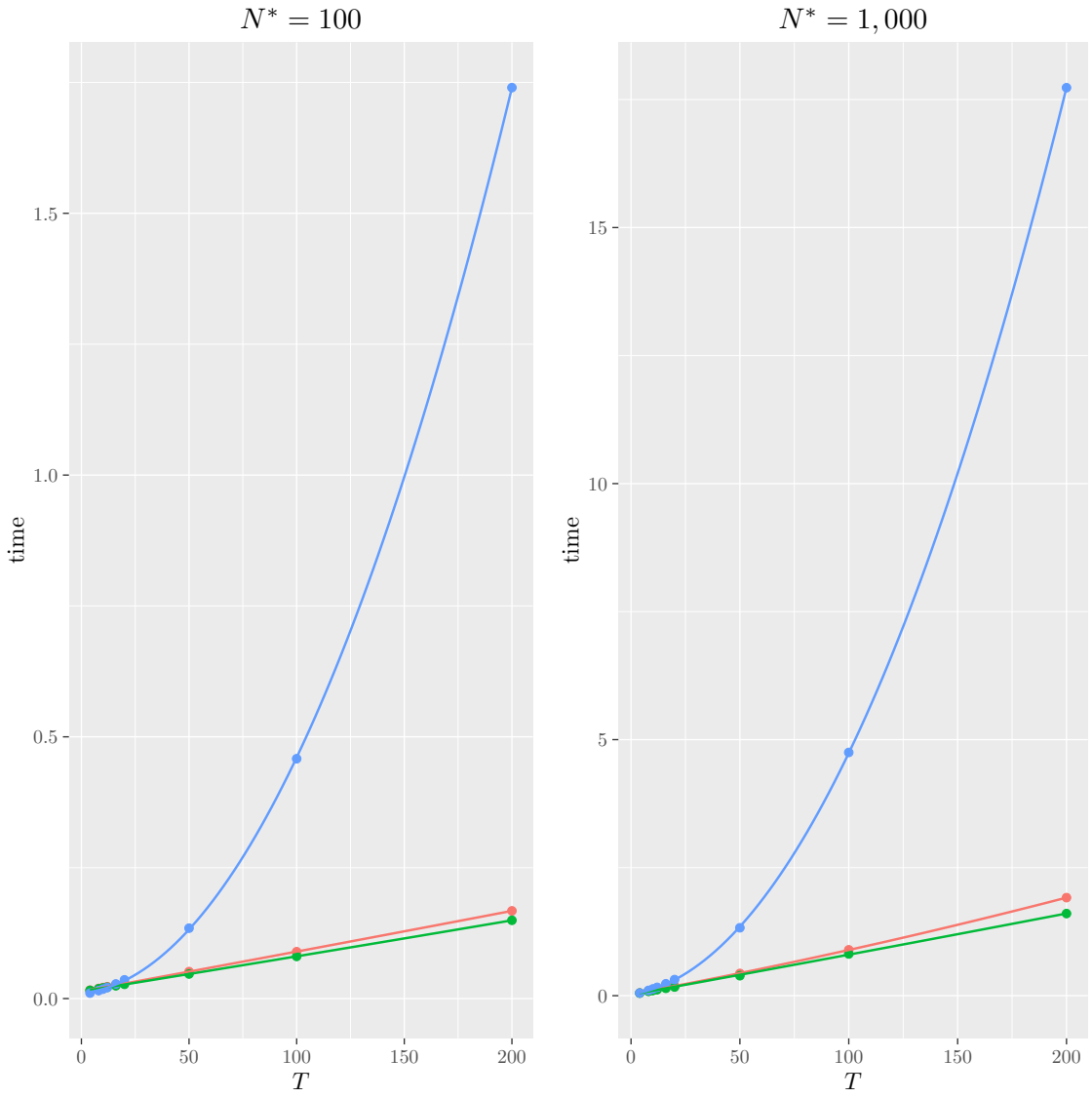
Figure 3 confirms that the computation time of both algorithms increases linearly in N^*. Besides, the figure already indicates that the computation time of CL increases drastically as T changes from $T = 10$ to $T = 100$. Figure 4 illustrates how computation time of CL and BCL evolve over T. It verifies the theoretical findings of a roughly quadratic computational cost of the CL algorithm and the linear one of the UCL algorithm. Since the bias-correction becomes redundant when T is sufficiently large, we further examine how much time the correction costs. Both figures depict that computing the unconditional estimator with bias-correction is not a big issue compared to the uncorrected unconditional estimator.

Figure 3: Computation Time



Note: Computation time in seconds. Blue line conditional logit algorithm approximated with linear regression, red line bias-corrected unconditional logit algorithm approximated with linear regression, green line unconditional logit algorithm approximated with linear regression.

Figure 4: Computation Time



Note: Computation time in seconds. Blue line conditional logit algorithm approximated with quadratic regression, red line bias-corrected unconditional logit algorithm approximated with linear regression, green line unconditional logit algorithm approximated with linear regression.

6. Conclusions

Fixed effects logit models for panel data allow for individual effects with an arbitrary distribution and a free correlation with the regressors. The two most widely used estimators suffer from substantial drawbacks.

