Estimation of Fixed Effects Logit Models with Large Panel Data

FLORIAN HEISS[*] DANIEL MCFADDEN[†] AMREI STAMMANN[‡]

December 18, 2019

We discuss different estimators for logistic regression models with individual fixed effects and tackle their short-comings. A simple unconditional maximum likelihood estimator (UCL) that uses a dummy variable for each of the N cross-sectional units suffers from prohibitive computational costs if N is large. Further, it can be severely biased due to the incidental parameter problem (IPP). Using the Frisch-Waugh-Lovell theorem we derive an intuitive and computationally efficient algorithm which can be easily combined with existing bias corrections to address the IPP. The popular conditional logit estimator (CL) is fixed T consistent, but it exhibits an enormous computational burden if T is large. We therefore propose a new conditional logit estimator that alleviates the computational burden of CL at costs of efficiency. Since conditional logit estimators do not allow to estimate average partial effects, we propose a novel hybrid approach. Extensive Monte-Carlo simulations confirm that the bias-corrected UCL estimator (BCL) and CL have similar statistical properties. However, combined with the pseudo-demeaning algorithm, UCL and BCL have a much lower computational burden, especially with large T. The algorithm is implemented in the R package `bife`.

JEL Classification: C13, C23, C55, C87
Keywords: Average Partial Effects, Bias-Reduction, Logit Model, High-Dimensional Fixed Effects, Panel Data

[*] Heinrich Heine University Duesseldorf
[†] University of California, Berkeley and USC, Los Angeles
[‡] Heinrich Heine University Duesseldorf

1 Introduction

The recent availability of long microeconomic panels like the Panel Study for Income Dynamics constitutes new computational challenges for the estimation of common econometric models. One of these is the logit model with individual fixed effects which is referred to hereinafter as the fixed effects logit model. The fixed effects logit model is a popular specification for analyzing panel data of binary variables, since it allows for unobserved individual heterogeneity like the variation in tastes with an arbitrary distribution.

There are two established approaches for the estimation of fixed effects logit models. On the one hand, it is possible to carry out a standard maximum likelihood estimation in which the regressor set is extended by one dummy variable per cross-sectional unit. We call this estimator the unconditional logit estimator and abbreviate it with UCL. The other estimator, a conditional logit estimator (CL), concentrates the individual heterogeneity out of the likelihood function by conditioning on a sufficient statistic. Both estimators suffer from substantial drawbacks which this article is intended to address.

UCL 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 most nonlinear fixed effects models suffer from the incidental parameters problem (IPP), which is reflected in a bias, first noted by Neyman and Scott (1948). This incidental parameters bias can be especially severe in models with a small number of observations T per individual. The reason is that only few observations contribute to the estimation of the fixed effects leading to noisy estimates. Due to the nonlinear nature of the logit model, the estimation noise of the fixed effects also contaminates the estimates of the structural parameters. Thus, UCL is inconsistent under fixed T asymptotics (see Arellano and Hahn 2007; Fernández-Val and Weidner 2018). Even increasing T does not necessary solve the incidental parameters bias because fixed effects estimators are asymptotically biased even if T grows at the same rate as N (see Hahn and Newey 2004).

CL has been derived by Rasch (1960) and Andersen (1970) and later generalized by Chamberlain (1980) as a solution to IPP. They show that CL is a fixed T consistent estimator for structural parameters. However it is not clear how interpretable values, such as average partial effects (APEs), can be estimated since CL does not deliver estimates of the fixed effects (see Hahn and Newey 2004; Arellano and Hahn 2007; Fernández-Val and Weidner 2018).[1] Another drawback is that CL is computationally very costly if T is large. Even if we use a more efficient recursion method proposed by Gail, Lubin, and Rubinstein (1981), the computational burden increases roughly quadratic in T which makes CL infeasible for panels with large time horizons.

The contributions of our article are manifold. We address the aforementioned problems of the different estimators. With respect to the UCL estimator this means that we first derive an intuitive and efficient algorithm based on the Frisch-Waugh-Lovell (FWL) theorem (Frisch and Waugh 1933; Lovell 1963).[2] We call this approach *pseudo-demeaning* because of its similarity to the within transformation in linear fixed effects models. The remaining challenge of UCL, which is the incidental parameter bias, is reduced by combining the pseudo-demeaning algorithm with an analytical bias correction proposed by Fernández-Val (2009).[3] To tackle the computational burden of CL for large T, we introduce a new estimator which we refer to as CLsub. This estimator is based on an estimator that has been designed by McFadden (1978) to overcome the curse of dimensionality problem in multinomial logit models. CLsub is essentially an adaption of this estimator to a binary dependent variable. The idea is to reduce the computational costs by using only a subset of all permutations of the observed choice sequence in the estimation routine. Furthermore, we propose a novel approach that uses estimates of the fixed effects obtained by an offset algorithm to compute APEs for conditional logit estimators. We also present an appropriate formula, based on a concentrated delta method, which can be used for conditional and unconditional logit estimators to calculate standard errors for APEs without having to use computationally

[1] Often partial effects are also called marginal or ceteris paribus effects.

[2] An alternative approach exploits the specific sparse structure of the Hessian (see Hall 1978; Prentice and Gloeckler 1978; Chamberlain 1980; Greene 2004).

[3] A comprehensive overview on different bias correction approaches is given by Arellano and Hahn (2007) and Fernández-Val and Weidner (2018). For our purposes only ex-post bias corrections are of interest, since they can be conveniently combined with our pseudo-demeaning approach. They can be analytical (e.g Hahn and Newey 2004; Fernández-Val 2009) or based on re-sampling methods (e.g Hahn and Newey 2004; Dhaene and Jochmans 2015).

demanding bootstrap methods. In extensive simulation experiments we finally investigate the finite sample properties of the different estimators with respect to structural parameters and APEs. In addition, we empirically verify the theoretical computational complexities that we have derived in advance. Finally, we use an empirical example from labor economics, to demonstrate a relevant field for the application of our pseudo-demeaning algorithm. In this example, T is even so large that conditional logit estimators are not feasible, whereas our pseudo-demeaning approach can easily estimate the model. In order to make our (bias-corrected) pseudo-demeaning algorithm accessible for applied work, we offer it in the *R*-package *bife*.[4]

Our simulation experiments confirm the findings of Greene (2004), who reports large distortions in the UCL estimator of the structural parameters for small T. Furthermore, the bias correction substantially reduces this distortion and, for sufficiently large values of T, it has similar desirable properties like the fixed T consistent CL estimator. Similar results regarding BCL are presented by Fernández-Val (2009), who focuses on probit models. Besides, our results, that UCL shows only little distortions in the APEs even for small values of T and that the bias correction works similarly well, are also in line with Fernández-Val (2009). Furthermore, we find that the CLsub estimator provides consistent estimates for the structural parameters only if the subset is large enough relative to the entire permutation set. However, compared to CL it is less efficient. The simulation results also demonstrate that estimates of the APEs obtained by conditional logit estimators can suffer from severe biases if the contributions of the fixed effects are ignored in their calculation. Our new strategy to estimate APEs for conditional logit estimators based on an offset algorithm is a substantial improvement over the aforementioned approach. However, even CL, which has the best properties among all conditional logit estimators, is slightly outperformed by UCL and BCL in estimating APEs. Moreover, the simulation experiments verify that the computational burden of UCL and BCL, both combined with the pseudo-demeaning approach, increase linearly with T, whereas the burden of recursive CL increases quadratically, which makes

[4] The package can estimate structural and incidental parameters, as well as average partial effects of fixed effects logit and probit models and provides the analytical bias correction of Fernández-Val (2009). The package also offers the corresponding standard errors. https://cran.r-project.org/web/packages/bife/.

a dramatic difference for large T. Besides, we demonstrate that CLsub can further reduce the computation time if the used subset of permutations is small. Considering the trade-off between statistical properties and computation time, we conclude that there is no advantage of using CLsub over CL. Especially if T is large, the speed advantage of a small subset comes at costs of high biases. Overall, UCL and BCL offer a clear computation time advantage over CL, which is particularly evident for samples with large T. Apart from that, (bias-corrected) UCL is also a promising candidate for practical applications in terms of statistical properties, especially when APEs are of main interest.

The paper is organized as follows. Section 2 presents a short recap of the fixed effects logit models along with its basic estimators. In section 3 we derive the pseudo-demeaning approach and present the entire Newton-Raphson pseudo-demeaning optimization routine. Section 4 introduces CLsub. It follows the description of different offset algorithms and the concentrated delta method in section 5. In section 6, the design and results of a series of Monte Carlo simulations are presented before section 7 demonstrates an empirical example. Finally, section 8 concludes.

2 The Fixed Effects Logit Model and Basic Estimators

2.1 The Fixed Effects Logit Model

For the sake of notational simplicity, we assume a balanced panel of $i = 1, \ldots, n$ individuals observed for $t = 1, \ldots, T$ time periods.[5] Suppose we observe a binary dependent variable y_{it}, such that $y_{it} = 1$ if an event occurs and $y_{it} = 0$ if it does not occur. Let $N = \sum_{i=1}^{n} \mathbf{1}[0 < \sum_{t=1}^{T} y_{it} < T]$ be the number of cross-sectional units for which y_{it} varies over time, where $\mathbf{1}[\cdot]$ is an indicator function. The $n - N$ individuals without varying y_{it} do not contribute to the identification and can be dropped from the analysis without affecting the estimator of the structural parameters. We refer to these observations as perfectly classified.

[5] The same type of model applies to unbalanced data and so-called *pseudo panels* where we include fixed effects for n groups each of size T_i.

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

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

with the conditional success probability

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

where $\eta_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}$ is the linear predictor and $\boldsymbol{\beta}$ is a vector of structural parameters corresponding to M regressors $\mathbf{x}_{it}$. The parameter α_i is called a fixed effect which is allowed to be arbitrarily correlated with the regressors. Throughout the paper, we assume that $N \gg M$ and that the regressor matrix $\mathbf{X}$ has full column rank.

The most common approach to estimate the fixed effects logit model is maximum likelihood. In the following subsections we depict the advantages and drawbacks of the two most popular estimators which are the conditional logit estimator (CL) and the unconditional logit estimator (UCL). Further, we address the problem of estimating APEs.

2.2 Basic Estimation Approaches for Structural Parameters

2.2.1 Unconditional Logit Estimator via Dummy Variables

The simplest estimator for the fixed effects logit model is a full maximum likelihood estimator which jointly estimates $\boldsymbol{\beta}$ and $\boldsymbol{\alpha} = [\alpha_1,\ldots,\alpha_N]'$. It can be conveniently estimated with standard statistical software by including a dummy variable for each individual as additional covariates.

This estimator is inconsistent as N increases and T is held constant, which is known as the incidental parameters problem (IPP) noted by Neyman and Scott (1948). However, several bias corrections have been proposed in the literature to reduce this bias (e.g. Hahn and Newey 2004; Carro 2007; Fernández-Val 2009; Dhaene and Jochmans 2015).

