Received: 19 December 2019 | Accepted: 10 May 2021

DOI: 10.1111/rssa.12722

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

ORIGINAL ARTICLE

Predicting individual effects in fixed effects panel probit models

Johannes S. Kunz[1] | **Kevin E. Staub**[2] | **Rainer Winkelmann**[3]

[1]Centre for Health Economics, Monash Business School, Monash University, 15 Innovation Walk, Clayton, VIC 3800, Australia

[2]Department of Economics, The University of Melbourne, 111 Barry Street, Parkville, VIC 3010, Australia

[3]Department of Economics, University of Zurich, Zürichbergstrasse 14, Zürich, CH-8032, Switzerland

Correspondence
Kevin E. Staub, Department of Economics, 111 Barry Street, The University of Melbourne, 3010 VIC, Australia.
Email: kevin.staub@unimelb.edu.au

Funding information
Australian Research Council, Grant/Award Number: DE170100644

Abstract

Many applied settings in empirical economics require estimation of a large number of individual effects, like teacher effects or location effects; in health economics, prominent examples include patient effects, doctor effects or hospital effects. Increasingly, these effects are the object of interest of the estimation, and predicted effects are often used for further descriptive and regression analyses. To avoid imposing distributional assumptions on these effects, they are typically estimated via fixed effects methods. In short panels, the conventional maximum likelihood estimator for fixed effects binary response models provides poor estimates of these individual effects since the finite sample bias is typically substantial. We present a bias-reduced fixed effects estimator that provides better estimates of the individual effects in these models by removing the first-order asymptotic bias. An additional, practical advantage of the estimator is that it provides finite predictions for all individual effects in the sample, including those for which the corresponding dependent variable has identical outcomes in all time periods over time (either all zeros or ones); for these, the maximum likelihood prediction is infinite. We illustrate the approach in simulation experiments and in an application to health care utilization.

KEYWORDS

bias reduction, binary response, doctor visits, fixed effects, health care utilization, incidental parameter bias, panel data, perfect prediction

JEL CLASSIFICATION

C23; C25; I11; I18

1 | INTRODUCTION

This paper addresses estimation and prediction of individual-specific effects in the fixed effects (FE) panel probit model when the individual dimension N is large and the time dimension T is relatively small. The view of individual—or, more generally, panel unit—effects not as nuisance parameters but rather as features of interest has received increased attention in applied work lately, for example in the context of neighbourhood effects (Chetty & Hendren, 2018), teacher effects (Chetty et al., 2014) and judge effects (Abrams et al., 2012), to name but a few. All these studies employ the linear fixed effects model to obtain individual effects, although the dependent variable is frequently binary.

In such cases, it clearly would be preferable to extract the individual effects from a proper (non-linear) binary response model, such as the logit or probit model. However, conventional estimation methods such as maximum likelihood (ML) face two major obstacles. First, to obtain a consistent and asymptotically normal estimator of all the parameters of the model requires an asymptotic theory where T increases without bound. In finite samples, the ML estimator suffers a non-negligible finite sample bias unless T is large. This problem is especially severe for the individual effects, whose estimation depends mainly on the T observations of each panel unit.

Second, methods like ML or non-linear least squares often fail to yield finite predictions for the individual effects. In short panels there are typically many individual units with identical outcomes in all time periods (either all zeros or ones). ML estimation will set the fitted probabilities to either zero or one, which requires the absolute value of the individual effect to go to infinity. The literature often describes this situation as 'perfect prediction' (see Maddala, 1983). In order to emphasize the distinction between prediction of fitted probabilities and those of individual effects, the main objective in this paper, we refer to such 'perfectly predicted' panel units instead as *concordant*.

The degree of concordance in typical applications can be very high. For instance, Autor et al. (2014) explain that they avoid using a non-linear binary response model in their panel analysis of US long-term disability policies specifically because 25 % of their sample is concordant and would be omitted in estimation. In our application to doctor visits with representative German survey data, between 29 and 45 % of the respective samples are concordant. Thus, for a sizeable share of individuals it would be impossible to obtain predictions of the individual effects using conventional FE binary response models, and units would necessarily be dropped from further descriptive and regression analyses involving the predicted individual effects. Such approaches discard informative variation stemming from what is often a rich set of covariates.

In this paper, we present and study an estimation approach that removes the first-order bias and obtains finite estimates for all individual effects, including the ones corresponding to concordant units. The concordance problem has, to the best of our knowledge, not been addressed directly by the previous literature on panel data models for binary responses, but it is closely related to well-known problems in the literature. Under asymptotic theories where T is considered fixed, the issue of finite sample bias becomes an identification problem. For instance, Honoré and Tamer (2006) and Browning and Carro (2014) studied the lack of point identification for dynamic discrete choice models and binary choice models with heterogeneous slopes, respectively. The concordant units, where the ML estimate of the individual effects is infinite, show the identification problem in its highest degree.

The identification problem is not unique to the ML estimator. For instance, for a sample consisting only of concordant units, the linear probability model would predict all individual effects as either zero or one, and estimate the structural parameters (the vector β, representing the slopes of all covariates) as zero. Moreover, the linear probability model also does not provide a consistent estimate of the average marginal effect, which is an average over the distribution of individual effects (Chernozhukov et al., 2013). A way of partially circumventing this problem in non-linear models is via conditional

maximum likelihood, which provides a consistent estimator of β. Such an estimator exists only for the FE logit model (or slightly modified FE logit models, see Bartolucci & Nigro, 2010) but not the FE probit model. In any case, the conditional approach is of no help here as it does not provide estimates of the individual effects.

Therefore, to obtain predictions of the individual effects we build on the literature that abandons the fixed-T assumption in favour of the alternative, where both N and T increase (Hahn & Newey, 2004, Arellano & Hahn, 2016). In this literature, consistent estimators are obtained by having N not grow too quickly relative to T and applying analytical or jackknife bias corrections of the first-order bias in the structural parameters such as β (e.g. Bester & Hansen, 2009, Fernández-Val, 2009, Bartolucci et al., 2016, Carro, 2007) or functionals of the individual effects such as average marginal effects (e.g. Fernández-Val, 2009, Dhaene & Jochmans, 2015). We contribute to this literature by focussing directly on the estimates of the individual effects. While a number of estimators remove the first-order bias in panel probit models and could potentially be used to improve over ML in terms of predicting individual effects, many fall short of the important practical issue of providing finite estimates for concordant units. For instance, all approaches that bias-correct the ML estimate fall into this categories, as the corrections cannot be applied to an infinite estimate.

We obtain the first-order bias of the maximum likelihood estimator for the individual effects, analytically under simplifying assumptions, as well as in simulation experiments, in the context of the FE panel probit model for binary outcomes. In order to remove the first-order, $O(1/T)$ bias, we follow Kosmidis and Firth (2009), who derive a general modified score function for models in the linear exponential family. Their approach can be tailored to the FE panel probit model. The resulting estimator is referred to as 'bias-reduced FE (BRFE) probit'. Importantly, we show that this estimator has finite support for all model parameters, including the individual effects of concordant observation units, by shrinking the individual effects towards zero. Shrinkage also reduces the variance of the predictions and unambiguously improves the mean squared error. Relatedly, all observations contribute to estimation of the structural parameters. In addition, the estimator can be easily implemented by using iteratively reweighted least squares, and it allows for straightforward inference using standard methods.