The conditional logit (CL) estimator is consistent in terms of model parameters under standard assumptions but does not offer estimates of individual effects. Standard estimates of partial effects assuming zero fixed effects can be severely biased. Its computational costs increases quadratically with the number of observations per fixed effect T and can quickly become prohibitive for long panel data sets or models in which the structure is used for example to include fixed effects for states, counties, or other sub-groups of the sample.

The most straightforward approach to implement the unconditional logit (UCL) estimator is to add a set of dummy variables to a standard logit model to capture the fixed effects. This estimator suffers from the incidental parameters bias, but bias corrections have been suggested in the literature. Numerical maximization of the likelihood function quickly becomes infeasible with a large number of fixed effects N. The literature proposes strategies to use the sparse structure of the Hessian to reduce the computational burden of the UCL estimator.

This paper proposes a similar strategy for a computationally efficient way to obtain the UCL estimator. It uses an iterative pseudo-demeaning approach to optimize the likelihood with respect to the model parameters and jointly estimates all individual fixed effects. Combined with a correction for the incidental parameter bias, this BCL estimator provides an attractive alternative for many relevant applications.

We compare the alternative estimators and their computational burden and present a series of Monte-Carlo simulations. We find that the uncorrected UCL estimator indeed suffers from substantial biases if T is small. This incidental parameter problem carries over to values like average partial effects, standard errors and significance tests.

In terms of parameter estimation, the bias corrected BCL estimator has very similar properties as the CL estimator for longer panels and is clearly dominated only for very short panels ($T = 4$). When it comes to average partial effects (APE), the most common strategy for CL estimation to set the fixed effects to zero introduces severe biases and the BCL with estimated fixed effects performs much better. We also discuss how to obtain estimates for the fixed effects after CL estimation based on our BCL alogorithm. The APEs obtained in this way perform similarly as our BCL estimates.

In summary, our bias-corrected unconditional logit estimator is especially attractive for empirical applications in which the time dimension is reasonably large

($T > 4$), the computational burden is an issue because of large samples, and partial effects are of interest. To allow the readers to use the algorithm in a straightforward and convenient way, we provide an implementation in an R package called `bife`.

Appendix

A. Methodological part

A.1. Partitioned Inverse

Greene (2004), among others, proposed an algorithm which avoids the inversion of the large Hessian in (2.11). He uses the partitioned inverse formula for the inversion of the Hessian. Additionally, he exploits the sparse structure of the Hessian, i.e. that $\mathbf{H}_{\alpha\alpha}$ in (2.10) is a diagonal matrix.

The partitioned inverse formula of a general matrix with sub-matrices $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$ and $\mathbf{A}_{22}$ is given by Greene (2012):

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}_{11}^{-1}(\mathbf{I} + \mathbf{A}_{12}\mathbf{F}_2\mathbf{A}_{21}\mathbf{A}_{11}^{-1}) & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{F}_2 \\ -\mathbf{F}_2\mathbf{A}_{12}\mathbf{A}_{11}^{-1} & \mathbf{F}_2 \end{pmatrix} \tag{A.1}$$

with

$$\mathbf{F}_2 = \left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}$$

Greene partitions the inverse Hessian:

$$\mathbf{H}^{-1} = \begin{pmatrix} \mathbf{H}_{\beta\beta} & \mathbf{H}_{\beta\alpha} \\ \mathbf{H}_{\alpha\beta} & \mathbf{H}_{\alpha\alpha} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{H}^{\beta\beta} & \mathbf{H}^{\beta\alpha} \\ \mathbf{H}^{\alpha\beta} & \mathbf{H}^{\alpha\alpha} \end{pmatrix} \tag{A.2}$$

The updates for $\boldsymbol{\beta}$ can be rewritten

$$(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) = -\left[\mathbf{H}^{\beta\beta}\mathbf{g}_\beta + \mathbf{H}^{\beta\alpha}\mathbf{g}_\alpha\right] \tag{A.3}$$

As a next step the partitioned inverse formula is applied to $\mathbf{H}^{\beta\beta}$ and $\mathbf{H}^{\beta\alpha}$:

$$\mathbf{H}^{\beta\beta} = \left(\mathbf{H}_{\beta\beta} - \mathbf{H}_{\beta\alpha}\mathbf{H}_{\alpha\alpha}^{-1}\mathbf{H}_{\alpha\beta}\right)^{-1} \tag{A.4}$$

$$\mathbf{H}^{\beta\alpha} = \mathbf{H}^{\beta\beta}\mathbf{H}_{\beta\alpha}\mathbf{H}_{\alpha\alpha}^{-1} \tag{A.5}$$

Note, since $\mathbf{H}_{\alpha\alpha}$ is a diagonal matrix, the formula of $\mathbf{H}^{\beta\beta}$ can be further simplified

$$\mathbf{H}^{\beta\beta} = \left[\mathbf{H}_{\beta\beta} - \sum_{i=1}^{N} \frac{1}{h_{\alpha_i\alpha_i}}\mathbf{h}_{\beta\alpha_i}\mathbf{h}'_{\beta\alpha_i}\right]^{-1} \tag{A.6}$$