Estimates of UCL can be obtained by maximizing the log-likehood function

$$L(\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_{i=1}^{N}\sum_{t=1}^{T} \log\left(f(y_{it}|\mathbf{x}_{it},\boldsymbol{\beta},\alpha_i)\right) = \sum_{i=1}^{N}\sum_{t=1}^{T} l_{it}\,. \tag{2}$$

The standard routine to optimize (2) is the Newton-Raphson algorithm, which has the following parameter update in iteration $(k-1)$

$$(\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}) = -\mathbf{H}^{-1}\mathbf{g}\,, \tag{3}$$

where $\mathbf{H}$ denotes the $(M+N)\times(M+N)$ Hessian, $\mathbf{g}$ denotes the $(M+N)\times 1$ gradient, and $\boldsymbol{\theta} = [\boldsymbol{\beta}', \boldsymbol{\alpha}']'$ is the parameter vector. To be more specific, (3) can be reformulated as follows:

$$(\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}) = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y}-\mathbf{p})\,, \tag{4}$$

where $\mathbf{Z} = [\mathbf{X},\mathbf{D}]$ denotes the entire regressor matrix, which includes the dummy variable matrix $\mathbf{D}$ and the remaining regressors $\mathbf{X}$, and $\mathbf{W}$ is a positive definite diagonal weighting matrix with $\text{diag}(\mathbf{W}) = \mathbf{p}(1-\mathbf{p})$. After convergence, the standard errors of $\hat{\boldsymbol{\theta}}$ can be obtained as the square-root of the diagonal of the inverse Hessian, $\widehat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = -\mathbf{H}^{-1}$. Details on the implementation are presented in appendix A.

Adding N dummy variables as covariates creates a substantial computational burden if N is large. Especially the computation and inversion of the Hessian needed for a Newton-Raphson optimization is demanding. As shown in appendix B, the computational costs of estimating UCL based on dummy variables is linear in T but cubic in N. This can quickly become prohibitive for large panel data sets. We discuss an algorithm that dramatically reduces the computational burden of this estimator in section 3.

2.2.2 The Conditional Logit Estimator

CL uses the individual number of successes $t_{1i} = \sum_t y_{it}$ as sufficient statistics to concentrate the incidental parameters out of the log-likelihood function. Thus, $\boldsymbol{\beta}$ obtained by CL is consistent for $N \to \infty$ and fixed T (see Chamberlain 1980).

The corresponding log-likelihood function is given by

$$L_c(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log\left(f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}, t_{1i})\right) , \tag{5}$$

where

$$f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}, t_{1i}) = \frac{\exp\left(\sum_{t=1}^{T} y_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)}{\sum_{b_i \in \mathscr{B}(t_{1i})} \exp\left(\sum_{t=1}^{T} b_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)} \tag{6}$$

is the joint probability of $\mathbf{y}_i$ conditional on t_{1i}, and $\mathscr{B}(t_{1i})$ is the set of all $c_i = \binom{T}{t_{1i}}$ permutations of $\mathbf{y}_i$. Just like (2), (5) can be maximized using a standard Newton-Raphson algorithm. In contrast to (3) the Hessian corresponding to CL is only of the dimension $M \times M$. Nevertheless, CL can become computationally intensive due to two other problems, stemming from the individual likelihood contributions given by (6). First, a large time-series dimension T implies substantial or even prohibitive computational costs, since $\mathscr{B}(t_{1i})$ quickly becomes huge. For example, $c_i = \binom{50}{20}$ is larger than 10^{13}. In total, a brute force implementation of CL requires $\approx O(\sum_{i=1}^{N} t_{1i}\binom{T}{t_{1i}})$ time, which is exponentially increasing in T (see appendix B). Second, the higher the number of permutations, the more likely the denominator in (6) becomes numerically hard to deal with.[6]

It is nowadays standard to mitigate the computational burden of CL by using a recursive algorithm proposed by Gail, Lubin, and Rubinstein (1981). As detailed in appendix B, the computational costs of this recursive implementation are $\approx O(\sum_{i=1}^{N} t_{1i}(T - t_{1i}))$. In the worst case, which is $t_{1i} = T/2$, they are quadratic in T.[7] We will discuss another strategy to reduce the computational burden of CL by considering only a random subset of $\mathscr{B}(t_{1i})$ in section 4.

2.3 Basic Estimation Approaches for Average Partial Effects

Since the structural parameters $\boldsymbol{\beta}$ do not have a direct interpretation, average partial effects (APEs) are often of major interest for applied work. When calculating APEs, a case distinction

[6] For instance, the largest value a computer can handle is $1.797693 \cdot 10^{308}$ in double precision.

[7] The recursion can also be accelerated by using a not completely recursive implementation which reuses results and thus decreases the number of arithmetic operations by a factor. This however comes along with a higher memory requirement compared to the fully recursive program (see Gaure 2012).

is made for discrete and continuous regressors. Suppose our k-th regressor is non-binary then we define the partial effect of individual i at time t based on the conditional success probability

$$\begin{aligned}\Delta_{it}^{k} &= \frac{\partial \Pr(y_{it} = 1 | \mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i)}{\partial x_{itk}} \\ &= \Pr(y_{it} = 1 | \mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i)[1 - \Pr(y_{it} = 1 | \mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i)]\beta_k \, .\end{aligned} \tag{7}$$

In the situation where the k-th regressor is binary, we consider the difference between the conditional success probabilities, where once all observations of the regressor are set to one and once all are set to zero

$$\Delta_{it}^{k} = \Pr(y_{it} = 1 | x_{itk} = 1, \mathbf{x}_{it\{-k\}}, \boldsymbol{\beta}, \alpha_i) - \Pr(y_{it} = 1 | x_{itk} = 0, \mathbf{x}_{it\{-k\}}, \boldsymbol{\beta}, \alpha_i) \, . \tag{8}$$

An estimator of the APEs can be formed by replacing (7) or (8) by their sample analogues and taking the average[8]

$$\hat{\delta}_k = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \hat{\Delta}_{it}^{k} \, . \tag{9}$$

This is straightforward for UCL because we can simply plug their estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ into the corresponding formulas (7) and (8).[9] However, CL does not provide any estimates of $\boldsymbol{\alpha}$ to form the plug-in estimator (9). A simple but inconsistent approach is to assume that all fixed effects estimates are zero.[10]

Another quantity of interest are the standard errors of APEs. They can be either estimated using bootstrap techniques or the delta method. If at least one of the panel dimensions is large, bootstrapping becomes impractical since we have to re-estimate the model multiple times. Thus the preferred strategy is the delta method. Using this approach, the corresponding

[8] When calculating the average, it is important to include those individuals who do not have a varying response. Since their log-likelihood contributions are zero, these individuals do not contribute to the identification of the structural parameters. However, these individuals are still informative about partial effects. The corresponding partial effects are zero (see appendix C).

[9] Note that APEs obtained by UCL are also affected by IPP, but bias corrections are available (e.g. Hahn and Newey 2004; Carro 2007; Fernández-Val 2009; Dhaene and Jochmans 2015).

[10] This approach is used for example by the software package *Stata* in post-estimation routines of *clogit* and *xtlogit*.

covariance matrix for APEs can be estimated as follows:

$$\widehat{\mathbf{V}}(\hat{\boldsymbol{\delta}}) = \hat{\mathbf{J}}\widehat{\mathbf{V}}(\hat{\boldsymbol{\theta}})\hat{\mathbf{J}}' \, ,$$

where $\hat{\mathbf{J}} = \partial\hat{\boldsymbol{\delta}}/\partial\hat{\boldsymbol{\theta}}'$ is the Jacobian and $\hat{\boldsymbol{\delta}} = [\hat{\delta}_1, \ldots, \hat{\delta}_M]'$ is the vector containing estimates of the APEs. In section 5 we present solutions to the aforementioned problems that are also feasible in case of large panel data.

3 Computationally Efficient Unconditional Logit Estimation

Greene (2004) and Chamberlain (1980), among others, propose an efficient algorithm which results in identical parameter estimates as the dummy variable approach. Their method avoids the inversion of the large Hessian in (3) by utilizing the partitioned inverse formula and exploiting the sparsity of the Hessian. We show how the Frisch-Waugh-Lovell (FWL) theorem (Frisch and Waugh 1933; Lovell 1963) can be applied alternatively.

Our basic idea is to use the fact that the parameter updates of the Newton-Raphson routine is the solution of a weighted least squares problem. This allows to apply the well-known FWL theorem to separate the updates of structural parameters from the ones of the fixed effects. Due to the sparsity of the corresponding projection matrix we can derive a straightforward and computationally efficient update formula based on transformed regressors. This transformation is comparable to the demeaning procedure of a linear fixed effects model. Since in our approach the demeaning involves weights and takes place in each iteration step of the optimization routine, we call the procedure *pseudo-demeaning*.

In order to derive the efficient pseudo-demeaning algorithm we need to reconsider the naive dummy variable approach presented in section 2. Since the weighting matrix $\mathbf{W}$ is positive definite and diagonal, (4) is equivalent to the solution of a regression of the dependent variable $\tilde{\mathbf{y}} = (\mathbf{y} - \mathbf{p}) \odot \tilde{\mathbf{w}}^{-1}$ on the independent variables $\widetilde{\mathbf{Z}} = \tilde{\mathbf{w}} \odot \mathbf{Z}$, where $\tilde{\mathbf{w}}$ is the square-root of the diagonal of $\mathbf{W}$. The corresponding regression model is

$$\tilde{\mathbf{y}} = \widetilde{\mathbf{X}}(\boldsymbol{\beta}_0^k - \boldsymbol{\beta}_0^{k-1}) + \widetilde{\mathbf{D}}(\boldsymbol{\alpha}_0^k - \boldsymbol{\alpha}_0^{k-1}) + \mathbf{u}\,, \tag{10}$$

where the subscript zero denotes the population parameters, $\widetilde{\mathbf{X}} = \tilde{\mathbf{w}} \odot \mathbf{X}$, $\widetilde{\mathbf{D}} = \tilde{\mathbf{w}} \odot \mathbf{D}$, and $\mathbf{u}$ is an error term. Using reformulation (10) we can apply the FWL theorem to separate the high-dimensional fixed effects update from the structural parameter update. In terms of our problem, the FWL theorem states that if we regress the residuals obtained from a regression of $\tilde{\mathbf{y}}$ on $\widetilde{\mathbf{D}}$ on the residuals from separate regressions of each column of $\widetilde{\mathbf{X}}$ on $\widetilde{\mathbf{D}}$, we get the same parameter estimates $(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})$ as if we estimate the original regression model (10). Thus, pre-multiplying (10) with the projection matrix $\mathbf{Q} = \mathbf{I}_{NT} - \mathbf{P} = \mathbf{I}_{NT} - \widetilde{\mathbf{D}}(\widetilde{\mathbf{D}}'\widetilde{\mathbf{D}})^{-1}\widetilde{\mathbf{D}}'$ eliminates the fixed effects and residualizes the remaining variables $\tilde{\mathbf{y}}$ and $\widetilde{\mathbf{X}}$. The resulting concentrated regression is

$$\mathbf{Q}\tilde{\mathbf{y}} = \mathbf{Q}\widetilde{\mathbf{X}}(\boldsymbol{\beta}_0^k - \boldsymbol{\beta}_0^{k-1}) + \mathbf{Q}\mathbf{u}$$

and has the solution

$$(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) = (\widetilde{\mathbf{X}}'\mathbf{Q}\mathbf{Q}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\mathbf{Q}\mathbf{Q}\tilde{\mathbf{y}}\,. \tag{11}$$