The identification problem and bias in finite samples can be considered two sides of the same coin. In this sense, there are limitations on what is achievable with the approach we present. A first limitation is of a theoretical nature. The ML estimator provides poor estimates of the individual effects in short panels. The BRFE improves over these estimates because it reduces the order of the asymptotic bias of the ML estimator and because it improves the finite sample performance. However, BRFE's ability to estimate individual effects is also negatively affected by the likelihood of concordance, although less extremely than ML estimation. As we show, the estimation of individual effects which are very large in absolute value (i.e. very likely to result in a concordant unit) is associated with a larger mean square error. A second limitation is of a more practical nature. The ML prediction for individual effects of concordant units, $\pm\infty$, is arbitrary and unreasonable for many practical purposes. The prediction given by BRFE in a model with no covariates is finite, but also arbitrary. The key advantage of BRFE over ML here lies in the BRFE prediction varying in sensible and intuitive ways with covariates and with the number of periods. For instance, BRFE will assign a lower probability (smaller individual effect) to a unit where the dependent variable is zero in eight periods versus four periods. Similarly, it will adjust the prediction of the individual effect according to the individual's covariates and corresponding β coefficients. The ML estimator, in contrast, remains unhelpfully infinite across these informative differences in the data.

In the next section, we formally introduce the problems of first-order bias and concordance in the context of binary response FE panel data models. We focus in our discussion on the FE probit

model where not even a conditional maximum likelihood solution exists. We then present the BRFE estimator, and show how it addresses these problems. In the Appendix, we show how the approach generalizes to other types of binary response models, including the FE logit model. In Section 3, we set up simulation experiments for predicting the individual effects across a number of differently shaped distributions from which the true individual effects are drawn. The BRFE estimator performs remarkably well in these simulations in terms of bias as well as mean squared error. The simulations also indicate that the BRFE estimator delivers reliable estimates of β in short panels. Overall, there is a clear recommendation for the use of the BRFE estimator in applications, in particular when there is a high degree of concordance.

In Section 4, we present an illustrative application related to health care utilization. Using panel data from the German Socio-Economic Panel for the period 2000–2014, we obtain predictions of the individual-specific effects in a model where the binary variable 'any doctor visit during the last 3 months (yes/no)' is regressed on a number of indicators of socio-economic status and health status. Estimating individual effects for separate subperiods, for example, 2000–2004 and 2010–2014, makes it possible to study their stability over time. We show that the percentile rank of the estimated individual effects for the first-period doctor visits are also predictive for a different outcome, the incidence of hospital admission 10 years later. Thus, the individuals' unobserved component is not only very persistent over time but also across different domains of health care utilization. Section 5 concludes.

2 | ECONOMETRIC METHODS

2.1 | Maximum likelihood estimation of the fixed effects probit model

Consider a panel probit model with individual-specific intercepts, α_i,

$$\Pr(y_{it} = 1 \,|\, \alpha_i, x_{it}) = \Phi(\alpha_i + x_{it}'\beta), \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{1}$$

where $y_{it} \in \{0, 1\}$, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, x_{it} is a vector of covariates and β a conformable vector of coefficients. The key parameters α_i are also referred to as individual or panel unit effects. N is assumed to be large and T is small. Observations are independent between individuals and, conditional on individual effects α_i and covariates x_{it}, serially uncorrelated. Regressors are strictly exogenous: $\Pr(y_{it} = 1 \,|\, x_{i1}, \ldots, x_{iT}, \alpha_i) = \Pr(y_{it} = 1 \,|\, x_{it}, \alpha_i)$. This model is prominent among practitioners, since it does not make any assumption on the distribution of the individual effects other than that they are finite, nor does it require the α_i's to be exogenous (uncorrelated with x_{it}). Estimators that do not place any restrictions on the distribution of individual effects are called FE estimators.

It is well-known that the maximum likelihood (ML) FE estimator, $(\widehat{\alpha}, \widehat{\beta}) = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_N, \widehat{\beta})$ has a number of deficiencies in this case. First, $\widehat{\beta}$ is inconsistent. This is a consequence of the incidental parameters problem. Abrevaya (1997) shows for the panel logit model with $T = 2$, that $\text{plim}\widehat{\beta} = 2\beta$. Greene (2004) provides Monte Carlo simulation results for the probit model showing that the upward bias persists for $T = 8$ and even $T = 20$. Second, $\widehat{\alpha}_i$ is inconsistent for fixed T and $N \to \infty$, and biased for small T. Third, $\widehat{\alpha}_i$ does not exist due to concordance, if $\bar{y}_i = 0$ or if $\bar{y}_i = 1$, where $\bar{y}_i = T^{-1} \sum_{t=1}^{T} y_{it}$.

We are here mostly concerned with the second and third issues, the small sample bias and the potential non-existence of $\widehat{\alpha}_i$. Our approach is based on an estimator developed by Kosmidis and Firth (2009) (see also Firth, 1993) for cross-section data and adapts it to the estimation of individual effects in a FE probit panel data model. We show that the resulting estimator does not suffer from the

concordance problem. It also is relatively easy to compute, as it can be obtained using an iteratively reweighted least squares estimator as shown by Kosmidis and Firth (2009).

2.2 | First-order bias

Non-linear maximum likelihood estimators have a finite sample bias. A formal derivation of the first-order bias of ML estimators is given in Cox and Snell (1971). For an illustration, consider a simple panel probit model with time-invariant regressors:

$$\Pr(y_{it} = 1 \mid \tilde{\alpha}_i, \overline{x}_i) = \Phi(\tilde{\alpha}_i + \overline{x}_i'\gamma) = \Phi(\alpha_i) \tag{2}$$

where $\alpha_i = \tilde{\alpha}_i + \overline{x}_i'\gamma$. In this case, $\overline{y}_i$ is the unbiased ML estimator for $\mu_i = \Phi(\alpha_i)$ and $\widehat{\alpha}_i^{ML} = \Phi^{-1}(\overline{y}_i)$. Expanding $\widehat{\alpha}_i^{ML}$ around α_i gives

$$\widehat{\alpha}_i^{ML} - \alpha_i \approx \frac{\partial \alpha_i(\mu_i)}{\partial \mu_i}(\overline{y}_i - \mu_i) + \frac{1}{2}\frac{\partial^2 \alpha_i(\mu_i)}{(\partial \mu_i)^2}(\overline{y}_i - \mu_i)^2$$

where $\alpha_i(\mu_i) = \Phi^{-1}(\mu_i)$, and

$$\frac{\partial \alpha_i}{\partial \mu_i} = \frac{1}{\partial \Phi(\alpha_i)/\partial \alpha_i} = \frac{1}{\phi(\alpha_i)}, \quad \frac{\partial^2 \alpha_i}{\partial \mu_i^2} = -\frac{1}{\phi(\alpha_i)^2} \times [-\alpha_i \phi(\alpha_i)] \times \frac{\partial \alpha_i}{\partial \mu_i} = \frac{\alpha_i}{\phi(\alpha_i)^2}.$$

Since $\mathrm{E}[(\overline{y}_i - \mu_i)^2] = \mu_i(1-\mu_i)/T = \Phi(\alpha_i)(1-\Phi(\alpha_i))/T$, it follows that

$$\mathrm{E}(\widehat{\alpha}_i^{ML} - \alpha_i \mid \alpha_i) = \frac{1}{2T}\frac{\alpha_i \Phi(\alpha_i)(1-\Phi(\alpha_i))}{\phi(\alpha_i)^2} + O(T^{-2}). \tag{3}$$

The first-order bias is positive if $\alpha_i > 0$, and hence $\Phi(\alpha_i) > 0.5$. It is negative for $\alpha_i < 0$. As $|\alpha_i|$ goes to infinity, so does the product of Mills ratios $\Phi(\alpha_i)(1-\Phi(\alpha_i))/\phi^2(\alpha_i)$ and hence the bias, both absolute and relative. Note that $\alpha_i \to \infty$ means that $\Pr(y_{it} = 1 \mid \alpha_i) \to 1$, while for $\alpha_i \to -\infty$, $\Pr(y_{it} = 1 \mid \alpha_i) \to 0$. Thus, the first-order bias increases with the likelihood of concordance.