Finally, we can plug our results in the update formula (A.3) for $\boldsymbol{\beta}$

$$\begin{aligned}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) &= -\left[\mathbf{H}^{\beta\beta}\mathbf{g}_\beta + \mathbf{H}^{\beta\alpha}\mathbf{g}_\alpha\right] \\ &= -\left[\mathbf{H}^{\beta\beta}\mathbf{g}_\beta + \mathbf{H}^{\beta\beta}\mathbf{H}_{\beta\alpha}\mathbf{H}_{\alpha\alpha}^{-1}\mathbf{g}_\alpha\right] \\ &= -\mathbf{H}^{\beta\beta}\left[\mathbf{g}_\beta - \mathbf{H}_{\beta\alpha}\mathbf{H}_{\alpha\alpha}^{-1}\mathbf{g}_\alpha\right] \\ &= -\left[\mathbf{H}_{\beta\beta} - \sum_{i=1}^{N}\frac{1}{h_{\alpha_i\alpha_i}}\mathbf{h}_{\beta\alpha_i}\mathbf{h}'_{\beta\alpha_i}\right]^{-1}\left(\mathbf{g}_\beta - \sum_{i=1}^{N}\frac{g_{\alpha_i}}{h_{\alpha_i\alpha_i}}\mathbf{h}_{\beta\alpha_i}\right)\end{aligned} \tag{A.7}$$

Thus, the problem of the $(N+M \times N+M)$ matrix is removed. The formula (A.7) only requires the computation of a $(M \times M)$ matrix and the N-fold summation of $(M \times 1)$ vectors. The updates for $(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1})$ are obtained in the same manner. We start with the partitioned matrices

$$(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) = -\left[\mathbf{H}^{\alpha\alpha}\mathbf{g}_\alpha + \mathbf{H}^{\alpha\beta}\mathbf{g}_\beta\right] \tag{A.8}$$

and apply the partitioned inverse formula (A.1) to $\mathbf{H}^{\alpha\alpha}$ and $\mathbf{H}^{\alpha\beta}$:

$$(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) = -\mathbf{H}_\alpha^{-1}\left(\mathbf{g}_\alpha + \mathbf{H}_{\alpha\beta}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})\right) \tag{A.9}$$

By using the diagonal characteristic of $\mathbf{H}_{\alpha\alpha}$ we end with the update formula for the i-th fixed effects:

$$(\alpha_i^k - \alpha_i^{k-1}) = -\frac{1}{h_{\alpha_i\alpha_i}}\left(g_{\alpha_i} + \mathbf{h}'_{\beta\alpha_i}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})\right) \tag{A.10}$$

A.2. Bias correction

Hahn and Newey (2004) derived a bias formula obtained from an asymptotic expansion as T grows. This formula is used to estimate the bias and finally to correct the estimator. In order to describe the bias-correction we need to introduce some notation:

$$l_{it} = \ln f(y_{it}, \mathbf{x}_{it}|\boldsymbol{\beta}, \alpha_i), \quad \mathbf{g}_{\boldsymbol{\beta}it}(\boldsymbol{\beta}, \alpha_i) = \frac{\partial l_{it}}{\partial \boldsymbol{\beta}}, \qquad g_{\alpha it}(\boldsymbol{\beta}, \alpha_i) = \frac{\partial l_{it}}{\partial \alpha_i}$$

Other partial derivatives are denoted with extra subscripts in the following fashion

$$\mathbf{h}_{\beta\beta it}(\boldsymbol{\beta}, \alpha_i) = \frac{\partial l_{it}}{\partial \boldsymbol{\beta}\boldsymbol{\beta}'} = \frac{\partial \mathbf{g}_{\beta it}}{\partial \boldsymbol{\beta}}$$

$$\hat{g}_{\alpha it}(\boldsymbol{\beta}) = g_{\alpha it}(\boldsymbol{\beta}, \hat{\alpha}_i), \hat{h}_{\alpha\alpha it}(\boldsymbol{\beta}) = h_{\alpha\alpha it}(\boldsymbol{\beta}, \hat{\alpha}_i)$$

$$\hat{\mathbf{g}}_{\beta it}(\boldsymbol{\beta}) = \mathbf{g}_{\beta it}(\boldsymbol{\beta}, \hat{\alpha}_i), \hat{\mathbf{h}}_{\beta\alpha it}(\boldsymbol{\beta}) = \mathbf{h}_{\beta\alpha it}(\boldsymbol{\beta}, \hat{\alpha}_i)$$

The bias correction can be expressed by

$$\bar{\mathbf{B}}(\boldsymbol{\beta}) = \bar{\mathbf{H}}(\boldsymbol{\beta})^{-1}\bar{\mathbf{b}}(\boldsymbol{\beta}), \tag{A.11}$$

with

$$\bar{\mathbf{H}}(\boldsymbol{\beta}) = \frac{-1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\mathbf{U}}_{it}(\boldsymbol{\beta})\hat{\mathbf{U}}_{it}(\boldsymbol{\beta})'$$

$$\bar{\mathbf{b}}(\boldsymbol{\beta}) = \frac{-1}{2N}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\mathbf{U}}_{it}(\boldsymbol{\beta})\hat{V}_{2it}(\boldsymbol{\beta})/(\sum_{t=1}^{T}\hat{g}_{\alpha it}(\boldsymbol{\beta})^2)$$

$$\hat{\mathbf{U}}_{it}(\boldsymbol{\beta}) = \hat{\mathbf{g}}_{\beta it}(\boldsymbol{\beta}) - \hat{g}_{\alpha it}(\boldsymbol{\beta})\sum_{t=1}^{T}\hat{\mathbf{g}}_{\beta it}(\boldsymbol{\beta})\hat{g}_{\alpha it}(\boldsymbol{\beta})/\sum_{t=1}^{T}\hat{g}_{\alpha it}(\boldsymbol{\beta})^2$$

$$\hat{V}_{2it} = \hat{g}_{\alpha it}(\boldsymbol{\beta})^2 + \hat{h}_{\alpha\alpha it}(\boldsymbol{\beta})$$