Since the matrix $\mathbf{Q}$ is idempotent and symmetric, (11) can be further transformed while retaining the same parameter estimates[11]

$$(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) = (\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}\ddot{\mathbf{X}}'\tilde{\mathbf{y}}\,, \tag{12}$$

where $\ddot{\mathbf{X}} = \mathbf{Q}\widetilde{\mathbf{X}}$. Noticing the special sparse structure of $\mathbf{Q}$, the projection $\mathbf{Q}\widetilde{\mathbf{X}}$ can be computed without having to create the $NT \times NT$ projection matrix. In fact, $\mathbf{Q}\widetilde{\mathbf{X}}$ translates into an efficiently implementable and intuitive weighted demeaning formula which allows to compute

[11] This transformation would not be useful in a linear regression model, since the residuals of (11) and (12) differ, and thus the standard errors would be incorrect.

the parameter updates given in (12) at minimal computational costs

$$(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{\mathbf{x}}_{it} \ddot{\mathbf{x}}_{it}' \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{\mathbf{x}}_{it} \tilde{y}_{it} \right) , \tag{13}$$

where $\tilde{\mathbf{x}}_{it} = \tilde{w}_{it}\mathbf{x}_{it}$, $\tilde{y}_{it} = (y_{it} - p_{it})/\tilde{w}_{it}$, and $\ddot{\mathbf{x}}_{it} = \tilde{\mathbf{x}}_{it} - (\tilde{w}_{it} \sum_{t=1}^{T} \tilde{w}_{it}\tilde{\mathbf{x}}_{it}) / \sum_{t=1}^{T} \tilde{w}_{it}^2$.

Unlike a linear regression model, we also need to recover the estimates of the fixed effects to update the weights of the iterative maximization algorithm. Re-arranging (10) yields the update formula of the fixed effects estimates

$$(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) = (\widetilde{\mathbf{D}}'\widetilde{\mathbf{D}})^{-1}\widetilde{\mathbf{D}}' \left(\tilde{\mathbf{y}} - \widetilde{\mathbf{X}}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) \right) , \tag{14}$$

which depends on the previously computed structural parameter updates. Similarly to the updates of the structural parameters, formula (14) can be simplified by the block-diagonal structure of $(\widetilde{\mathbf{D}}'\widetilde{\mathbf{D}})^{-1}\widetilde{\mathbf{D}}'$ as follows:

$$(\alpha_i^k - \alpha_i^{k-1}) = \frac{\sum\limits_{t=1}^{T} \tilde{w}_{it}\tilde{y}_{it}}{\sum\limits_{t=1}^{T} \tilde{w}_{it}^2} - \frac{\sum\limits_{t=1}^{T} \tilde{w}_{it}\tilde{\mathbf{x}}_{it}'}{\sum\limits_{t=1}^{T} \tilde{w}_{it}^2} (\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) \, . \tag{15}$$

After we have derived all components to update the model parameters $\boldsymbol{\theta}$ efficiently, we can now introduce the entire optimization algorithm, which is linear in N and T.[12] This estimation routine is concisely summarized in algorithm 1.

Algorithm 1 Newton-Raphson with Pseudo-Demeaning

```
Initialize $\boldsymbol{\beta}^0$, $\boldsymbol{\alpha}^0$, and $k = 0$.
repeat
    Set $k = k + 1$.
    Compute $\mathbf{p}^{k-1}$ (see formula (1)).
    Compute $\tilde{\boldsymbol{y}}^{k-1}$ and $\ddot{\mathbf{X}}^{k-1}$ to update $\boldsymbol{\beta}^k$ (see formula (13)).
    Update $\boldsymbol{\alpha}^k$ (see formula (15)).
until convergence.
```

Finally, we show how to obtain the standard errors of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ after convergence of

[12] A detailed derivation of the computational complexity is presented in appendix B.

algorithm 1. Instead of estimating the covariance matrix of $\hat{\boldsymbol{\beta}}$ as the inverse of the entire negative Hessian, it can be easily obtained by its concentrated counterpart

$$\widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \left(\ddot{\mathbf{X}}'\ddot{\mathbf{X}}\right)^{-1} = -\ddot{\mathbf{H}}^{-1} \, .$$

Similarly, the variance of $\hat{\boldsymbol{\alpha}}$ can be computed as

$$\widehat{\text{Var}}(\hat{\alpha}_i) = \frac{1}{\sum_{t=1}^{T} \tilde{w}_{it}^2} + \left(\frac{\sum_{t=1}^{T} \tilde{w}_{it} \tilde{\mathbf{x}}_{it}}{\sum_{t=1}^{T} \tilde{w}_{it}^2} \right)' \widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) \left(\frac{\sum_{t=1}^{T} \tilde{w}_{it} \tilde{\mathbf{x}}_{it}}{\sum_{t=1}^{T} \tilde{w}_{it}^2} \right) .$$

Additionally we would like to draw the reader's attention to the fact that our pseudo-demeaning approach can be combined with different post-estimation bias corrections to reduce the incidental parameters bias; e.g. the analytical ones of Hahn and Newey (2004) and Fernández-Val (2009) or the jack-knife approaches of Hahn and Newey (2004) and Dhaene and Jochmans (2015). Especially if the panel is large the analytical corrections are advantageous because they only require to estimate the model once and the entire estimation procedure remains linear in N and T.

4 Conditional Logit with Random Subsets

As discussed above, CL can be attractive since it delivers fixed T consistent estimates for the structural parameters $\boldsymbol{\beta}$. However, it suffers from large computational costs with a long individual time series T. In this section, we introduce a new estimator that reduces this burden at the costs of efficiency.

Similar to the binary case, the multinomial logit estimator (CML) faces a huge computational burden in the presence of many alternatives. McFadden (1978) introduced a consistent but less efficient estimator for the multinomial logit model that overcomes this curse of dimensionality. We denote this estimator as CMLsub. Contrary to CML, it uses only random subsets of all possible permutations. Recently, D'Haultfœuille and Iaria (2016) analyzed the behavior of this estimator for a five-alternative multinomial logit model in a simulation study with respect to bias and computation time. Their key findings are that CMLsub is asymptotically less efficient than CML and that increasing the number of sampled permutations increases the precision. Thus, CMLsub becomes especially attractive when CML is either computationally too costly or not feasible at all.

For the binary fixed effects logit model, this approach is very similar, and we denote the corresponding estimator as CLsub. Instead of using the entire set $\mathscr{B}(t_{1i})$ of all permutations in the denominator of equation (6), we only use a random subset $\mathscr{D}(t_{1i})$ which contains m elements of $\mathscr{B}(t_{1i})$ where we make sure that the observed sequence is included. For brevity, we denote $\mathscr{B}(t_{1i})$ and $\mathscr{D}(t_{1i})$ as $\mathscr{B}$ and $\mathscr{D}$, respectively. Suppose that $\mathscr{D}$ is drawn conditionally on the observed choice $\mathbf{y}_i$ according to a probability $\pi(\mathscr{D}|\mathbf{y}_i)$. The key condition that we have to respect when creating the subset is the *uniform conditioning property* of McFadden (1978) which states: if $\mathbf{y}_i, \mathbf{d}_i \in \mathscr{D} \subseteq \mathscr{B}$, then $\pi(\mathscr{D}|\mathbf{y}_i) = \pi(\mathscr{D}|\mathbf{d}_i)$.[13] This condition holds if all remaining possible permutations of the observed choice sequence have the same probability of being selected in the subset, regardless of which choice sequence is observed.

In the following, the log-likelihood function of CLsub is derived. Given the joint success probability $\Pr(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta})$ of $\mathbf{y}_i \in \mathscr{B}$ conditioned on covariates $\mathbf{x}_i$ and given the probability $\pi(\mathscr{D}|\mathbf{y}_i)$ of selecting a subset $\mathscr{D} \subseteq \mathscr{B}$, the joint probability of $(\mathbf{y}_i, \mathscr{D})$ is $\pi(\mathscr{D}|\mathbf{y}_i)\Pr(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta})$ and hence the conditional probability of $\mathbf{y}_i$ given $\mathscr{D}$ is

$$\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathscr{D}, \boldsymbol{\beta}) = \frac{\pi(\mathscr{D}|\mathbf{y}_i)\Pr(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta})}{\sum_{\mathbf{d}_i \in \mathscr{D}} \pi(\mathscr{D}|\mathbf{d}_i)\Pr(\mathbf{d}_i|\mathbf{x}_i, \boldsymbol{\beta})} \tag{16}$$

with $\Pr(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}) = \prod_{t=1}^{T} \exp(y_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i))/(1 + \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i))$. Equation (16) can be rewritten to

$$\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathscr{D}, \boldsymbol{\beta}) = \frac{\pi(\mathscr{D}|\mathbf{y}_i)\prod_{t=1}^{T} \exp(y_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i))}{\sum_{\mathbf{d}_i \in \mathscr{D}} \pi(\mathscr{D}|\mathbf{d}_i)\prod_{t=1}^{T} \exp(d_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i))}\,. \tag{17}$$

[13] The validity of the uniform conditioning property can be shown as follows: $\mathscr{D}$ is selected to contain $\mathbf{y}_i$ plus $m-1$ random permutations $\mathbf{d}_i$ of $\mathbf{y}_i$. There are $c_i = \binom{T}{k(\mathbf{y}_i)}$ ways to place $k(\mathbf{y}_i) = \sum_{t=1}^{T} y_{it}$ ones in T slots, and hence $(c_i - 1)!/((m-1)!(c_i - m)!)$ ways to randomly select $m-1$ permutations of $\mathbf{y}_i$ without replacement. Thus $\pi(\mathscr{D}|\mathbf{y}_i) = ((m-1)!(c_i - m)!)/(c_i - 1)!$ depends only on $k(\mathbf{y}_i)$. Since any permutation $\mathbf{d}_i$ of $\mathbf{y}_i$ has the same c_i, it follows $\pi(\mathscr{D}|\mathbf{d}_i) = \pi(\mathscr{D}|\mathbf{y}_i)$, which is the uniform conditioning property of McFadden (1978).