2.3 | Concordance

Concordance in the general model (1) means that the first-order conditions for the ML estimator do not have a finite solution due to the lack of variation in the binary outcome in the panel probit model. The *K*+*N* first-order conditions of the panel probit model are:

$$s^{ML}(\alpha_i) = \frac{\partial \log L}{\partial \alpha_i} = \sum_{t=1}^{T}(y_{it} - \Phi(\eta_{it}))\frac{\phi(\eta_{it})}{\Phi(\eta_{it})(1-\Phi(\eta_{it}))} = 0, \quad i = 1, \ldots, N, \tag{4}$$

$$s^{ML}(\beta_k) = \frac{\partial \log L}{\partial \beta_k} = \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \Phi(\eta_{it}))\frac{\phi(\eta_{it})}{\Phi(\eta_{it})(1-\Phi(\eta_{it}))}x_{k,it} = 0, \quad k = 1, \ldots, K, \tag{5}$$

where $\eta_{it} = \alpha_i + x'_{it}\beta$, and K is the number of regressors in x_{it}. Suppose that $\bar{y}_i = 0$ for some i. Then (4) simplifies to

$$\sum_{t=1}^{T} \frac{\phi(\eta_{it})}{1 - \Phi(\eta_{it})} = 0, \tag{6}$$

which does not have a solution since the inverse Mills ratio $\lambda_{it} = \phi(\eta_{it})/(1 - \Phi(\eta_{it})) > 0$ for finite η_{it}. Similarly, if $\bar{y}_i = 1$ for some i, (4) simplifies to

$$\sum_{t=1}^{T} \frac{\phi(\eta_{it})}{\Phi(\eta_{it})} = 0, \tag{7}$$

which does not have a solution either. In the first case, $\widehat{\alpha}_i$ will tend to minus infinity, while it will tend to plus infinity in the second.

Note that the estimator for β still exists. As long as there are some panel units with variation in y_{it} (i.e. some *discordant* units), β can be estimated using those observations, based on (5). Concordant observations do not contribute to the (concentrated) score, since

$$\lim_{\widehat{\alpha}_i^{ML}(\beta)\to-\infty} \left(- \sum_t \frac{\phi(\widehat{\alpha}_i^{ML}(\beta) + x'_{it}\beta)}{1 - \Phi(\widehat{\alpha}_i^{ML}(\beta) + x'_{it}\beta)} x_{it} \right) = 0 \text{ if } \bar{y}_i = 0,$$
$$\lim_{\widehat{\alpha}_i^{ML}(\beta)\to+\infty} \left(- \sum_t \frac{\phi(\widehat{\alpha}_i^{ML}(\beta) + x'_{it}\beta)}{\Phi(\widehat{\alpha}_i^{ML}(\beta) + x'_{it}\beta)} x_{it} \right) = 0 \text{ if } \bar{y}_i = 1.$$

The problem of concordance is most severe for short panels: as T increases, it becomes less and less likely to obtain panel units with $\bar{y}_i = 0$ or $\bar{y}_i = 1$, provided that $0 < \Pr(y_{it} = 1) < 1$. For example, in the simple time-invariant model (2),

$$\begin{aligned} \Pr(\bar{y}_i = 0|\alpha_i) + \Pr(\bar{y}_i = 1|\alpha_i) &= \Pr\left(\sum_{t=1}^{T} y_{it} = 0 \,\middle|\, \alpha_i \right) + \Pr\left(\sum_{t=1}^{T} y_{it} = T \,\middle|\, \alpha_i \right) \\ &= (1 - \Phi(\alpha_i))^T + \Phi(\alpha_i)^T, \end{aligned} \tag{8}$$

Hence, the probability of concordance decreases in T. For a given T, it has a minimum at $\alpha_i = 0$. A larger absolute value of α_i leads to both a larger first-order bias and a higher incidence of concordance. Discarding concordant units from an analysis of the distribution of the individual effects, α_i, might lead to flawed conclusions.

2.4 | A bias-reduced fixed effects (BRFE) probit estimator

Firth (1993) showed that for linear exponential family models with canonical link function, the first-order bias of the maximum likelihood estimator can be removed by maximizing a modified log-likelihood function that includes a term based on the log-determinant of the information matrix (see also Ehm, 1991). The probit model is a linear exponential family model with non-canonical link where such a modified likelihood function does not exist. However, Kosmidis and Firth (2009) derived a related adjustment to the score function that achieves the same first-order bias reduction for

general cross-sectional models. Such a re-centring of the estimating equations to eliminate bias at the assumed model is also one of the features of the approach to robustness based on the influence function (Hampel et al., 1986).

Applying and extending this approach to the FE panel probit model achieves three things: first, it resolves the incidental parameters problem, in the sense that the estimator of the structural parameter is asymptotically normal and centred at the truth, contingent on asymptotics where T grows faster than $N^{1/3}$ (Hahn & Newey, 2004). Sartori (2003) establishes the same sufficient condition for the absence of the incidental parameters problem in the context of general profile likelihood functions, that extend to the Firth correction in exponential families (see also Lunardon, 2018). Second, as the focus is here on estimation of α_i, removing its first-order bias has the obvious direct benefit of providing better estimates of these parameters of interest. Third, and equally important, it ensures that finite estimates of α_i exist even for concordant observation units, which makes it possible to study the whole distribution of individual effects rather than a selected subsample.

For the FE probit panel model, the adjusted score function with respect to α_i is given by

$$\begin{aligned} s^{BRFE}(\alpha_i) &= \sum_{t=1}^{T} \left[y_{it} - \Phi(\eta_{it}) - \frac{1}{2} h_{it}\eta_{it} \frac{\Phi(\eta_{it})(1-\Phi(\eta_{it}))}{\phi(\eta_{it})} \right] \frac{\phi(\eta_{it})}{\Phi(\eta_{it})(1-\Phi(\eta_{it}))} \\ &= s^{ML}(\alpha_i) - \sum_{t=1}^{T} \frac{1}{2} h_{it}\eta_{it}, \end{aligned} \tag{9}$$

where h_{it} are the it-th diagonal elements of the $NT \times NT$ projection matrix

$$H = W^{1/2} X (X'WX)^{-1} X' W^{1/2}, \tag{10}$$

with X the $NT \times (K + N)$ matrix of the K regressors and N individual-unit indicator vectors, and W is the $NT \times NT$ diagonal matrix with typical element $w_{it} = \phi(\eta_{it})^2 / [\Phi(\eta_{it})(1 - \Phi(\eta_{it}))]$.

To provide an intuition for the way that the ML score is re-centred in (9), consider again the bias of $\widehat{\alpha}_i^{ML}$ in a simple constant-only model. In that model, the term that is subtracted from the ML score in (9) is equal to (3), the bias of $\widehat{\alpha}_i^{ML}$, times the curvature of the ML score function. Since the bias in $\widehat{\alpha}_i^{ML}$ is a result of the combination of the ML score being both unbiased and non-linear, by re-centring the score by 'bias in parameter × curvature of score' BRFE produces an estimator for α_i which is precisely free of this (first-order) bias.

The β_k-terms of the score vector are adjusted accordingly as

$$s^{BRFE}(\beta) = s^{ML}(\beta) - \sum_{t=1}^{T} \frac{1}{2} h_{it}\eta_{it} x_{it}. \tag{11}$$

From (9) and (11), it can be seen that if we define

$$y_{it}^{*} = y_{it} - \frac{1}{2} h_{it}\eta_{it} \frac{\Phi(\eta_{it})(1 - \Phi(\eta_{it}))}{\phi(\eta_{it})}, \tag{12}$$

then (9) and (11) are in the form of the standard ML scores $s^{ML}(\alpha_i)$ and $s^{ML}(\beta)$, where y_{it} is replaced by the pseudo-response y_{it}^{*}.