The bias corrected coefficient of $\boldsymbol{\beta}$ is computed by the formula

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \frac{1}{T}\bar{\mathbf{B}}(\hat{\boldsymbol{\beta}}) \tag{A.12}$$

The bias-corrected estimator $\tilde{\boldsymbol{\beta}}$ decreases the incidental parameter bias enough if T grows faster than $N^{1/3}$. But this reduction is not enough for small T. The $\hat{\boldsymbol{\beta}}$ is heavily biased and affects $\tilde{\boldsymbol{\beta}}$. In order to decrease the bias for small T one could use $\tilde{\boldsymbol{\beta}}$ to get a new bias correction $\bar{\mathbf{B}}$:

$$\tilde{\boldsymbol{\beta}}^{(\mathbf{2})} = \hat{\boldsymbol{\beta}} - \frac{1}{T}\bar{\mathbf{B}}(\tilde{\boldsymbol{\beta}})$$

Hahn and Newey (2004) propose to repeat this procedure until convergence. However, we did not find any improvements using the iterative procedure in our simulations but rather an increase of the bias compared to the one-step correction. The same finding has been noted by Juodis (2015) for the panel probit model.

A.3. Fast computation of $\tilde{\boldsymbol{\alpha}}$

After the estimated structural parameters $\hat{\boldsymbol{\beta}}$ are bias-corrected, the estimated fixed effects $\hat{\boldsymbol{\alpha}}$ have to be adjusted. Therefore, the bias-corrected estimates $\tilde{\boldsymbol{\beta}}$ are used as a linear predictor $X\tilde{\boldsymbol{\beta}}$ in the IRWLS procedure (3.1) such that we get the new IRWLS procedure with the $k-th$ step

$$\begin{aligned}\breve{\boldsymbol{\alpha}}^{k} &= (\mathbf{D}'\mathbf{W}\mathbf{D})^{-\mathbf{1}}\mathbf{D}'\mathbf{D}(\mathbf{D}\breve{\boldsymbol{\alpha}}^{\mathbf{k-1}} - \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{W}^{-\mathbf{1}}(\mathbf{y}-\mathbf{p})) \\ &= (\mathbf{D}'\mathbf{W}\mathbf{D})^{-\mathbf{1}}\mathbf{D}'\mathbf{W}(\boldsymbol{\gamma} - \mathbf{X}\tilde{\boldsymbol{\beta}})\end{aligned} \tag{A.13}$$

with $\boldsymbol{\gamma} = \mathbf{D}\breve{\boldsymbol{\alpha}}^{\mathbf{k-1}} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})$.
Thus, $\breve{\boldsymbol{\alpha}}^k$ is a solution to the regression $\sqrt{\mathbf{W}}(\boldsymbol{\gamma} - \mathbf{X}\hat{\tilde{\boldsymbol{\beta}}}) = \sqrt{\mathbf{W}}\mathbf{D}\breve{\boldsymbol{\alpha}}^{\mathbf{k}}$.

The implementation of (A.13) with the dummy matrix and weighting matrix is not efficient if N is large. A closer look at (A.13) suggests to compute each $\breve{\boldsymbol{\alpha}}_i^k$ sequentially with the formula

$$\breve{\alpha}_i^k = \frac{\sum_{t=1}^T w_{it}(\gamma_{it} - \mu_{it})}{\sum_{t=1}^T w_{it}} \tag{A.14}$$

with $\gamma_{it} = \breve{\alpha}_i^{k-1} + \frac{y_{it} - p_{it}}{w_{it}}$ and $\mu_{it} = \mathbf{x_{it}}\tilde{\boldsymbol{\beta}}$.
The IRWLS procedure based on (A.14) is repeated until convergence. Lets denote the converged $\breve{\boldsymbol{\alpha}}^{\boldsymbol{k}}$ with $\tilde{\boldsymbol{\alpha}}$.