Let $k(\mathbf{y}_i) = \sum_{t=1}^{T} y_{it}$, then equation (17) can be further simplified since $k(\mathbf{y}_i) = k(\mathbf{d}_i)$ and thus the fixed effect α_i is conditioned out

$$\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathscr{D}, \boldsymbol{\beta}) = \frac{\pi(\mathscr{D}|\mathbf{y}_i)\prod_{t=1}^{T}\exp(y_{it}\mathbf{x}_{it}'\boldsymbol{\beta})}{\sum_{\mathbf{d}_i \in \mathscr{D}} \pi(\mathscr{D}|\mathbf{d}_i)\prod_{t=1}^{T}\exp(d_{it}\mathbf{x}_{it}'\boldsymbol{\beta})} \,. \tag{18}$$

The application of the uniform conditioning property, reduces (18) to

$$\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathscr{D}, \boldsymbol{\beta}) = \frac{\prod_{t=1}^{T}\exp(y_{it}\mathbf{x}_{it}'\boldsymbol{\beta})}{\sum_{\mathbf{d}_i \in \mathscr{D}} \prod_{t=1}^{T}\exp(d_{it}\mathbf{x}_{it}'\boldsymbol{\beta})} \,,$$

which can be finally used to form the log-likelihood function of CLsub

$$L_{sub}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log\left(\frac{\exp\left(\sum_{t=1}^{T}\mathbf{x}_{it}' y_{it}\boldsymbol{\beta}\right)}{\sum_{d_i \in \mathscr{D}(t_{1i})}\exp\left(\sum_{t=1}^{T}\mathbf{x}_{it}' d_{it}\boldsymbol{\beta}\right)}\right) .$$

Next, we encounter a practical problem with the implementation of CLsub. A naive approach would first generate $\mathscr{B}$ to sample $\mathscr{D}$ from it. However, this approach has two shortcomings: it requires a lot of memory and for data sets with large T the computation of $\mathscr{B}$ is infeasible. Therefore we recommend to randomly shuffle the observed choice sequence $m-1$ times and to store the positions of the successes on each occasion. Multiple permutations are deleted and the process is repeated until the subset contains m unique permutations.

Compared to CL, which uses the entire permutation set $\mathscr{B}$ in the log-likelihood, CLsub reduces the number of arithmetic operations per individual from $c_i t_{1i} - 1$ to $m t_{1i} - 1$. Hence, it can be derived that CLsub requires $O(m \sum_{i=1}^{N} t_{1i})$ time, which means that the shape of the computational complexity depends on the choice of m (see appendix B).[14] Note that the theoretical derivation of the computational complexity is based on the assumption that $\mathscr{D}$ is already generated. From a practical point of view the total computation time, including the sampling of $\mathscr{D}$, is of interest. This will be the subject of our simulation experiments presented in section 6.

[14] For example if m is a linear function of T the computational complexity evolves roughly quadratically in T.

5 Feasible Estimation of Average Partial Effects

5.1 Efficient Offset Algorithm

So far, we have dealt with the problems of estimating structural parameters. In this section we tackle the remaining problems associated with the estimation of average partial effects.

Remember that one of the drawbacks of CL and CLsub is that they do not provide estimates of the fixed effects, so that the APE plug-in estimator (9) cannot be formed. In the following, we propose a simple ex-post estimation strategy to obtain estimates of the fixed effects. This is usually done by a so-called offset algorithm which in our case maximizes the log-likelihood function (2) while keeping the estimates of the structural parameters fixed at their values obtained by any conditional logit estimator.[15] The estimates obtained by this algorithm can in turn be used to calculate the APEs according to (9). The same type of algorithm is also required to alleviate the IPP using analytical bias corrections for average partial effects.[16]

We now turn to the derivation of an efficient offset algorithm, which is linear in N and T. Let $\tilde{\boldsymbol{\beta}}$ denote known estimates of the structural parameters. Maximizing (2) with $\mathbf{X}\tilde{\boldsymbol{\beta}}$ being fixed yields the Newton-Raphson update in iteration $(k-1)$

$$(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}) = (\widetilde{\mathbf{D}}'\widetilde{\mathbf{D}})^{-1}\widetilde{\mathbf{D}}'\tilde{\mathbf{y}} \, . \tag{19}$$

Thus, (19) can be efficiently computed according to

$$(\alpha_i^k - \alpha_i^{k-1}) = \frac{\sum_{t=1}^{T} \tilde{w}_{it}\tilde{y}_{it}}{\sum_{t=1}^{T} \tilde{w}_{it}^2} \tag{20}$$

and the whole procedure is repeated until convergence.[17]

[15] In an *offset* algorithm an additional variable is added to the linear predictor whose parameter is constrained to the value one (see Nelder and Wedderburn 1972).

[16] Analytical bias corrections of the APEs require, among other steps, that the fixed effects have to be re-estimated after bias-correcting the structural parameter estimates (see among others Hahn and Newey 2004).

[17] Note that $\mathbf{X}\tilde{\boldsymbol{\beta}}$ is still part of the linear predictor and thus has to be incorporated when updating the weights

In the context of CL, Bartolucci and Pigini (2019) suggest a refined version of our offset approach presented above. They use a strategy proposed by Firth (1993) to obtain an estimate of $\boldsymbol{\alpha}$ with improved finite sample properties by solving the following modified score equations

$$s^{Firth}(\boldsymbol{\alpha}) = \sum_{t=1}^{T} (y_{it} - p_{it}) + \frac{\sum_{t=1}^{T} p_{it}(1-p_{it})(1-2p_{it})}{2\sum_{t=1}^{T} p_{it}(1-p_{it})} = 0\,, \tag{21}$$

where $p_{it} = 1/(\exp(-\alpha_i - \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}}))$.

Solving the system (21) has the drawback that it becomes computationally demanding if N increases. Therefore, we follow Kosmidis and Firth (2009), who have shown that the solution of (21) can be obtained equivalently by using a standard Newton-Raphson algorithm with a modified dependent variable $\mathbf{y}^* = \mathbf{y} + \mathrm{diag}(\mathbf{S})(0.5 - \mathbf{p})$, where $\mathbf{S} = \mathbf{D}(\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}$. The sparse structure of $\mathbf{S}$ in turn suggests to compute the adjusted dependent variable as follows $y_{it}^* = y_{it} + (\tilde{w}_{it}^2 / \sum_{t=1}^{T} \tilde{w}_{it}^2)(0.5 - p_{it})$. Thus, we can use the same kind of efficient offset algorithm described previously by simply replacing the dependent variable in (20). Another modification compared to Bartolucci and Pigini (2019) is that we estimate the fixed effects of all n individuals.[18] We draw on a very recent result of Kunz, Staub, and Winkelmann (2018), who have proven that Firth's method can be used to obtain finite estimates of the fixed effects in probit models for perfectly classified individuals. It is straightforward to show that the same applies to logit models with fixed effects. Although the article of Kunz, Staub, and Winkelmann (2018) is about predicting fixed effects, we have found that their approach is also useful to obtain non-zero estimates of partial effects in the case of perfect classification.

5.2 Concentrated Delta Method

Next, we address the estimation of the standard errors for the APEs. The attentive reader might have noticed that using the brute force delta method as described in section 2 is problematic, because it requires the entire covariance matrix of $\hat{\boldsymbol{\theta}}$. However, with our pseudo-

and the adjusted dependent variable.

[18] Bartolucci and Pigini (2019) seem to estimate fixed effects only for individuals with varying responses. The specific approach is not clear from the methodological part of their article. However, a replication of their simulation results indicates that they only consider the APEs obtained from non-perfectly classified observations.

demeaning approach for UCL estimation described in section 3, we have only a reduced covariance matrix corresponding to the structural parameters and the variance of the fixed effects. The same obstacle occurs when we estimate the APEs for conditional logit estimators using the (modified) offset algorithm.

A solution to this problem consists of a concentrated delta method, which we derive from a combination of the results of Fernández-Val and Weidner (2016) and our pseudo-demeaning approach. To be more precise, Fernández-Val and Weidner (2016) have suggested an estimator for the covariance of APEs for nonlinear models with individual and time fixed effects. Thanks to the fact that our approach is based on the FWL theorem, it is straightforward to translate their estimator to the case of individual fixed effects and to exploit the sparsity of several terms included.[19] Assuming that the individual fixed effects are independent, the variance estimator of the APEs is given by

$$\widehat{\mathbf{V}}(\hat{\boldsymbol{\delta}}) = \frac{1}{N^2T^2}\left(\sum_{i=1}^{N}\sum_{t=s=1}^{T}\hat{\bar{\boldsymbol{\Delta}}}_{it}\hat{\bar{\boldsymbol{\Delta}}}'_{is} + \sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{\boldsymbol{\Gamma}}_{it}\widehat{\boldsymbol{\Gamma}}'_{it}\right),$$

where

$$\widehat{\boldsymbol{\Gamma}}_{it} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial\widehat{\boldsymbol{\Delta}}_{it}}{\partial\beta} - \frac{\bar{\mathbf{x}}_{it}}{\tilde{w}_{it}}\frac{\partial\widehat{\boldsymbol{\Delta}}_{it}}{\partial\alpha_i}\right)'\widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}})\ddot{\mathbf{x}}_{it}\tilde{y}_{it} - \frac{\bar{\boldsymbol{\psi}}_{it}}{\tilde{w}_{it}}\frac{\partial l_{it}}{\partial\alpha_i},$$

$\boldsymbol{\psi}_{it}$ and $\ddot{\boldsymbol{\psi}}_{it}$ are the it-th rows of $\boldsymbol{\Psi}$ and $\mathbf{Q}\boldsymbol{\Psi}$, $\boldsymbol{\psi}_{it} = (\partial\widehat{\boldsymbol{\Delta}}_{it}/\partial\alpha_i)/\tilde{w}_{it}^2$, $\widehat{\boldsymbol{\Delta}}_{it} = [\widehat{\Delta}_{it}^1,\ldots,\widehat{\Delta}_{it}^M]'$, $\bar{\mathbf{x}}_{it} = \mathbf{x}_{it} - \ddot{\mathbf{x}}_{it}$, $\bar{\boldsymbol{\psi}}_{it} = \boldsymbol{\psi}_{it} - \ddot{\boldsymbol{\psi}}_{it}$, and $\hat{\bar{\boldsymbol{\Delta}}}_{it} = \widehat{\boldsymbol{\Delta}}_{it} - \hat{\boldsymbol{\delta}}$. Note that the first part of the variance estimator takes into account the variation induced by estimating sample instead of population means and the second term is a concentrated version of the delta method. Especially the former is not very well known from the standard textbook literature, but it substantially improves the finite sample properties of the estimator (see Fernández-Val and Weidner 2016).

[19] This relationship is not so obvious when we use the partitioned inverse formula instead of the FWL theorem.