The BRFE probit estimator $(\widehat{\beta}^{BRFE}, \widehat{\alpha}^{BRFE}) = (\widehat{\beta}^{BRFE}, \widehat{\alpha}_1^{BRFE}, \ldots, \widehat{\alpha}_N^{BRFE})$ is obtained by jointly solving the $N+K$ first-order conditions $s^{BRFE}(\alpha_i) = 0$ (i=1, …, N) and $s^{BRFE}(\beta) = 0$. In practice, this can be

done using iteratively reweighted least squares common for generalized linear models (as in Kosmidis & Firth, 2009), or Newton–Raphson type pseudo-ML estimators for the probit model, where at iteration s, the vector $\{\widehat{y}_{it}^{*}\}$ is computed from the values $\widehat{h}_{it}(\widehat{\beta}^{s-1}, \widehat{\alpha}^{s-1})$ and $\widehat{\eta}_{it}(\widehat{\beta}^{s-1}, \widehat{\alpha}^{s-1})$ estimated at the previous iteration.

The resulting estimator follows an asymptotically normal distribution with an asymptotic variance equal to the inverse of the Fisher information; when estimates are obtained via iteratively reweighted least squares, estimated standard errors can be conveniently extracted from the square root of the diagonal elements in $(X'\widehat{W}X)^{-1}$, where $\widehat{W}$ is W evaluated at the estimates of the final iteration (Kosmidis, 2007).

2.4.1 | Obtaining finite individual effects for all units

To prove that the BRFE probit estimator of the individual effects is always finite, consider the case where all observations of unit i are equal to one, $\bar{y}_i = 1$. Then, we can write the bias-reduced score (9) as

$$s^{BRFE}(\alpha_i) = \left(\sum_t \frac{\phi(\eta_{it})}{\Phi(\eta_{it})} \right) - \frac{\alpha_i}{2} \left(\sum_t h_{it} \right) - \frac{1}{2} \left(\sum_t h_{it} x_{it}' \beta \right) = g_1(\alpha_i) - \alpha_i g_2(\alpha_i) - g_3(\alpha_i). \quad (13)$$

When α_i becomes large, the first term in the score, $g_1(\alpha_i)$, approaches zero, because each inverse Mills ratio in the sum approaches zero. Because h_{it} is an element of the diagonal of a projection matrix, we have that $0 < h_{it} \leq 1$ for each h_{it}, so that $g_2(\alpha_i)$ is bounded. Thus, as α_i tends to plus infinity, the second term, $-\alpha_i g_2(\alpha_i)$, tends to minus infinity. The third term, $g_3(\alpha_i)$, tends to some finite constant because it is a sum of T finite summands. Thus, the whole score tends to minus infinity when α_i tends to plus infinity. When α_i tends to minus infinity, $g_1(\alpha_i)$ grows without bound, and so does $-\alpha_i g_2(\alpha_i)$, while $g_3(\alpha_i)$ tends to some other finite constant. Thus, the whole score tends to plus infinity. Since the score is continuous, a finite solution must exist. Similar arguments can be made to show that a solution exists for the other concordance case, $\bar{y}_i = 0$, as well.

Without time varying regressors, that is, in the constants-only model, $\sum_t h_{it} = 1$. From (9), we see that if $y_{i1} = \ldots = y_{iT} = 1$, $\widehat{\alpha}_i$ is obtained as solution to the non-linear equation

$$\widehat{\alpha}_i^{BRFE} = 2T \frac{\phi(\widehat{\alpha}_i^{BRFE})}{\Phi(\widehat{\alpha}_i^{BRFE})}, \quad (14)$$

whereas for $y_{i1} = \ldots = y_{iT} = 0$,

$$\widehat{\alpha}_i^{BRFE} = -2T \frac{\phi(\widehat{\alpha}_i^{BRFE})}{1 - \Phi(\widehat{\alpha}_i^{BRFE})}. \quad (15)$$

The two cases differ only in sign. For $T = 2, 4, 8, 12$ the estimates for α_i are about ± 1.06, ± 1.37, ± 1.67, and ± 1.84 respectively. The solutions for Equations (14) and (15) may be approximated by $\widehat{\alpha}_i^{BRFE} \approx 0.8 + 0.413 \log T_i$ and $\widehat{\alpha}_i^{BRFE} \approx 0.007 + 0.293/T$, respectively.

When $\bar{y}_i = 0$, the associated predicted probabilities $\widehat{\Pr(y_{it}=0)} = 1 - \Phi(\hat{\alpha}_i)$ are 0.144, 0.086, 0.048 and 0.033. These values illustrate how the BRFE estimator bounds the predicted probabilities away from the perfect-prediction probability of 0. By how much probabilities are bounded away from zero is inversely related to the number of individual observations T. The ML solution, of course, is a probability of exactly zero, which, while unbiased, might be an unreasonable prediction for many applications: it means that an event that has not occurred in two, three or four periods is deemed impossible. Note that in models with time-varying covariates x_{it}, the ML solution remains an exact zero irrespective of the value of the covariates. In contrast, the BRFE estimator adjusts the predicted probabilities according to the value of x_{it} allowing distinctions between different underlying propensities.

The problem of non-existence of the ML estimator for the individual effect is not limited to the probit model. All binary response models of the general form $P(y_{it} = 1 \,|\, x_{it}, \alpha_i) = G(x_{it}'\beta + \alpha_i)$, where $G(\cdot)$ is a smooth strictly increasing function, suffer from it. As we show in the appendix, the BRFE estimator of α_i, in contrast, is guaranteed to obtain finite estimates of the individual effects for all units for this general class of models (Appendix B.3), which includes all the commonly used models in the literature, such as logit, complementary log–log, Weibull, etc. Of particular interest is logit, which is the canonical parametrization for generalized linear models with a binary response variable. For this special case, the bias-reduced estimator has a penalized likelihood representation (Firth, 1993).

Also note that the problem of non-existing ML estimates can arise in cross-sections as well. An example is a model for the probability of treatment with a binary instrument Z and no 'always takers', that is, when Z=0 implies non-treatment. Heinze and Schemper (2002) have shown for the logit model that Firth's (1993) modified likelihood estimator provides finite estimates in such a case.

2.4.2 | Mean squared error

After having established the finiteness of the BRFE estimator for α_i, we now briefly consider the quality of the estimation. For the simple model without covariates, the mean square error (MSE) of the estimated individual effects in the BRFE probit model can be written as

$$\begin{aligned}\mathrm{MSE}(\hat{\alpha}_i^{BRFE}, \alpha_i) &= \mathrm{E}_{\bar{y}_i}(\hat{\alpha}_i^{BRFE} - \alpha_i)^2 \\ &= \sum_{k=0}^{T} \Pr(\bar{y}_i = k/T)(\hat{\alpha}_i^{BRFE}(k/T) - \alpha_i)^2 \\ &= \sum_{k=0}^{T} \frac{T!}{k!(T-k)!}\Phi(\alpha_i)^k(1-\Phi(\alpha_i))^{T-k}(\hat{\alpha}_i^{BRFE}(k/T) - \alpha_i)^2, \end{aligned} \tag{16}$$

The first and last terms of the sum represent concordant observations. Hence, the MSE of the ML estimator $\hat{\alpha}_i^{ML}$ cannot be finite. With first-order bias reduction, the MSE of $\hat{\alpha}_i^{BRFE}$ is always finite.