A.4. Technical note on pseudo-demeaning

The update formula of the fixed effects (3.3)

$$(\alpha_i^k - \alpha_i^{k-1}) = \frac{\sum_{t=1}^T \tilde{w}_{it}\tilde{y}_{it}}{\sum_{t=1}^T \tilde{w}_{it}^2} - \frac{\sum_{t=1}^T \tilde{w}_{it}\tilde{\mathbf{x}}_{it}}{\sum_{t=1}^T \tilde{w}_{it}^2}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})$$

can be derived by rewriting (3.2) in matrix notation

$$\tilde{\mathbf{y}} = \tilde{\mathbf{W}}(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) + \tilde{\mathbf{X}}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) \tag{A.15}$$

with $\tilde{\mathbf{W}} = \sqrt{\mathbf{W}}\mathbf{I}_{NT\times N}$.
Using this trick, we can directly apply the well-known Frisch-Waugh-Lovell Theorem, Frisch and Waugh (1933), Lovell (1963):

1. $\boldsymbol{\beta^k - \beta^{k-1}} = (\tilde{\mathbf{X}}'\mathbf{Q}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{Q}\tilde{\mathbf{y}}$ with $\mathbf{Q} = I - \tilde{\mathbf{W}}(\tilde{\mathbf{W}}'\tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}'$
2. $\boldsymbol{\alpha^k - \alpha^{k-1}} = (\tilde{\mathbf{W}}'\tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}'(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})$

Since $\mathbf{Q}$ is idempotent and symmetric, (1.) can be rewritten in

$$\boldsymbol{\beta^k - \beta^{k-1}} = (\tilde{\mathbf{X}}'\mathbf{Q}'\mathbf{Q}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{Q}'\mathbf{Q}\tilde{\mathbf{y}}$$

that is a regression of the weighted demeaned dependent variable $\bar{\mathbf{y}} := \mathbf{Q}\tilde{\mathbf{y}}$ on the weighted demeaned regressors $\bar{\mathbf{X}} := \mathbf{Q}\tilde{\mathbf{X}}$.
A more detailed derivation is given in the following. It is based on a similar approach as for the linear fixed effects model described in (Batalgi 2013) where the estimation

equation is directly manipulated with projections. In order to eliminate the fixed effects out of (A.15), we define the projection $\mathbf{P}$, such that $\mathbf{P} = \mathbf{P}^2$, $\mathbf{P} = \mathbf{P}'$ and $\mathbf{Q} = \mathbf{I} - \mathbf{P}$; i.e. $\mathbf{P}$ and $\mathbf{Q}$ are idempotent and symmetric. It follows, $\text{rank}(\mathbf{Q}) = \text{trace}(\mathbf{Q}) = N(T-1)$, $\text{rank}(\mathbf{P}) = \text{trace}(\mathbf{P}) = N$, $\mathbf{PQ} = 0$.
The projection $\mathbf{P}$ is an endomorphism

$$\begin{aligned}\mathbf{P}\tilde{\mathbf{y}} &= \mathbf{P}\tilde{\mathbf{W}}(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) + \mathbf{P}\tilde{\mathbf{x}}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) \\ &= \tilde{\mathbf{W}}(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) + \mathbf{P}\tilde{\mathbf{x}}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})\end{aligned}$$

with $\mathbf{P} = \tilde{\mathbf{W}}(\tilde{\mathbf{W}}'\tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}'$.
The matrix $\mathbf{Q}$ wipes out the weighted individual effects:

$$\begin{aligned}\mathbf{Q}\tilde{\mathbf{y}} &= \mathbf{Q}\tilde{\mathbf{W}}(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) + \mathbf{Q}\tilde{\mathbf{x}}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) \\ &= (\mathbf{I} - \mathbf{P})\tilde{\mathbf{W}}(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) + \mathbf{Q}\tilde{\mathbf{x}}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) \\ &= (\mathbf{I}\tilde{\mathbf{W}}(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) - \mathbf{P}\tilde{\mathbf{W}}(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1})) + \mathbf{Q}\tilde{\mathbf{x}}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) \\ &= (\tilde{\mathbf{W}}(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) - \tilde{\mathbf{W}}(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1})) + \mathbf{Q}\tilde{\mathbf{x}}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) \\ &= \mathbf{Q}\tilde{\mathbf{x}}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})\end{aligned}$$

since $\mathbf{P}\tilde{\mathbf{W}} = \tilde{\mathbf{W}}$ and $\mathbf{P} + \mathbf{Q} = \mathbf{I}$.

Finally, we can solve for $(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1})$ by using the projection $\tilde{\mathbf{P}}$, such that $\tilde{\mathbf{P}}\tilde{\mathbf{W}} = \mathbf{I}_{N\times N}$.

$$\begin{aligned}\tilde{\mathbf{P}}\tilde{\mathbf{y}} &= \tilde{\mathbf{P}}\tilde{\mathbf{W}}(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) + \tilde{\mathbf{P}}\tilde{\mathbf{x}}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) \\ &= (\tilde{\mathbf{W}}'\tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}'\tilde{\mathbf{W}}(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) + \tilde{\mathbf{P}}\tilde{\mathbf{x}}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) \\ &= (\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) + \tilde{\mathbf{P}}\tilde{\mathbf{x}}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})\end{aligned}$$

with $\tilde{\mathbf{P}} = (\tilde{\mathbf{W}}'\tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}'$, i.e. $\mathbf{P} = \tilde{\mathbf{W}}\tilde{\mathbf{P}}$ and $\mathbf{Q} = \mathbf{I} - \tilde{\mathbf{W}}\tilde{\mathbf{P}}$.