6 Simulation Experiments

6.1 Simulation Design

In this section we analyze the statistical properties of UCL, BCL, CL, and CLsub in terms of structural parameters and APEs. Further, we investigate the computation times of the different estimation routines. BCL refers to the bias-corrected UCL estimator suggested by Fernández-Val (2009).[20] For CLsub we consider two variants which differ by the size of the random subset $\mathcal{D}$. To be more precise, we choose $m^* \in \{1, T/2\}$, where m^* denotes the size of $\mathcal{D}$ without the observed choice sequence $\mathbf{y}_i$. All estimators analyzed in the simulation study are implemented by ourselves in the same programming language to guarantee comparability.[21] UCL and BCL are estimated using the Newton-Raphson pseudo-demeaning approach introduced in section 3. For the recursive CL and the CLsub algorithm we use a standard Newton-Raphson optimization routine with numerical derivatives to make it comparable to the estimation routine used for UCL and BCL without unnecessarily blowing up the memory.[22]

For our simulation experiments we generate the data according to Greene (2004) as follows:

$$y_{it} = \mathbf{1}[\alpha_i + \beta_1 x_{it} + \beta_2 d_{it} + v_{it} > 0]\,, \tag{22}$$

where $v_{it} = \log(u_{it}/(1-u_{it}))$, $u_{it} \sim \mathcal{U}(0,1)$, $\beta_1 = \beta_2 = 1$ $x_{it} \sim \mathcal{N}(0,1^2)$, $d_{it} = \mathbf{1}[x_{it} + h_{it} > 0]$, $h_{it} \sim \mathcal{N}(0,1^2)$, $\alpha_i = \sqrt{T}\bar{\mathbf{x}}_i + a_i$, $\bar{\mathbf{x}}_i = T^{-1}\sum_t x_{it}$, $a_i \sim \mathcal{N}(0,1^2)$. This design is well suited to analyze the behavior of the various fixed effects estimators, as it introduces an approximately constant correlation between the unobserved heterogeneity and the regressors for different T.

[20] In an earlier version of this article we use the bias correction of Hahn and Newey (2004). However, we find that the bias correction of Fernández-Val (2009) has better finite sample properties, although both approaches are asymptotically equivalent.

[21] All estimators and replication scripts are available on request.

[22] We do not investigate the brute-force implementations of UCL and CL because their computational costs are unreasonable high and in most of our analyzed setups they are even infeasible. In addition, we do not use analytical first and second order derivatives of the recursive CL due to its enormous memory requirement especially for large T (see appendix A). Also note that we are not using a full recursive implementation of CL, but an algorithm that exploits the usage of previous results in the recurrence that is substantially faster (see Gaure 2012).

Throughout our experiments, we analyze several model specifications with different n and T. We also consider panels with unusual large T, which can be justified when we think about so-called *pseudo panels*, where n groups, each consisting of T statistical units, are observed. For instance, n could be the number of postal code areas and T the number of households living in each area.

6.2 Finite Sample Properties

First of all, we focus on the statistical properties of the different estimators for the structural parameters and APEs. In order to investigate the biases and inference accuracies, all tables report the bias and standard deviations (SD) in percent relative to the true parameter value, the ratio between the average standard errors and the standard deviation, as well as the empirical coverage probabilities at a nominal value of 95%. All results are obtained by 1,000 replications of 9 model specifications with $n = 1,000$ and $T \in \{4, 8, 10, 12, 16, 20, 50, 100, 200\}$. For the sake of brevity, we only report the results of the continuous regressor, since we make similar findings for the discrete regressor.[23]

Table 1 shows the corresponding results for the structural parameter β_1. The UCL estimator is strongly distorted by the incidental parameter bias, but the distortion decreases as the T increases. At $T = 50$ the estimator still suffers a percentage distortion of 2.51 and even at $T = 200$ the coverage probabilities are too low, although its bias is below one percent. On the other hand, we find that the bias correction considerably reduces the bias of the UCL estimator. However, since the bias correction is based on a large-T expansion, it also requires a sufficiently large T to eliminate most of the distortion (see Fernández-Val 2009). Whereas for $T = 4$ there is still a bias of 13.01 percent, for $T = 8$ it is already only 0.49 percent and finally disappears with increasing T. Furthermore, the bias correction already brings the coverage probabilities close to their nominal level for $T = 8$. As expected, the CL estimator is unaffected by the incidental parameter bias. It delivers almost undistorted estimates across all T and the coverage probabilities are almost at the desired 95 percent. Thus, we can consider CL as a benchmark for the bias correction and find that the properties of CL

[23] The results of the discrete regressor are available on request.

and BCL for the structural parameters become closer as T increases. CLsub delivers similar results as CL if its subset $\mathscr{D}$ is not too small relative to the entire permutation set $\mathscr{B}$. In the case of $m^* = 1$, CLsub is almost undistorted for $T \leq 12$ but the bias increases rapidly from $T = 16$. While the distortion for $T = 12$ is still 0.25 percent, it rises to 85.12 percent for $T = 200$. We observe a similar behavior for CLsub with $m^* = T/2$, albeit in a delayed form. Since m^* does not depend on the size of the entire permutation set, it is not surprising that the bias of of CLsub increases considerably for large T. The optimization problem also becomes very unstable and produces unreliable results if m^* is small relative to T. Altogether, these findings suggest a careful choice of m^*. In a direct comparison to CL, we also find that CLsub is less efficient and precise, which is reflected by a larger standard error and higher distortion. Interestingly, CLsub can maintain coverage probabilities of about 95 percent even in cases where it exhibits extreme distortions.

Next, we consider the statistical properties of the different estimators regarding the APEs. First, we discuss table 2, which summarizes the results for UCL, BCL and CL. Then we will look at CLsub, whose results are given in table 3. For CL we analyze the performance of the three different approaches to estimate APEs. We denote the first approach which neglects the contributions of the fixed effects as *naive*, the second approach which uses the offset algorithm to recover estimates of the fixed effects as *score*, and the third approach which is based on the modified score as *Firth*. Remarkably, the incidental parameter bias present in the UCL estimator of the structural parameters hardly transfers to the APEs. For $T = 4$ we find a distortion of 2.23 percent and for $T \geq 8$ the distortion is close to zero. This notable result is consistent with the finding of Fernández-Val (2009).[24] The bias correction delivers comparatively good results like UCL, with the exception of $T = 4$, where the distortion is substantially higher with 8.16 percent. In addition, both estimators provide coverage probabilities close to the level of 95 percent for $T \geq 8$. We now turn to CL. The *naive* approach has a persistent high bias that ranges between 19.61 and 29.25 percent across all T. The

[24] Fernández-Val (2009) shows that the components that drive the bias of uncorrected APEs are the variation of the individual effects and their impact on the regressors. He finds that the bias is small, even in panels with a short time dimension, for a wide range of different distributions of individual effects and regressors. On the other hand Fernández-Val (2009) motivates the need of bias corrections in models with lagged dependent variables, where the small bias property of static binary-choice models disappears.

Table 1: *Finite sample properties of* $\hat{\beta}_1$

		UCL	BCL	CL	CLsub	
					$m^* = 1$	$m^* = T/2$
$T = 4$	Bias	47.08	-13.01	0.17	0.81	0.35
	SD	11.48	5.13	7.34	11.08	8.78
	SE/SD	0.81	1.42	1.01	0.99	0.99
	CP .95	0.00	0.60	0.96	0.95	0.94
$T = 8$	Bias	19.69	-0.49	0.20	0.74	0.27
	SD	5.51	4.38	4.47	9.62	5.99
	SE/SD	0.91	1.06	1.02	0.98	1.01
	CP .95	0.03	0.96	0.96	0.95	0.95
$T = 10$	Bias	14.94	-0.29	0.03	0.82	0.29
	SD	4.54	3.82	3.85	9.05	5.40
	SE/SD	0.94	1.05	1.03	1.03	1.02
	CP .95	0.06	0.96	0.95	0.95	0.95
$T = 12$	Bias	11.91	-0.33	-0.14	0.25	-0.05
	SD	4.11	3.58	3.60	9.18	5.49
	SE/SD	0.92	1.01	0.99	1.01	0.94
	CP .95	0.14	0.95	0.95	0.95	0.94
$T = 16$	Bias	9.00	0.17	0.26	1.81	0.32
	SD	3.32	3.00	3.01	9.98	4.72
	SE/SD	0.95	1.02	1.00	0.98	1.01
	CP .95	0.20	0.95	0.95	0.95	0.95
$T = 20$	Bias	6.84	-0.03	0.02	1.06	0.07
	SD	2.90	2.68	2.68	10.62	4.71
	SE/SD	0.95	1.00	0.99	0.98	0.98
	CP .95	0.31	0.94	0.94	0.96	0.95
$T = 50$	Bias	2.51	-0.07	-0.06	8.19	0.82
	SD	1.77	1.72	1.72	23.82	5.70
	SE/SD	0.94	0.95	0.95	0.92	1.00
	CP .95	0.66	0.93	0.93	0.97	0.96
$T = 100$	Bias	1.29	0.03	0.03	60.13	2.40
	SD	1.16	1.14	1.14	165.22	10.27
	SE/SD	0.99	1.00	1.00	3.70	0.97
	CP .95	0.80	0.95	0.94	0.97	0.95
$T = 200$	Bias	0.66	0.04	0.04	85.12	5.55
	SD	0.80	0.80	0.80	184.48	22.21
	SE/SD	1.00	1.00	1.00	113.49	0.87
	CP .95	0.87	0.95	0.95	0.95	0.95

Note: Bias and SD denote biases and standard deviations in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; results based on 1,000 repetitions.

score approach leads to a considerable reduction of the distortion with increasing T and also improves the coverage probabilities. At $T = 200$ the bias is only 0.32 percent and the coverage probability is 94 percent. Firth's approach brings a further substantial improvement. Overall, it performs similar as UCL, but always a bit worse for $T \geq 8$. Whereas CL based on Firth's method still has a distortion of 1.01 percent in the case of $T = 8$, UCL is almost undistorted. Next we compare the different conditional logit estimator combined with Firth's method to each other in table 3.[25] With regard to the distortion of the APEs, we make similar observations as with the structural parameters. CLsub provides comparable low distortions as CL, as long as $T \leq 20$. However, if T becomes too large, the CLsub collapses and its distortions increase. This is again particularly extreme in the case of $m^* = 1$. A crucial difference to the structural parameters is that the inference of the APEs obtained with CLsub is invalid. We conjecture that this is due to the fact that the higher dispersion of the structural parameters carries over to the estimation of the standard errors of the APEs with the delta method.

[25] A complete table with the *naive* and *score* approach can be provided upon request. Overall, they perform substantially worse compared to Firth's approach.