Figure 1 visualizes the relationship between the MSE of $\hat{\alpha}_i^{BRFE}$ and the severity of concordance. The graph plots the MSE of $\hat{\alpha}_i^{BRFE}$ against the share (or probability) of concordant units for $T = 4$ and $T = 12$ respectively. For example, if $\alpha_i = 0$, it follows that $\Phi(0) = 0.5$, and for $T = 4$ and $N \to \infty$, $2 \times 0.5^4 = 12.5\,\%$ of all observations can be expected to be concordant (see also Equation (8)). Other values for α_i necessarily lead to a higher degree of concordance.

In these graphs, the MSE is decomposed into a part attributable to concordant observations (the summands corresponding to $k = 0$ and $k = T$ on the right-hand side of Equation (16)) and a part attributable to discordant observations (the summands corresponding to $0 < k < T$). The relative

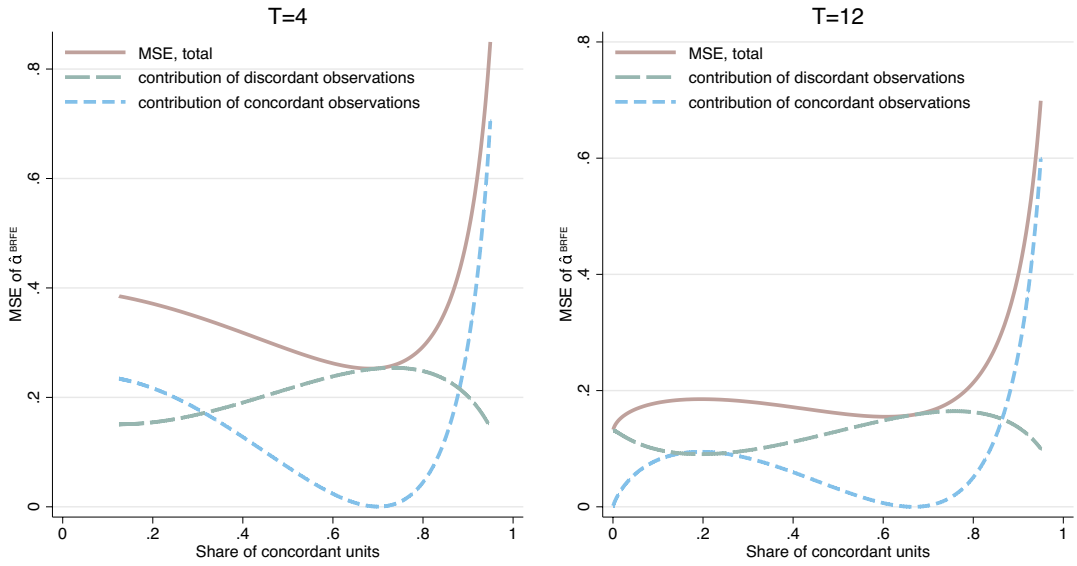

FIGURE 1 Share of concordant units and MSE in BRFE probit models without covariates *Notes*: The graphs in the figure plot MSE($\hat{\alpha}_i^{BRFE}$) against the share of concordant units, $P(\bar{y}_i = 0) + P(\bar{y}_i = 1)$, in panel probit models without covariates, $P(y_{it} = 1) = \Phi(\alpha_i)$. The decomposition of MSE into contributions from concordant and discordant observations refers to Equation (16), where the concordant contribution is equal to the sum of the first ($t = 0$) and last ($t = T$) addends of the right-hand side of Equation (16), and the discordant contribution is equal to the remainder.

contribution of concordant observations to the total MSE first decreases and then increases, as concordance becomes more prevalent. Eventually, the MSE is almost exclusively driven by concordant observations and becomes large. But we know from results in section 2.4.1 that it remains finite. No closed-form MSE results are available once covariates are introduced into the model. We therefore conduct a number of simulation experiments where individual effects are obtained from a number of different distributions

3 | SIMULATION EXPERIMENTS

3.1 | Set-up

The focus of this simulation study is the performance of the BRFE approach for estimating individual effects in fixed effects panel probit models with a small to moderate number of time periods and a varying prevalence of concordance. The simulation experiments contrast the behaviours of the ML and BRFE estimators for different true underlying distributions of the individual effects, in terms of bias and mean squared error.

In our simulations, the time-invariant individual effects α_i are drawn from four alternative distributions: uniform, beta, Gaussian and Bernoulli, as plotted in Figure 2. The distributions have been rescaled and shifted to make them more comparable. All distributions have a mean of zero, or close to zero, and all, or most, of their probability mass lies within the interval [-1,1]. The distributions vary starkly, however, in their shape. The data generating processes (DGPs) correspond to a 'random effects' model as the distribution of α_i does not depend on the regressor. This is the simplest type of DGP and it allows us to focus on biases purely related to small samples and the concordance problem.

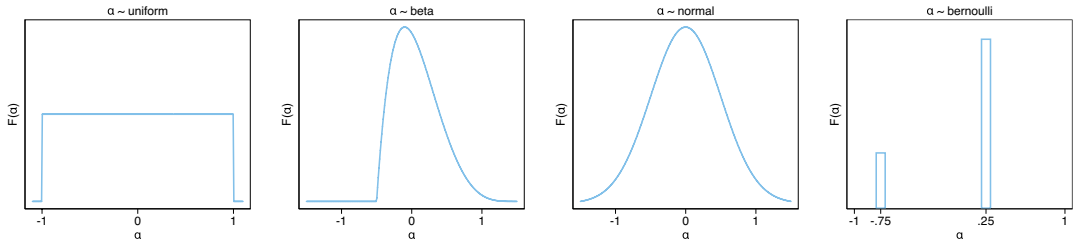

FIGURE 2 Distributions of α_i

Notes: Distributions from which α_i were drawn for the Monte Carlo simulation: 'uniform' corresponds to a uniform distribution on the interval [−1,1]; 'beta', to a Beta distribution with shape parameters 2 and 5, rescaled to the interval [−1,1] by multiplying the variable by 2 and subtracting 0.5; 'bernoulli', to a modified Bernoulli distribution taking the value −0.75 with probability 0.25, and the value 0.25 with probability 0.75; and 'normal', to a Normal distribution with mean 0 and variance 0.5.

While additional dependence on regressors can potentially exacerbate or attenuate those biases, we found that our conclusions are robust to introducing positive and negative correlation between α_i and x_{it} (and some results from such DGPs are presented in the Appendix, Figure A1).

Below, we report simulation results for $N = 100$ and $T \in \{2, 4, 8, 12\}$. For each of the four distributions from Figure 2, we draw one hundred values of α_i first, and keep them fixed through all Monte Carlo replications. There is a single regressor, x_{it}, which is drawn from a uniform distribution with support [−1,1]. Again, this is done once for each T and kept fixed over replications. Finally, the binary dependent variables y_{it} are obtained as

$$y_{it}^{(r)} = \mathbb{1}(\alpha_i + \beta x_{it} + \varepsilon_{it}^{(r)} > 0), \quad i = 1, \ldots, 100 \quad t = 1, \ldots, T,$$

where $\varepsilon_{it}^{(r)}$ has a standard normal distribution, $\beta = 1$, and r=1, …, 500 denotes Monte Carlo replications.

In each of the 500 replications, we keep track of the fraction of concordant, or perfectly predicted, observations, that is, the fraction of cross-sectional units for which $\bar{y}_i^{(r)} = 0$ or $\bar{y}_i^{(r)} = 1$. For instance, with $T = 4$ and a uniformly distributed α_i, the average fraction of concordant observations over the 500 replications amounts to 24 %. This fraction is somewhat lower for the beta (15 %) and Bernoulli (20 %) distributions, and higher for the normal distribution (28 %). Plots and summary statistics of our results are based on all finite estimates: since the maximum likelihood estimator of α_i does not exist for concordant observations, the effective replication sample size is below 500 in these cases.