The projection matrix $\tilde{\mathbf{P}}$ is a $(N \times NT)$ block-diagonal matrix $\tilde{\mathbf{P}} = (\mathbf{I}_N \otimes \bar{\tilde{\mathbf{W}}}_i)'$ with the N individual specific blocks $\bar{\tilde{\mathbf{W}}}_i$, where $\bar{\tilde{\mathbf{W}}}_\mathbf{i} := \tilde{\mathbf{W}}_\mathbf{i} / \sum_{t=1}^{T} \tilde{\mathbf{W}}_{it}^2$ is a vector of dimension T. Further, $\text{rank}(\tilde{\mathbf{P}}) = N$.

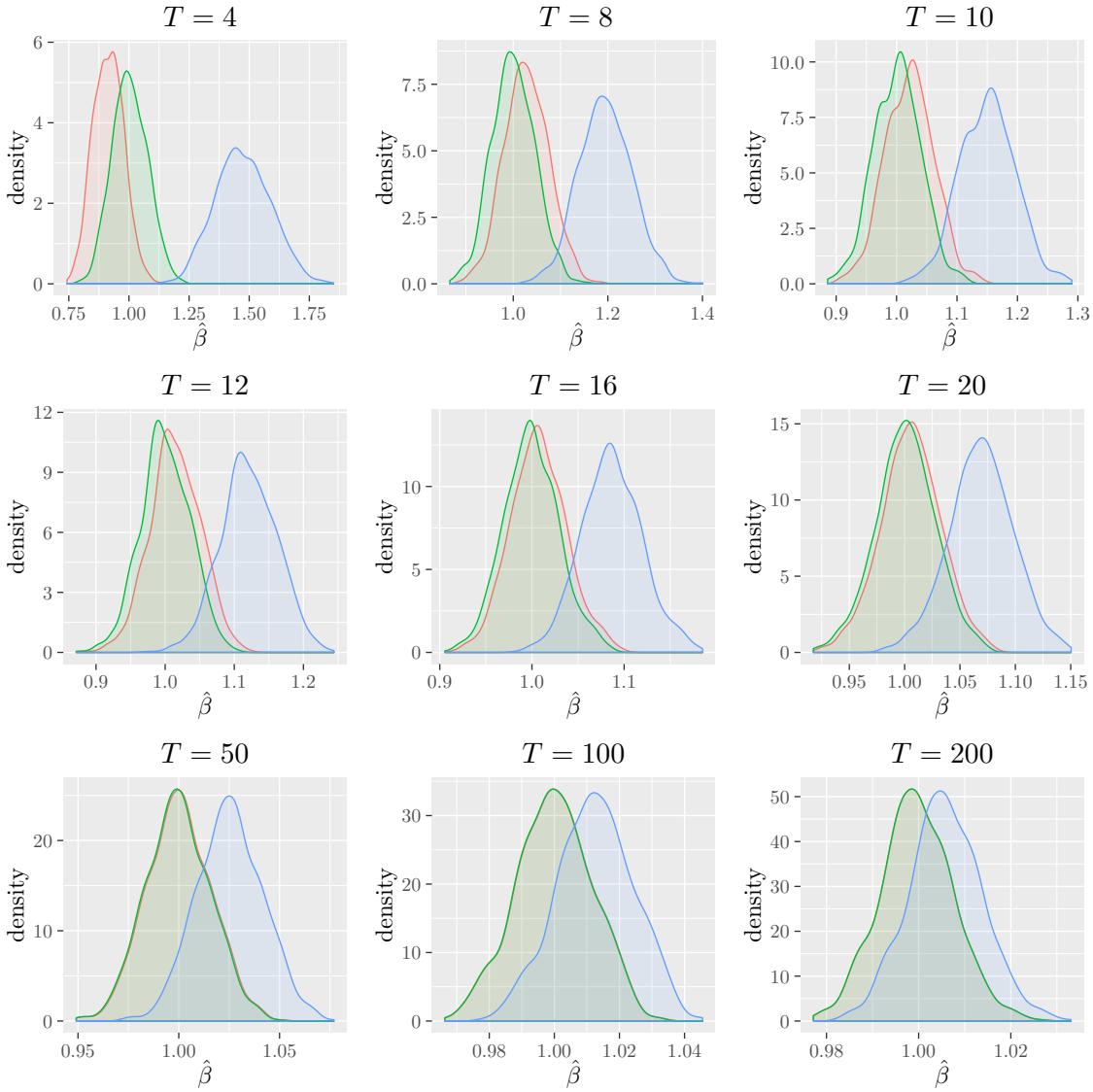
B. Tables

Table B.1: Mean estimates of β and δ

	β			δ		
	UCL	BCL	CL	UCL	BCL	CL
$T = 4$	1.4743	0.9149	1.0039	1.4562	0.9477	1.0006
$T = 8$	1.1948	1.0281	1.0003	1.1841	1.0227	1.0014
$T = 10$	1.1506	1.0216	1.0014	1.1384	1.0130	0.9994
$T = 12$	1.1212	1.0148	1.0004	1.1165	1.0127	1.0039
$T = 16$	1.0866	1.0073	0.9995	1.0812	1.0044	1.0004
$T = 20$	1.0692	1.0056	1.0010	1.0604	0.9996	0.9973
$T = 50$	1.0254	1.0003	0.9997	1.0249	1.0012	1.0010
$T = 100$	1.0125	1.0000	0.9999	1.0109	0.9992	0.9992
$T = 200$	1.0058	0.9996	0.9996	1.0057	0.9999	0.9999

Note: 1,000 replications; $N^* = 1,000$; *UCL* denotes the unconditional logit estimator; *BCL* denotes the bias-corrected estimator of Hahn and Newey (2004) based on Bartlett equalities; *CL* denotes the conditional logit estimator.