Table 2: *Finite sample properties of* $\hat{\delta}_1$

		UCL	BCL	CL		
				naive	score	Firth
$T = 4$	Bias	-2.23	-8.16	29.25	-21.34	1.76
	SD	6.43	5.47	8.28	5.28	6.48
	SE/SD	0.92	0.93	1.24	0.89	0.89
	CP .95	0.92	0.63	0.14	0.01	0.91
$T = 8$	Bias	0.16	-0.22	26.27	-9.71	-1.01
	SD	4.13	4.04	5.08	3.77	3.95
	SE/SD	0.93	0.91	1.14	0.95	0.95
	CP .95	0.93	0.92	0.00	0.24	0.93
$T = 10$	Bias	0.12	-0.15	25.21	-7.68	-1.30
	SD	3.53	3.49	4.39	3.29	3.39
	SE/SD	0.96	0.95	1.14	0.99	0.98
	CP .95	0.94	0.93	0.00	0.34	0.93
$T = 12$	Bias	-0.11	-0.32	24.37	-6.52	-1.53
	SD	3.29	3.27	4.07	3.11	3.18
	SE/SD	0.94	0.93	1.09	0.96	0.96
	CP .95	0.94	0.93	0.00	0.44	0.90
$T = 16$	Bias	0.29	0.13	23.99	-4.47	-1.05
	SD	2.82	2.81	3.43	2.71	2.74
	SE/SD	0.95	0.94	1.09	0.98	0.97
	CP .95	0.93	0.93	0.00	0.60	0.92
$T = 20$	Bias	0.08	-0.04	23.18	-3.68	-1.11
	SD	2.53	2.53	3.01	2.46	2.46
	SE/SD	0.96	0.95	1.09	0.98	0.98
	CP .95	0.95	0.94	0.00	0.65	0.91
$T = 50$	Bias	-0.04	-0.08	21.12	-1.50	-0.65
	SD	1.67	1.67	1.89	1.66	1.65
	SE/SD	0.99	0.98	1.05	0.99	0.99
	CP .95	0.94	0.94	0.00	0.85	0.93
$T = 100$	Bias	0.02	0.00	20.26	-0.70	-0.31
	SD	1.23	1.23	1.26	1.22	1.22
	SE/SD	1.05	1.04	1.09	1.05	1.05
	CP .95	0.96	0.96	0.00	0.93	0.95
$T = 200$	Bias	0.03	0.03	19.61	-0.32	-0.14
	SD	1.05	1.05	0.86	1.05	1.05
	SE/SD	1.01	1.01	1.11	1.01	1.01
	CP .95	0.94	0.94	0.00	0.94	0.94

Note: Bias and SD denote biases and standard deviations in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; results based on 1,000 repetitions.

Table 3: *Finite sample properties of $\hat{\delta}_1$ (based on Firth's method)*

		CL	CLsub	
			$m^* = 1$	$m^* = T/2$
$T = 4$	Bias	1.76	2.10	1.84
	SD	6.48	9.60	7.70
	SE/SD	0.89	1.24	1.00
	CP .95	0.91	0.98	0.94
$T = 8$	Bias	-1.01	-0.89	-1.02
	SD	3.95	7.73	5.05
	SE/SD	0.95	1.87	1.22
	CP .95	0.93	1.00	0.98
$T = 10$	Bias	-1.30	-1.00	-1.20
	SD	3.39	7.29	4.53
	SE/SD	0.98	2.19	1.30
	CP .95	0.93	1.00	0.98
$T = 12$	Bias	-1.53	-1.45	-1.49
	SD	3.18	7.44	4.52
	SE/SD	0.96	2.39	1.27
	CP .95	0.90	1.00	0.98
$T = 16$	Bias	-1.05	-0.18	-1.01
	SD	2.74	7.54	3.92
	SE/SD	0.97	3.03	1.47
	CP .95	0.92	1.00	0.99
$T = 20$	Bias	-1.11	-0.76	-1.16
	SD	2.46	8.07	3.81
	SE/SD	0.98	3.61	1.57
	CP .95	0.91	1.00	1.00
$T = 50$	Bias	-0.65	2.35	-0.23
	SD	1.65	14.50	4.27
	SE/SD	0.99	13.05	3.32
	CP .95	0.93	1.00	1.00
$T = 100$	Bias	-0.31	8.65	0.66
	SD	1.22	38.56	7.34
	SE/SD	1.05	> 1000	8.22
	CP .95	0.95	1.00	1.00
$T = 200$	Bias	-0.14	17.56	1.54
	SD	1.05	54.44	14.16
	SE/SD	1.01	> 1000	22.44
	CP .95	0.94	1.00	1.00

Note: Bias and SD denote biases and standard deviations in percentage relative to the truth; SE/SD and CP. 95 refer to average ratios of standard errors and standard deviations and empirical coverage probabilities of 95 % confidence intervals; results based on 1,000 repetitions.

6.3 Computational Costs

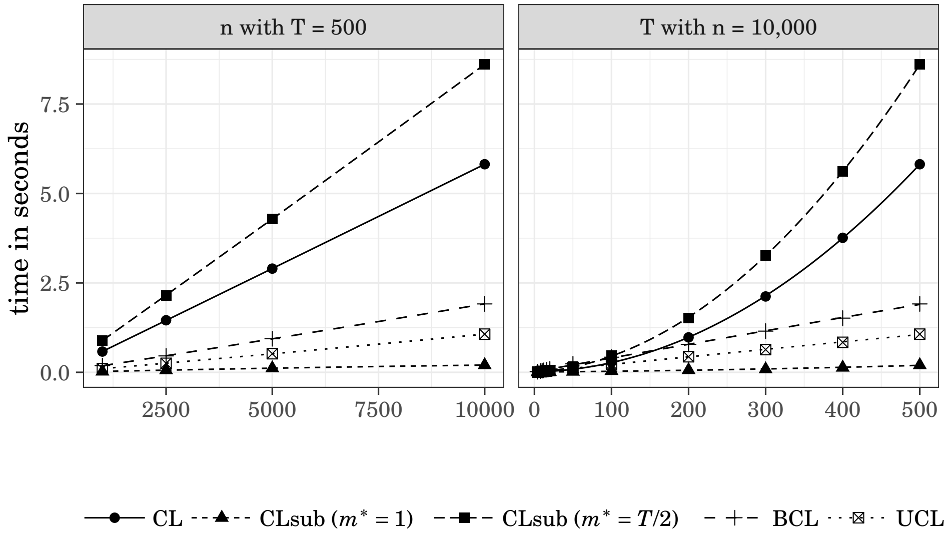
Aside from the statistical properties of the estimators, their computation times also matter for their application in practice. Whereas the theoretical computational complexities derived earlier in this article give a rough impression of the relationship between the panel dimension and the computation time, they do not reveal anything about the total magnitude of time required by an algorithm.

The computation times reported in table 4 are the averages of the respective fitting processes over 30 different data sets per $n-T$ combination generated according to (22). Furthermore, we investigate whether the theoretical computational complexities hold up empirically. To this end we measure the average computation times per iteration, since the estimators sometimes require a different number of iterations for each data set and $n-T$ combination. All calculations were done with the software R (R Core Team 2019) version 3.6.1 on a Linux Workstation with Intel Xeon E5-2640 v3 and 64 GB RAM.

Altogether our theoretical findings about the shape of the computational complexities are also verified empirically as shown in figure 1. The left figure depicts exemplary for $T = 500$ that all estimators evolve linearly in n. Moreover, UCL, BCL, and CLsub with $m^* = 1$ rise linear in T whereas CLsub with $m^* = T/2$ rises quadratically as demonstrated in the right figure for $n = 10,000$.

Table 4 depicts the enormous speed advantage of UCL and BCL compared to CL, especially when T becomes large. BCL takes on average 5.47 seconds when $T = 200$ and $n = 10,000$, whereas CL takes 9.77 seconds. The difference becomes even more dramatic when $T = 500$. In this case BCL requires 13.25 seconds and CL roughly 1 minute. As indicated in table 4, CLsub with $m^* = 1$ is faster than CL when $T \geq 100$ and CLsub with $m^* = T/2$ is not able to outperform CL. Furthermore, table 4 depicts that CLsub with $m^* = 1$ is negligibly slower than CL if T is small, but when T increases it outperforms CL by far. Table 4 also reveals two other notable results about CLsub. First, CLsub with $m^* = T/2$ is always much slower than the other two conditional logit estimators. On the other hand, the creation of the subset of the entire permutation set, whose computation time is shown in parentheses, accounts for

Figure 1: *Empirical Computational Complexities*



Note: Dots: average computation times per iteration in seconds; Curves: quadratic polynomial approximations.

Table 4: *Average Computation Times*

		UCL	BCL	CL	CLsub	
					$m^* = 1$	$m^* = T/2$
$n = 10{,}000$	$T = 4$	0.04	0.11	0.04	0.81 (0.75)	1.30 (1.23)
	$T = 8$	0.10	0.22	0.10	0.91 (0.83)	1.61 (1.49)
	$T = 10$	0.12	0.27	0.13	0.92 (0.83)	1.69 (1.52)
	$T = 12$	0.15	0.32	0.17	0.96 (0.85)	1.79 (1.59)
	$T = 16$	0.22	0.44	0.26	1.02 (0.89)	2.07 (1.76)
	$T = 20$	0.29	0.57	0.35	1.09 (0.91)	2.35 (1.95)
	$T = 50$	0.81	1.42	1.03	1.38 (1.00)	6.73 (4.01)
	$T = 100$	1.56	2.84	2.87	1.91 (1.08)	17.50 (8.55)
	$T = 200$	3.05	5.47	9.77	2.62 (1.20)	50.95 (22.32)
	$T = 300$	4.38	7.94	21.21	3.38 (1.31)	108.40 (42.47)
	$T = 400$	5.78	10.44	37.60	4.37 (1.43)	187.46 (68.35)
	$T = 500$	7.40	13.25	58.17	4.10 (1.50)	287.42 (98.08)
$T = 500$	$n = 1{,}000$	0.69	1.23	5.81	0.31 (0.15)	31.49 (9.98)
	$n = 2{,}500$	1.67	3.05	14.55	1.14 (0.37)	73.48 (24.66)
	$n = 5{,}000$	3.55	6.41	28.99	2.37 (0.75)	144.03 (49.06)
	$n = 10{,}000$	7.40	13.25	58.17	4.10 (1.50)	287.42 (98.08)

Note: Computation times in seconds; time needed for generating $\mathscr{D}$ in parentheses; results based on 30 repetitions.

a large part of the total computation time. Even after subtracting this time from the total computation time, the CL estimator is still much faster, especially at large T.

Summarizing the findings from the simulation experiments, we conclude that CLsub is not an option to CL. If we sample a sufficiently large subset from the entire permutation set, the estimator is still computationally more demanding and additionally less precise than CL. As we have shown in theory and simulation, CL quickly encounters computational challenges when T rises, although we have already employed the efficient recursive implementation. Moreover, the conditional logit estimators are outperformed by UCL and BCL in the estimation of APEs and computation times in general. Thus, UCL and BCL offer attractive alternatives.

7 Empirical Illustration

In this section we demonstrate the advantage of or pseudo-demeaning approach by providing an illustration from labor economics, where the brute-force dummy approach as well as the recursive CL approach fail due to computational limitations. We investigate the labor force participation of women using a data set from the *American Community Survey* (2017 ACS 1-YEAR PUMS). The data set can be interpreted as a pseudo panel where the cross-sectional units are *Public Use Microdata Areas* (PUMAs) and the time dimension translates into groups of women in these PUMAs. The data set consists of $1,294,938$ women in $N = 982$ PUMAs, where the smallest PUMA includes $T_i = 230$ and the largest $T_i = 26,772$ women.