Our choice of simulation DGP is a special case of the DGP introduced by Heckman (1991) and widely adopted in the bias correction literature for discrete choice models (e.g. Hahn & Newey, 2004, Fernández-Val; 2009, and others). There are two differences regarding the original DGP. First, unlike in the original DGP, x_{it} is not serially correlated in our DGP, but we specify a larger support for the regressor's uniform distribution ([−1,1] instead of the original [−0.5,0.5]). Second, Heckman's DGP only considers $\alpha_i \sim N(0, 1)$, where we consider several distributions, albeit with a lower baseline variance. Additional simulations, discussed below, show results for increasing the variance in our baseline (beyond Heckman's DGP) as well as results for Heckman's DGP.

3.2 | Results for quantities other than $\widehat{\alpha}_i$

While our interest is obtaining individual predictions of $\widehat{\alpha}_i$, we first show results for the estimated β as well as for marginal or partial effects of x_{it}. Table 1 presents means and standard deviations

TABLE 1 MC simulation: mean and standard deviation [SD] of $\widehat{\beta}$ ($\beta = 1500$ replications)

	$T = 2$		$T = 4$		$T = 8$		$T = 12$	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
$\alpha_i \sim$ *Bernoulli*								
ML	2.105	0.673	1.400	0.256	1.154	0.122	1.092	0.089
BRFE	0.953	0.240	1.006	0.169	1.007	0.103	1.002	0.080
$\alpha_i \sim$ *Uniform*								
ML	2.206	0.747	1.427	0.272	1.163	0.122	1.098	0.091
BRFE	0.928	0.242	0.997	0.173	1.005	0.102	1.004	0.082
$\alpha_i \sim$ *Beta*								
ML	2.075	0.716	1.364	0.231	1.143	0.125	1.084	0.086
BRFE	0.942	0.268	1.013	0.159	1.004	0.107	0.999	0.078
$\alpha_i \sim$ *Normal*								
ML	2.195	0.990	1.410	0.263	1.163	0.126	1.103	0.090
BRFE	0.889	0.250	0.977	0.165	0.997	0.105	1.001	0.080

Notes: Cells contain the average and standard deviation, over 500 replications, of the estimated β for each of the two estimators, ML and BRFE. The true value of β is 1.

of the estimated $\widehat{\beta}^{ML}$ and $\widehat{\beta}^{BRFE}$ across different distributions of α_i and different numbers of time periods. The true value is 1. The corresponding entries in the table confirm that the ML estimator for the common parameter suffers from incidental parameters bias. The bias is sizeable regardless of the distribution of α_i, and it amounts to about 110, 40, 15 and 10 % for T equal to 2, 4, 8 and 12 respectively. In contrast, BRFE effectively removes much of the bias in $\widehat{\beta}$. Even for $T = 2$, only a bias of about -10 % is left. At $T = 4$ the bias falls to between 0.6 % (Bernoulli) and 2.3 % (normal).

One of the differences between BRFE and ML estimation is that concordant observations contribute to estimation of β in the former but not in the latter case. This affects the relative precision of the two estimators. For $T = 2$, where the degree of concordance is most severe, the standard deviation of $\widehat{\beta}^{BRFE}$ is much smaller than that of $\widehat{\beta}^{ML}$, by a factor varying between 2.5 and almost 4, depending on data generating process. For $T = 12$, the difference is reduced to around 10 %.

Table 2 shows results for marginal effects and predicted probabilities. The table entries contain ratios of the average estimates to the true value. In line with other studies (e.g. Alexander & Breunig, 2016, and references therein), the ML estimates of the average marginal effect are less biased than the β coefficient. In contrast, the BRFE estimates are more biased than their corresponding β estimates. (This is a consequence of the bias reduction which corrects for bias in the coefficients, introducing bias in any non-linear transformation of them). However, overall, BRFE still is substantially less biased than ML for short panels. For $T = 2$ and $T = 4$, the average marginal effects estimated by ML have biases between 30–50 % (BR: 15–20 %) and 15–30 % (BR: around 10 %). For longer panels, the differences are smaller, and it appears that ML converges to the true quantities at a somewhat faster rate. The table also shows the marginal effect at the median, for which the ML estimates display generally larger biases than for the average marginal effect. BRFE estimates, on the contrary, tend to be only minimally worse. Finally, the table also shows predicted probabilities. For these, biases seem to be substantially smaller overall and BRFE estimates consistently outperform ML estimates.

TABLE 2 MC simulation: ratios of estimated to true marginal effects and predicted probabilities, N=100, 500 replications

	$T = 2$				$T = 4$				$T = 8$				$T = 12$			
	Pred.prob.		**Marg.eff.**		**Pred.prob.**		**Marg.eff.**		**Pred.prob.**		**Marg.eff.**		**Pred.prob.**		**Marg.eff.**	
	Avg	**Med**	**Avg**	**Med**	**Avg**	**Med**	**Avg**	**Med**	**Avg**	**Med**	**Avg**	**Med**	**Avg**	**Med**	**Avg**	**Med**
$\alpha_i \sim$ *Bernoulli*																
ML	1.045	1.061	1.385	1.811	1.009	1.008	1.166	1.259	1.014	1.015	1.032	1.047	1.003	1.003	1.005	1.011
BRFE	1.002	1.000	0.831	0.804	1.001	1.000	0.899	0.881	1.000	0.999	0.940	0.928	0.998	0.998	0.955	0.946
$\alpha_i \sim$ *Uniform*																
ML	1.060	1.082	1.464	1.983	1.003	1.000	1.216	1.333	1.003	1.003	1.055	1.077	0.998	0.997	1.014	1.021
BRFE	1.001	1.002	0.842	0.823	1.004	1.005	0.906	0.893	1.000	1.001	0.946	0.939	0.999	0.999	0.960	0.954
$\alpha_i \sim$ *Beta*																
ML	0.937	0.917	1.311	1.632	0.971	0.966	1.108	1.172	0.991	0.990	1.014	1.024	0.997	0.997	0.999	1.003
BRFE	0.985	0.982	0.803	0.769	0.994	0.992	0.890	0.866	0.997	0.996	0.934	0.917	0.998	0.997	0.951	0.939
$\alpha_i \sim$ *Normal*																
ML	0.935	0.918	1.510	1.993	0.971	0.967	1.272	1.391	0.987	0.986	1.101	1.130	1.001	1.001	1.050	1.063
BRFE	0.978	0.975	0.840	0.822	0.992	0.990	0.915	0.900	0.994	0.993	0.952	0.945	0.995	0.994	0.967	0.962

Notes: Cells contain the ratios of estimated quantities to the true quantity. Estimated quantities are averages over 500 replications. Quantities are the average marginal effect of x_{it} (Marg.eff., Avg) the marginal effect for the median x_{it}(Marg.eff., Med), the average predicted probability marginal of x_{it} (Pred.prob., Avg) and the predicted probability at the median x_{it} (Pred.prob., Med).

Since β coefficients and marginal effects have been the object of interest in the literature on bias correction, we can compare the performance of BRFE to that of other such estimators. In the Appendix (Table A1), we present results from the Heckman (1991) DGP for BRFE and ML, and contrast this to the results obtained for ML and six different bias correction estimators in this DGP by Alexander and Breunig (2016). These include three estimators proposed by Hahn and Newey (2004), as well as the estimators proposed by Fernández-Val (2009), Dhaene and Jochmans (2015), and Bartolucci et al. (2016). The results show that BRFE performs well within the range of these estimators from the literature in terms of estimating β and marginal effects. However, as mentioned previously, when the interest lies in obtaining predictions of α_i, BRFE has the advantage that it obtains finite $\widehat{\alpha}_i$ for all i including those of concordant units.