Figure 5: Density



Note: 1,000 replications; $N^* = 1,000$; blue refers to the unconditional logit estimator; green refers to the bias-corrected estimator of Hahn and Newey (2004) based on Bartlett equalities; red refers the conditional logit estimator.

Table B.2: Kolmogorov-Smirnov test

		UCL vs. BCL	UCL vs. CL	BCL vs. CL
β	$T = 4$	0.0000	0.0000	0.0000
	$T = 8$	0.0000	0.0000	0.0000
	$T = 10$	0.0000	0.0000	0.0000
	$T = 12$	0.0000	0.0000	0.0000
	$T = 16$	0.0000	0.0000	0.0000
	$T = 20$	0.0000	0.0000	0.0024
	$T = 50$	0.0000	0.0000	0.8280
	$T = 100$	0.0000	0.0000	1.0000
	$T = 200$	0.0000	0.0000	1.0000
δ	$T = 4$	0.0000	0.0000	0.0000
	$T = 8$	0.0000	0.0000	0.0000
	$T = 10$	0.0000	0.0000	0.0000
	$T = 12$	0.0000	0.0000	0.0097
	$T = 16$	0.0000	0.0000	0.2877
	$T = 20$	0.0000	0.0000	0.5361
	$T = 50$	0.0000	0.0000	1.0000
	$T = 100$	0.0000	0.0000	1.0000
	$T = 200$	0.0000	0.0000	1.0000

Note: 1,000 replications; $N^* = 1,000$; reported p-values of two-sided Kolmogorov-Smirnov test; *UCL* denotes the unconditional logit estimator; *BCL* denotes the bias-corrected estimator of Hahn and Newey (2004) based on Bartlett equalities; *CL* denotes the conditional logit estimator.

Table B.3: SE and SE/SD

		SE			SE/SD		
		UCL	BCL	CL	UCL	BCL	CL
β	$T = 4$	0.0927	0.0742	0.0745	0.8121	1.1760	1.0194
	$T = 8$	0.0504	0.0470	0.0454	0.9062	1.0006	1.0057
	$T = 10$	0.0429	0.0406	0.0395	0.9348	1.0084	1.0169
	$T = 12$	0.0379	0.0363	0.0355	0.9295	0.9938	0.9967
	$T = 16$	0.0316	0.0305	0.0301	0.9581	1.0075	1.0075
	$T = 20$	0.0276	0.0269	0.0266	0.9532	0.9927	0.9907
	$T = 50$	0.0165	0.0164	0.0163	1.0088	1.0261	1.0236
	$T = 100$	0.0115	0.0114	0.0114	0.9786	0.9868	0.9857
	$T = 200$	0.0080	0.0080	0.0080	1.0068	1.0109	1.0106
δ	$T = 4$	0.1450	0.1301	0.1195	0.8476	1.2114	1.0350
	$T = 8$	0.0814	0.0789	0.0747	0.8859	0.9935	0.9670
	$T = 10$	0.0699	0.0682	0.0654	0.9487	1.0401	1.0157
	$T = 12$	0.0622	0.0610	0.0589	0.9699	1.0488	1.0256
	$T = 16$	0.0522	0.0515	0.0502	0.9848	1.0441	1.0252
	$T = 20$	0.0459	0.0453	0.0444	0.9446	0.9915	0.9753
	$T = 50$	0.0278	0.0277	0.0275	0.9857	1.0046	0.9979
	$T = 100$	0.0194	0.0193	0.0193	0.9895	0.9989	0.9957
	$T = 200$	0.0136	0.0136	0.0136	1.0058	1.0103	1.0089

Note: 1,000 replications; $N^* = 1,000$; *SE* denotes the average standard error of the estimator; *SE/SD* denotes the ratio of the average standard error and the standard deviation of the estimator; *UCL* denotes the unconditional logit estimator; *BCL* denotes the bias-corrected estimator of Hahn and Newey (2004) based on Bartlett equalities; *CL* denotes the conditional logit estimator.

Table B.4: Rejection frequencies

		SE			*SD*		
		UCL	BCL	CL	UCL	BCL	CL
β	$T = 4$	0.999	0.192	0.040	0.992	0.287	0.051
	$T = 8$	0.967	0.088	0.051	0.947	0.095	0.053
	$T = 10$	0.945	0.080	0.047	0.911	0.084	0.048
	$T = 12$	0.878	0.074	0.047	0.842	0.074	0.051
	$T = 16$	0.774	0.059	0.055	0.752	0.062	0.057
	$T = 20$	0.709	0.062	0.059	0.670	0.062	0.058
	$T = 50$	0.323	0.040	0.042	0.327	0.047	0.047
	$T = 100$	0.201	0.051	0.052	0.193	0.046	0.047
	$T = 200$	0.107	0.045	0.045	0.110	0.049	0.050
δ	$T = 4$	0.842	0.032	0.038	0.748	0.079	0.047
	$T = 8$	0.612	0.069	0.063	0.518	0.065	0.058
	$T = 10$	0.511	0.046	0.051	0.459	0.051	0.053
	$T = 12$	0.476	0.038	0.044	0.453	0.052	0.047
	$T = 16$	0.341	0.040	0.040	0.334	0.043	0.044
	$T = 20$	0.272	0.053	0.058	0.240	0.047	0.049
	$T = 50$	0.147	0.041	0.044	0.146	0.043	0.042
	$T = 100$	0.090	0.056	0.057	0.088	0.055	0.055
	$T = 200$	0.067	0.048	0.048	0.070	0.055	0.055