We specify our model as follows:

$$\begin{aligned} work_{it} &= \mathbf{1}\left[\eta_{it} \geq v_{it}\right] , \\ \eta_{it} &= \alpha_i + \sum_j \gamma_j educ_{jit} + \beta_1 age_{it} + \beta_2 mar_{it} + \beta_3 inc_{it} + \beta_4 kids6_{it} , \end{aligned}$$

where i and t refer to the t-th woman in PUMA i, $work$ denotes the labor force participation status, age refers to the age in years, mar is the marital status, inc is the household income without the labor earnings of the woman in thousand dollars, $educ_j$ are indicators of different educational attainments[26], and $kids6$ is an indicator of the presence of children under the age of 6 years.

Table 5 shows the estimates of the structural parameters (left panel), APEs (right panel), and the corresponding standard errors (in parentheses). We observe that the bias-corrected and uncorrected estimates are almost identical, which is as expected due to large T_i. Overall the results are intuitive. For instance, higher education has a significant positive impact on the probability to participate in the labor force. Having a high school degree increases the probability by 26.2 percentage points relative to a woman with no high school degree. Further the presence of young and new born children lowers the probability to participate. Interestingly, the transitory non-labor household income does not affect the participation decision, which is in line with Hyslop (1999).

To demonstrate that the bias correction also works with real data, we extract a subset from the entire data set by randomly drawing $T_i = 8$ observations from each PUMA. Now that T is small enough to be handled by the CL estimator, we can use it as a benchmark for the performance of the bias correction due to its fixed T consistency property in case of structural parameter estimation. Furthermore, the small T makes a bias correction of the UCL estimator necessary. The results shown in table 6 are in line with the findings in the simulation study. Whereas the structural parameter estimates of CL and BCL are close to each other, the corresponding estimates obtained by UCL differ remarkably. However, although the structural parameter estimates obtained by the UCL estimator are clearly distorted, its estimated APEs hardly differ from the bias-corrected ones. With the CL estimator, we get substantially lower estimated APEs in terms of magnitude.

[26] More precisely, $educ$ has three levels: no high-school degree, high-school degree, college and/or university degree.

Table 5: *Estimation results based on the entire sample*

	$\hat{\boldsymbol{\beta}}$		$\hat{\boldsymbol{\delta}}$	
	UCL	BCL	UCL	BCL
college / university	2.229 (0.008)	2.227 (0.008)	0.375 (0.001)	0.375 (0.001)
highschool	1.549 (0.007)	1.548 (0.007)	0.262 (0.001)	0.262 (0.001)
age	-0.057 (0.000)	-0.057 (0.000)	-0.011 (0.000)	-0.011 (0.000)
married	0.285 (0.004)	0.284 (0.004)	0.053 (0.001)	0.053 (0.001)
kids6	-0.741 (0.007)	-0.741 (0.007)	-0.139 (0.001)	-0.139 (0.001)
nlinc	-0.003 (0.000)	-0.003 (0.000)	-0.001 (0.000)	-0.001 (0.000)

Note: $\hat{\boldsymbol{\beta}}$ denotes estimates of the structural parameters; $\hat{\boldsymbol{\delta}}$ denotes estimates of APEs; standard errors in parenthesis; standard errors of $\hat{\boldsymbol{\delta}}$ are computed with the delta method.

Table 6: *Estimation results based on a subsample*

	$\hat{\boldsymbol{\beta}}$			$\hat{\boldsymbol{\delta}}$		
	UCL	BCL	CL	UCL	BCL	CL Firth
college / university	2.292 (0.108)	1.939 (0.102)	1.952 (0.099)	0.344 (0.013)	0.339 (0.013)	0.323 (0.012)
highschool	1.619 (0.097)	1.367 (0.092)	1.376 (0.089)	0.247 (0.014)	0.243 (0.014)	0.230 (0.013)
age	-0.062 (0.002)	-0.053 (0.002)	-0.053 (0.002)	-0.010 (0.000)	-0.010 (0.000)	-0.009 (0.000)
married	0.363 (0.065)	0.307 (0.063)	0.310 (0.060)	0.060 (0.010)	0.058 (0.010)	0.055 (0.010)
kids6	-1.003 (0.103)	-0.855 (0.100)	-0.862 (0.095)	-0.164 (0.017)	-0.163 (0.017)	-0.154 (0.016)
income	-0.003 (0.000)	-0.003 (0.000)	-0.003 (0.000)	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)

Note: $\hat{\boldsymbol{\beta}}$ denotes estimates of the structural parameters; $\hat{\boldsymbol{\delta}}$ denotes estimates of APEs; standard errors in parenthesis; standard errors of $\hat{\boldsymbol{\delta}}$ are computed with the delta method.

8 Conclusion

This paper discussed and addressed the disadvantages of the two most commonly used estimators for logit models with fixed effects, especially in the case of data sets where many cross-sectional units are observed for long time horizons. These are the conditional and the unconditional logit estimators. In a series of simulation experiments we found that the (bias-corrected) unconditional logit estimator has desirable finite sample properties with respect to structural parameters and average partial effects. Furthermore, by combining the estimator with our novel pseudo-demeaning approach, our algorithm is linear in both panel dimensions.

Thus, the (bias-corrected) unconditional logit estimator is a promising candidate for many relevant applications based on large panel data. To allow the readers to use our algorithm in a straightforward and convenient way, we provide an implementation in our R-package *bife*.

We would like to draw the attention of our readers to the fact that our pseudo-demeaning paves the way to derive algorithms for more complex nonlinear fixed effects models. Stammann (2018) combines the pseudo-demeaning with the method of alternating projections (MAP) to develop a feasible algorithm for the estimation of generalized linear models with multiple high-dimensional fixed effects. The combination of MAP and pseudo-demeaning can also be extended to bias corrections with multiple fixed effects, as shown by Czarnowske and Stammann (2019) for binary choice models with additive unobservable individual and time effects.

References

Andersen, Erling Bernhard. 1970. “Asymptotic properties of conditional maximum-likelihood estimators.” *Journal of the Royal Statistical Society. Series B:* 283–301.

Arellano, Manuel, and Jinyong Hahn. 2007. “Understanding bias in nonlinear panel models: Some recent developments.” *Econometric Society Monographs* 43:381.

Bartolucci, Francesco, and Claudia Pigini. 2019. “Partial effects estimation for fixed-effects logit panel data models.” *Working Paper.*

Carro, Jesus M. 2007. “Estimating dynamic panel data discrete choice models with fixed effects.” *Journal of Econometrics* 140 (2): 503–528.

Chamberlain, Gary. 1980. “Analysis of Covariance with Qualitative Data.” *Review of Economic Studies* 47:225–238.

Czarnowske, Daniel, and Amrei Stammann. 2019. “Binary Choice Models with High-Dimensional Individual and Time Fixed Effects.” *arXiv preprint:1904.04217.*

D’Haultfœuille, Xavier, and Alessandro Iaria. 2016. “A convenient method for the estimation of the multinomial logit model with fixed effects.” *Economics Letters* 141:77–79.

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

Fernández-Val, Iván. 2009. “Fixed effects estimation of structural parameters and marginal effects in panel probit models.” *Journal of Econometrics* 150:71–85.

Fernández-Val, Iván, and Martin Weidner. 2016. “Individual and time effects in nonlinear panel models with large N, T.” *Journal of Econometrics* 192 (1): 291–312.

———. 2018. “Fixed Effects Estimation of Large-T Panel Data Models.” *Annual Review of Economics* 10 (1): 109–138.

Firth, David. 1993. "Bias reduction of maximum likelihood estimates." *Biometrika* 80 (1): 27–38.

Frisch, Ragnar, and Frederick V. Waugh. 1933. "Partial Time Regressions as Compared with Individual Trends." *Econometrica* 1 (4): 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.

Gaure, Simen. 2012. "A Faster Algorithm for Computing the Conditional Logit Likelihood." *Unpublished Note.*

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.

Hahn, Jinyong, and Whitney 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.

Hyslop, Dean R. 1999. "State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women." *Econometrica* 67 (6): 1255–1294.

Kosmidis, Ioannis, and David Firth. 2009. "Bias reduction in exponential family nonlinear models." *Biometrika* 96 (4): 793–804.

Kunz, Johannes S., Kevin E. Staub, and Rainer Winkelmann. 2018. "Predicting fixed effects in panel probit models." *Working Paper.*

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

McFadden, Daniel. 1978. “Modeling the choice of residential location.” *Transportation Research Record,* no. 673.

Nelder, John Ashworth, and Robert William Maclagan Wedderburn. 1972. “Generalized linear models.” *Journal of the Royal Statistical Society: Series A (General)* 135 (3): 370–384.

Neyman, Jerzy, and Elizabeth L Scott. 1948. “Consistent estimates based on partially consistent observations.” *Econometrica* 16 (1): 1–32.

Prentice, Ross L., and Lynn A. Gloeckler. 1978. “Regression analysis of grouped survival data with application to breast cancer data.” *Biometrics:* 57–67.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rasch, George. 1960. “Probabilistic models for some intelligence and attainment tests: Danish institute for Educational Research.” *Denmark Paedogiska, Copenhagen.*

Reid, Stephen, and Rob Tibshirani. 2014. “Regularization paths for conditional logistic regression: The clogitl1 package.” *Journal of Statistical Software* 58 (12).

Stammann, Amrei. 2018. “Fast and Feasible Estimation of Generalized Linear Models with High-Dimensional k-way Fixed Effects.” *arXiv preprint:1707.01815v3.*

Appendix

A Details on the Implementations

A.1 Brute-Force UCL Estimation

Let $\mathbf{Z} = [\mathbf{X}, \mathbf{D}]$ denote the $NT \times (M+N)$ regressor matrix, where $\mathbf{D}$ is the $NT \times N$ dummy variable matrix corresponding to the fixed effects and $\mathbf{X}$ is the $NT \times M$ matrix of the remaining regressors. In a similar way as Greene (2004) we define the gradient and the Hessian of UCL. The $(N+M) \times 1$ gradient is given by

$$\mathbf{g} = [\mathbf{g}_\beta', \mathbf{g}_\alpha']'$$

with

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

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

and the $(N+M) \times (N+M)$ Hessian takes the following form

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

with

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

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

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

Thus, the $(k-1)$-th Newton-Raphson update in (3) can be rewritten as follows:

$$(\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}) = -\mathbf{H}^{-1}\mathbf{g} = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y}-\mathbf{p})\,, \tag{23}$$

where the $NT \times NT$ matrix $\mathbf{W}$ serves as a weighting matrix. $\mathbf{W}$ is a diagonal-matrix with strictly positive weights $w_{it} = p_{it}(1-p_{it})$ where p_{it} is defined in (1). Note that the weights and all dependent quantities are evaluated at $\boldsymbol{\theta}^{k-1}$.