3.3 | Predicting individual effects $\widehat{\alpha}_i$

Before looking at the individual predictions of $\widehat{\alpha}_i$, we start by examining the estimation of overall features of the distribution of α_i. Table 3 lists, for each of the four distributions of α_i, the estimated mean, standard deviation, skewness and kurtosis of the $N = 100$ predictions, averaged over 500 replications. These can be benchmarked against the corresponding moments of the distribution of the (once) simulated α_i's, for instance −0.030, 0.451, −0.980 and 1.960 in the case of the Bernoulli data generating process.

For the longer time horizons, $T = 12$ and to a lesser extent also $T = 8$, there is not much of a difference between ML and BRFE. Both estimators are centred roughly at the true mean with the right variance. The third and fourth moment are likewise estimated quite accurately. For the shorter panels, the BRFE estimator performs much better than ML. For example, the mean of the ML predictions has a sign opposite to that of the true mean in seven of eight cases (for $T = 2$ and $T = 4$). Regarding variation, we find that for short panels (in particular for $T = 2$), the BRFE estimator underestimates the variance of the generated individual effects somewhat, reflecting the shrinkage property of the BRFE estimator. The ML estimator also displays a poor performance in estimating the skewness of the distribution with $T = 2$ where BRFE gives averages which are much closer to the true values.

We have seen in Table 3 that the BRFE estimates $\widehat{\alpha}_i^{BRFE}$ are approximately centred at the true mean $\overline{\alpha}_i$ for each of the four distributions (normal, Bernoulli, uniform and beta). However, from the point of view of estimating individual α_i's, this result is not very informative. For example, an upward bias for large α_i's and a downward bias for small α_i's might simply offset each other.

Figures 3, 5 and 6 therefore plot the mean prediction, over the $r = 500$ replications, separately for each true value of α_i. All results are based on a cross-section size of $N = 100$. Points on the 45-degree line indicate the absence of bias. Figure 4, showing the results from the Bernoulli DGP, is an exception. Here, α_i can only take two values, and we compare the entire empirical distribution function of $\widehat{\alpha}_i$ to the step function implied by the true Bernoulli DGP (with jumps at −0.75 and 0.25, see Figure 2).

Each figure has a 2×2 format, one for the estimation method (ML versus BRFE) and one for the time dimension ($T = 4$ versus $T = 8$).

Figure 3 shows results for the normally distributed individual effect. The scale on the y-axis indicates that the true draws of α_i varied from −2.0 to 1.6 in this case. Naturally, the density of observations increases as we move to the centre of the distribution. In order to compute the mean of the ML estimates, we had to drop all concordant observations which are more prevalent for α_i's located in the tail of the distribution. Given this necessary adjustment, and for $T = 4$, we find that the ML means display a substantial bias. The bias is positive for α_i' below zero, and negative for α_i's above zero, and

TABLE 3 MC Simulation: Estimates of $E(\alpha_i)$ [Mean], $SD(\alpha_i)$ [SD], Skewness of α [Skew] and Kurtosis of α [Kurt]; N=100, 500 replications

	$T = 2$				$T = 4$				$T = 8$				$T = 12$			
	Mean	**SD**	**Skew**	**Kurt**	**Mean**	**SD**	**Skew**	**Kurt**	**Mean**	**SD**	**Skew**	**Kurt**	**Mean**	**SD**	**Skew**	**Kurt**
$\alpha_i \sim$ *Bernoulli*																
True	−0.030	−0.451	−0.980	*1.960*												
ML	0.025	0.778	−0.201	2.643	−0.030	0.361	−0.921	2.937	−0.019	0.452	−0.993	2.048	−0.032	0.482	−0.980	1.970
BRFE	−0.018	0.353	−0.952	2.138	−0.023	0.420	−0.992	2.039	−0.031	0.446	−0.978	1.969	−0.034	0.453	−0.978	1.962
$\alpha_i \sim$ *Uniform*																
True	−0.045	*0.585*	−0.009	*1.827*												
ML	0.025	0.819	−0.203	2.689	−0.048	0.445	−0.151	2.391	−0.047	0.573	0.015	1.783	−0.053	0.620	−0.015	1.791
BRFE	−0.044	0.431	−0.072	1.895	−0.041	0.540	−0.008	1.787	−0.047	0.576	0.007	1.796	−0.050	0.588	−0.016	1.816
$\alpha_i \sim$ *Beta*																
True	*0.034*	*0.296*	*0.455*	*2.380*												
ML	−0.147	0.836	−0.041	2.373	−0.014	0.317	0.180	3.021	0.025	0.305	0.337	2.181	0.034	0.318	0.436	2.300
BRFE	0.007	0.229	0.439	2.716	0.027	0.290	0.432	2.428	0.032	0.295	0.441	2.296	0.033	0.296	0.461	2.349
$\alpha_i \sim$ *Normal*																
True	*0.045*	*0.733*	−0.303	*3.009*												
ML	−0.138	0.831	−0.049	2.213	−0.000	0.473	−0.066	2.909	0.028	0.638	−0.252	2.357	0.047	0.710	−0.147	2.336
BRFE	0.009	0.502	−0.340	2.782	0.038	0.629	−0.203	2.490	0.040	0.696	−0.293	2.708	0.043	0.715	−0.221	2.728

Notes: Rows labelled 'True' contain the (true) mean, standard deviation, skewness and kurtosis of the 100 drawn α_i for each of the four distributions (Bernoulli, uniform, beta, and normal). Cells in rows ML and BRFE contain the average, over 500 replications, of the mean, standard deviation, skewness and kurtosis of the estimated α_i for each of the two estimators.