Note: 1,000 replications; $N^* = 1,000$; Rejection frequencies of two-sided t-test $H_0 : \mu = 1$ with the nominal value $p = 0.05$ based on the standard errors *SE* and based on the standard deviation *SD*; *UCL* denotes the unconditional logit estimator; *BCL* denotes the bias-corrected estimator of Hahn and Newey (2004) based on Bartlett equalities; *CL* denotes the conditional logit estimator.

Table B.5: Average Partial Effects

	β				δ			
	UCL	BCL	CL1	CL2	UCL	BCL	CL1	CL2
$T = 4$	1.3093	0.9872	1.1408	1.0534	1.3184	1.0068	1.1662	1.0388
$T = 8$	1.0919	1.0020	1.1727	0.9847	1.0873	0.9896	1.2081	0.9770
$T = 10$	1.0607	0.9914	1.1839	0.9788	1.0525	0.9766	1.2180	0.9692
$T = 12$	1.0412	0.9842	1.1893	0.9751	1.0395	0.9767	1.2329	0.9724
$T = 16$	1.0222	0.9799	1.1977	0.9749	1.0186	0.9724	1.2406	0.9707
$T = 20$	1.0147	0.9811	1.2043	0.9782	1.0072	0.9711	1.2432	0.9703
$T = 50$	1.0010	0.9886	1.2059	0.9883	1.0012	0.9880	1.2566	0.9881
$T = 100$	1.0002	0.9947	1.2012	0.9947	0.9989	0.9930	1.2507	0.9930
$T = 200$	0.9997	0.9976	1.1950	0.9975	0.9997	0.9973	1.2468	0.9973

Note: 1,000 replications; $N^* = 1,000$; *UCL* denotes the unconditional logit estimator; *BCL* denotes the bias-corrected estimator of Hahn and Newey (2004) based on Bartlett equalities; *CL1* denotes the conditional logit estimator with $\widehat{\alpha}_i = 0 \quad \forall i = 1, ..., N$; *CL2* denotes the conditional logit estimator with $\widehat{\alpha}_i \quad \forall i = 1, ..., N$ computed with the f.o.c. of *UCL*.

References

Andersen, Erling Bernhard (1970) 'Asymptotic properties of conditional maximum-likelihood estimators.' *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 283–301

Arellano, Manuel, and Jinyong Hahn (2006) 'A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects.' *Documentos de Trabajo (CEMFI)* (13), 1–49

Batalgi, Badi H. (2013) *Econometric Analysis Of Panel Data,* 5th ed. (Wiley: New York)

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

Coupé, Tom (2005) 'Bias in conditional and unconditional fixed effects logit estimation: A correction.' *Political Analysis* 13(3), 292–295

Dhaene, Geert, and Koen Jochmans (2015) 'Split-panel jackknife estimation of fixed-effect models.' *The Review of Economic Studies* 82(3), 991–1030

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

Frisch, Ragnar, and Frederick V Waugh (1933) 'Partial time regressions as compared with individual trends.' *Econometrica: Journal of the Econometric Society* pp. 387–401

Gail, Mitchell H., Jay H. Lubin, and Lawrence V. Rubinstein (1981) 'Likelihood calculations for matched case-control studies and survival studies with tied death times.' *Biometrika* 68(3), 703–707

Greene, William (2004) 'The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects.' *Econometrics Journal* 7, 98–119

—— (2012) *Econometric Analysis,* 7th ed. (Englewood Cliffs: Prentice Hall)

Hahn, J., and W. Newey (2004) 'Jackknife and analytical bias reduction for nonlinear panel models.' *Econometrica* 72(4), 1295–1319

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

Hospido, Laura (2012) 'Estimating nonlinear models with multiple fixed effects: A computational note.' *Oxford Bulletin of Economics and Statistics* 74(5), 760–775

Juodis, Arturas (2015) 'Iterative bias correction procedures revisited: A small scale monte carlo study.' Discussion Paper 2015/02, Amsterdam School of Economics

Lovell, Michael C (1963) 'Seasonal adjustment of economic time series and multiple regression analysis.' *Journal of the American Statistical Association* 58(304), 993–1010

Neyman, Jerzy, and Elizabeth L Scott (1948) 'Consistent estimates based on partially consistent observations.' *Econometrica: Journal of the Econometric Society* pp. 1–32

Prentice, Ross L, and Lynn A Gloeckler (1978) 'Regression analysis of grouped survival data with application to breast cancer data.' *Biometrics* pp. 57–67

Reid, Stephen, and Rob Tibshirani (2014) 'Regularization paths for conditional logistic regression: The clogitl1 package.' *Journal of statistical software*