A.2 Recursive CL Estimation

Gail, Lubin, and Rubinstein (1981) proposed an recursive implementation of CL. This approach accelerates the computation while retaining the exactness of the brute force approach presented in section 2.

The individual likelihood contribution in (6) can be rewritten as follows:

$$\exp(L_{ci}) = \frac{\prod_{k=1}^{t_{1i}} \exp(\mathbf{x}_k'\boldsymbol{\beta})}{\sum_{h=1}^{c_i}\prod_{k_h=1}^{t_{1i}} \exp(\mathbf{x}_{k_h}'\boldsymbol{\beta})}\,, \tag{24}$$

where the index k denotes the observed data and the index k_h the h-th possible assignment. Lets define the denominator in (24) as follows:

$$f_i(t_{1i}, T) = \sum_{h=1}^{c_i}\prod_{k_h=1}^{t_{1i}} \exp(\mathbf{x}_{k_h}'\boldsymbol{\beta}) = \sum_{h=1}^{c_i}\prod_{k_h=1}^{t_{1i}} U_{k_h}\,.$$

The recursion can be specified by

$$f_i(t_{1i}, T) = f_i(t_{1i}, T-1) + U_T f_i(t_{1i}-1, T-1)$$

with $f_i(0,T) = 1$ for $T \geq 0$, $f_i(t_{1i}, T) = 0$ for $t_{1i} > T$ and $U_T = \exp(\mathbf{x}_{iT}'\boldsymbol{\beta})$. Finally, the conditional log-likelihood in (5) can be rewritten to

$$L_c = \sum_{i=1}^{N} L_{ci} = \sum_{i=1}^{N}\left(\sum_{t=1}^{T} y_{it}\mathbf{x}'_{it}\boldsymbol{\beta} - \log(f_i(t_{1i}, T))\right). \tag{25}$$

The maximization of the conditional log-likelihood (25) is usually solved iteratively with gradient based maximization techniques. It is possible to apply the recurrence to the computation of the gradient and Hessian as well. However, this has not been proven to be useful since the recurrence is very time and memory consuming. Gail, Lubin, and Rubinstein (1981) proposed to implement the estimator based on numerical first and second order derivatives.

B Computational Complexities

For the following derivations we assume a balanced panel with $N \gg T \gg M$.

B.1 Brute-Force UCL Estimation

Remember the $(k-1)$-th Newton-Raphson step is

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

The most demanding part is the computation of the $(M+N)\times(M+N)$ Hessian.[27] The multiplication of the $NT \times (M+N)$ matrix $\mathbf{Z}$ with the $NT \times NT$ diagonal matrix $\mathbf{W}$ can be done in $\approx O(N^2T)$. Suppose we have already generated the variable $\mathbf{Z}_w = \mathbf{W}\mathbf{Z}$. The matrix multiplication $\mathbf{Z}'\mathbf{Z}_w$ costs $\approx O(N^3T)$, matrix multiplication $\mathbf{Z}'\mathbf{Y}$ costs $\approx O(N^2T)$, matrix inversion $(\mathbf{Z}'_w\mathbf{Z})^{-1}$ costs $\approx O(N^3)$ and finally the product of the Hessian and the gradient costs $\approx O(N^2)$. Since $O(N^3T) > O(N^2T) > O(N^2)$ the computation time increases cubically in N and linear in T.

[27] Note that some software routines, such as `glm()` in R, include n dummies instead of N and the computation becomes even more costly. Remember, $N = \sum_{i=1}^{n} \mathbf{1}[0 < \sum_{t=1}^{T} y_{it} < T]$ where n denotes the total number of individuals in the data set.

B.2 Computationally Efficient UCL Estimation

The computational complexity of the pseudo-demeaning can be derived by considering the most extensive part, which is the computation of the structural parameter updates

$$(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) = (\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}\ddot{\mathbf{X}}'\tilde{\mathbf{y}} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{x}}_{it}\ddot{\mathbf{x}}_{it}'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{x}}_{it}\tilde{y}_{it}\right).$$

Although $\ddot{\mathbf{X}}$ consists out of MNT elements, its computation requires only $\approx O(MNT)$ time, since $\sum_{t=1}^{T}\tilde{w}_{it}\tilde{x}_{it}$ is different for MN elements and $\sum_{t=1}^{T}\tilde{w}_{it}^2$ is different for N elements. Thus, $\sum_{t=1}^{T}\tilde{w}_{it}\tilde{x}_{it}$ requires $MN(T-1+T)$ arithmetic operations and $\sum_{t=1}^{T}\tilde{w}_{it}^2$ requires $N(T-1)$ arithmetic operations. The matrix multiplication $\ddot{\mathbf{X}}'\ddot{\mathbf{X}}$ costs $\approx O(M^2NT)$, matrix multiplication $\ddot{\mathbf{X}}'\widetilde{\mathbf{Y}}$ costs $\approx O(MNT)$, matrix inversion $(\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}$ costs $\approx O(M^3)$ and finally the product of the Hessian and the gradient costs $\approx O(M^2)$. Altogether, $O(M^2NT) > O(MNT) > O(M^3)$. Thus, the computation time of the pseudo-demeaning is linear in T and N.

B.3 Brute-Force CL Estimation

Next, we demonstrate the computational complexity of brute force implementation of CL. Taking into account that y_{it} is binary, (5) can be rewritten to

$$L_c(\boldsymbol{\beta}) = \sum_{i=1}^{N}\log\left(\frac{\exp\left(\sum_{k=1}^{t_{1i}}\mathbf{x}_{ik}'\boldsymbol{\beta}\right)}{\sum_{h=1}^{c_i}\exp\left(\sum_{k_h=1}^{t_{1i}}\mathbf{x}_{ik_h}'\boldsymbol{\beta}\right)}\right), \tag{26}$$

where the index k denotes the observed data and the index k_h the h-th possible assignment. Lets consider the individual likelihood contribution

$$\exp(L_{ci}) = \frac{\prod_{k=1}^{t_{1i}}\exp(\mathbf{x}_k'\boldsymbol{\beta})}{\sum_{h=1}^{c_i}\prod_{k_h=1}^{t_{1i}}\exp(\mathbf{x}_{k_h}'\boldsymbol{\beta})}. \tag{27}$$

A direct evaluation of the denominator in (27) requires the summation of c_i terms and becomes prohibitive if T increases (see Gail, Lubin, and Rubinstein 1981). The computation of the denominator involves roughly $t_{1i}c_i$ arithmetic operations: there are $(c_i - 1)$ outer additions and $(t_{1i}-1)$ inner multiplications. Thus, evaluating the log-likelihood, as it is required by a numerical optimization routine, costs $\approx O(\sum_{i=1}^{N} t_{1i}\binom{T}{t_{1i}})$. Since t_{1i} is a proportion of T, which usually grows with T, the complexity is exponential in T.

B.4 Recursive CL Estimation

In order to determine the computational complexity of the recursive implementation of CL, we consider how it tackles the problem of computing the denominator of (27). Reid and Tibshirani (2014) have shown that the denominator can be computed in $\approx O(t_{1i}(T-t_{1i}))$ time. Thus evaluating the log-likelihood, takes $\approx O(\sum_{i=1}^{N} t_{1i}(T-t_{1i}))$. Hence, the computational complexity is linear in N and roughly quadratic in T since t_{1i} is a proportion of T, which usually grows with T. Even if one follows Simen Gaure's recommendation not to set up the program completely recursively, but to reuse intermediate results, nothing changes in the form of computational complexity, since it is reduced only by a factor (see Gaure 2012).

B.5 CL Estimation with Random Subsets

CLsub reduces the number of arithmetic operations per individual from roughly $t_{1i}c_i$ with the brute force CL algorithm to $t_{1i}m$, since CLsub requires only $(m-1)$ outer additions and still $(t_{1i}-1)$ inner multiplications. Hence, CLsub requires $\approx O(m\sum_{i=1}^{N} t_{1i})$ to evaluate the log-likelihood function. Therefore, the shape of the computational complexity depends on the choice of m. If m is a function of T the computational complexity evolves roughly quadratically in T, else linearly.

C Details on Average Partial Effects

Remember that we estimate $N \leq n$ fixed effects since we use the reduced sample in the optimization of the log-likelihood. For those individuals who don't change their status over time (perfectly classified) the estimates of the fixed effects are unbounded. Thus, their estimates of partial effects are zero as shown in the following.

Non-binary regressor:

$$\lim_{\hat{\alpha}_i\to\infty}\hat{\Delta}^k_{it} = \underbrace{\lim_{\hat{\alpha}_i\to\infty}\Pr(y_{it}=1|\mathbf{x}_{it},\hat{\boldsymbol{\beta}},\hat{\alpha}_i)}_{=1}\underbrace{\lim_{\hat{\alpha}_i\to\infty}[1-\Pr(y_{it}=1|\mathbf{x}_{it},\hat{\boldsymbol{\beta}},\hat{\alpha}_i)]}_{0}\hat{\beta}_k = 0$$

$$\lim_{\hat{\alpha}_i\to-\infty}\hat{\Delta}^k_{it} = \underbrace{\lim_{\hat{\alpha}_i\to-\infty}\Pr(y_{it}=1|\mathbf{x}_{it},\hat{\boldsymbol{\beta}},\hat{\alpha}_i)}_{=0}\underbrace{\lim_{\hat{\alpha}_i\to-\infty}[1-\Pr(y_{it}=1|\mathbf{x}_{it},\hat{\boldsymbol{\beta}},\hat{\alpha}_i)]}_{1}\hat{\beta}_k = 0$$

Binary regressor:

$$\lim_{\hat{\alpha}_i\to\infty}\hat{\Delta}^k_{it} = \underbrace{\lim_{\hat{\alpha}_i\to\infty}\Pr(y_{it}=1|x_{itk}=1,\mathbf{x}_{it\{-k\}},\hat{\boldsymbol{\beta}},\hat{\alpha}_i)}_{=1} -$$

$$\underbrace{\lim_{\hat{\alpha}_i\to\infty}\Pr(y_{it}=1|x_{itk}=0,\mathbf{x}_{it\{-k\}},\hat{\boldsymbol{\beta}},\hat{\alpha}_i)}_{=1} = 0$$

$$\lim_{\hat{\alpha}_i\to-\infty}\hat{\Delta}^k_{it} = \underbrace{\lim_{\hat{\alpha}_i\to-\infty}\Pr(y_{it}=1|x_{itk}=1,\mathbf{x}_{it\{-k\}},\hat{\boldsymbol{\beta}},\hat{\alpha}_i)}_{=0} -$$

$$\underbrace{\lim_{\hat{\alpha}_i\to-\infty}\Pr(y_{it}=1|x_{itk}=0,\mathbf{x}_{it\{-k\}},\hat{\boldsymbol{\beta}},\hat{\alpha}_i)}_{=0} = 0$$