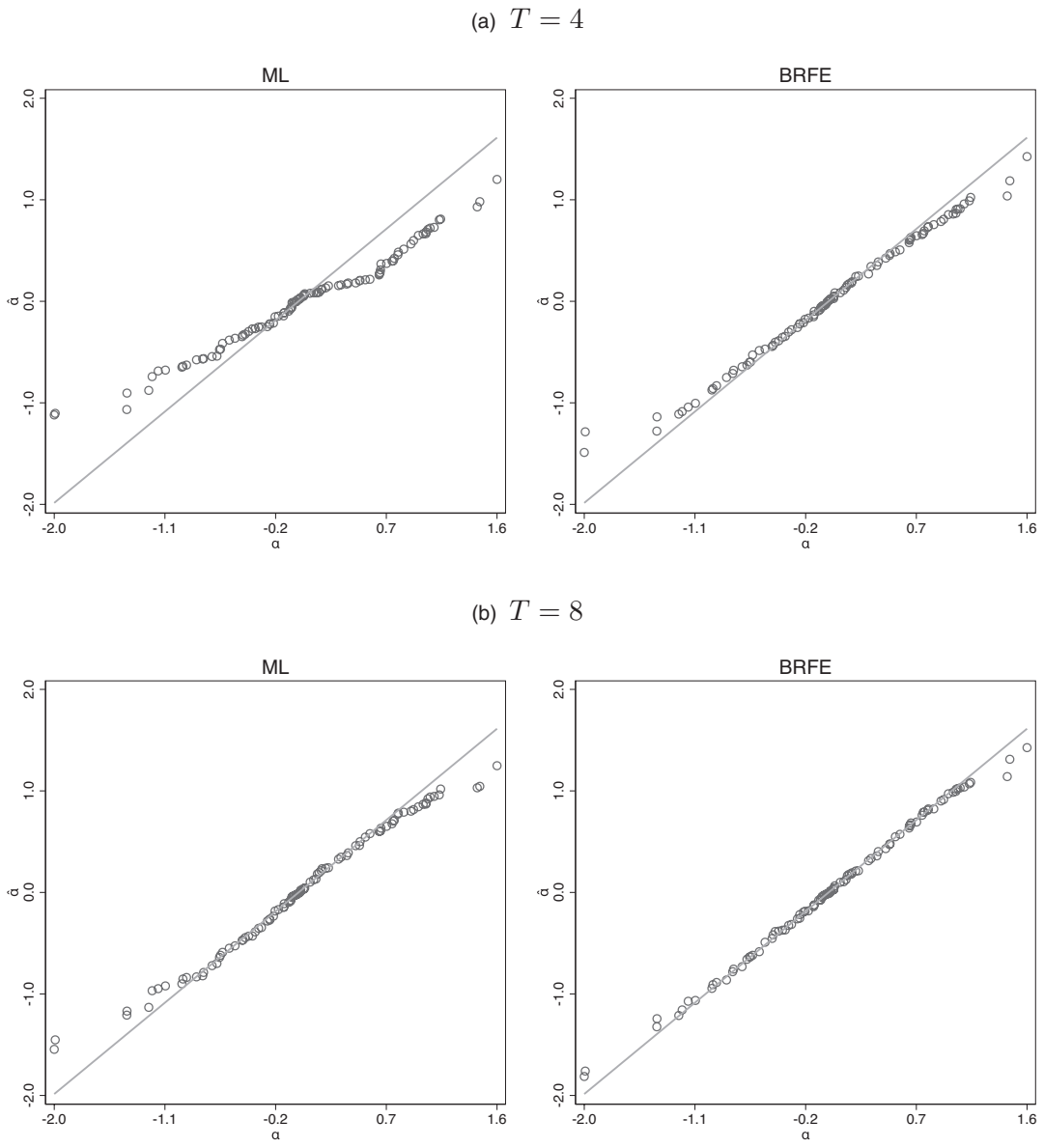

FIGURE 3 Mean of estimated individual effects by true α_i ($\alpha_i \sim$ *normal*)
Notes: Graphs show average estimates of α_1, …, α_{100} over 500 replications against their true values.

increasing, in absolute terms, as α_i's move towards the tails. Doubling the sample size from $T = 4$ to $T = 8$ improves the ML estimates, although bias remains in the tails.

The BRFE means are computed on the full sample, as they include values for concordant observations. They are close to the 45-degree line, even for the small panel, although some bias emerges for α_i's more than one standard deviation above or below its mean. Here, the shrinkage towards zero starts having a noticeable effect, with an upward bias being observed for large negative α_i's and a downward bias for large positive α_i's. The pattern is similar when the sample size is doubled, although the magnitude of the biases in the tails decreases. The fit in the centre of the distribution is even tighter.

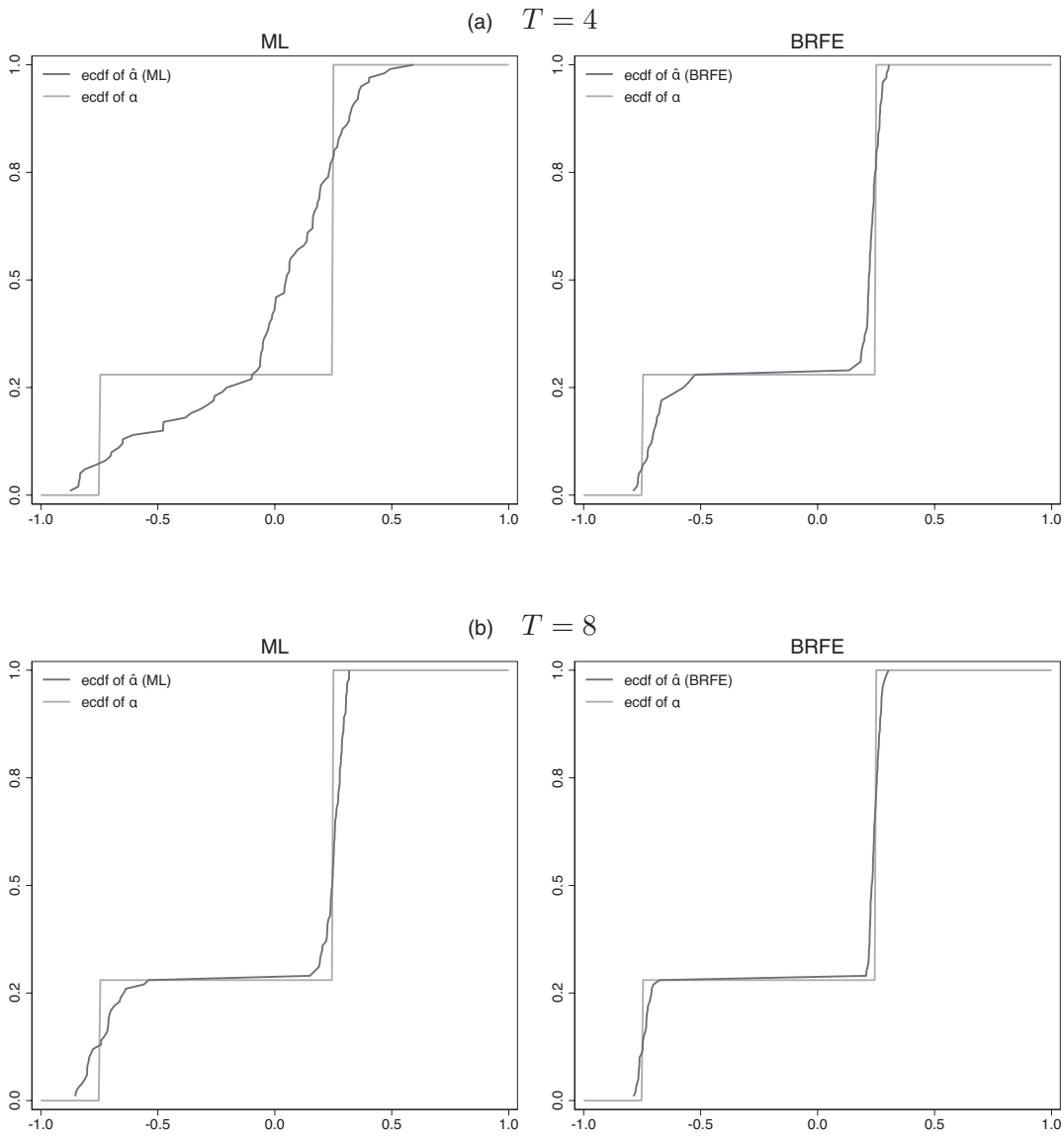

FIGURE 4 Estimated versus true distributions of α_i ($\alpha_i \sim$ *Bernoulli*)
Notes: Graphs show the empirical cdf of the hundred true α_i against the empirical cdf of the hundred average $\widehat{\alpha}_i$ estimated over 500 replications.

A comparison of the empirical distribution functions in Figure 4 shows a good recovery of the binary true distribution function through the BRFE estimator, with two separate clusters of estimates closely centred around the true values −0.75 and 0.25. The fit is already quite good for $T = 4$ and further improves for $T = 8$. By contrast, the ML estimates in the small sample are spread from 0.8 to 0.5 and 'almost uniformly', which would be the 45-degree line. The binary nature of the true α_i's only emerges in the longer panels.

Finally, Figures 5 and 6 are based on bounded but continuous distributions for the individual effect, a uniform(−1,1) distribution in case of Figure 5 and a scaled beta distribution in case of Figure 6, the latter with values between −0.5 and 0.7. In the context of the probit model, this means that

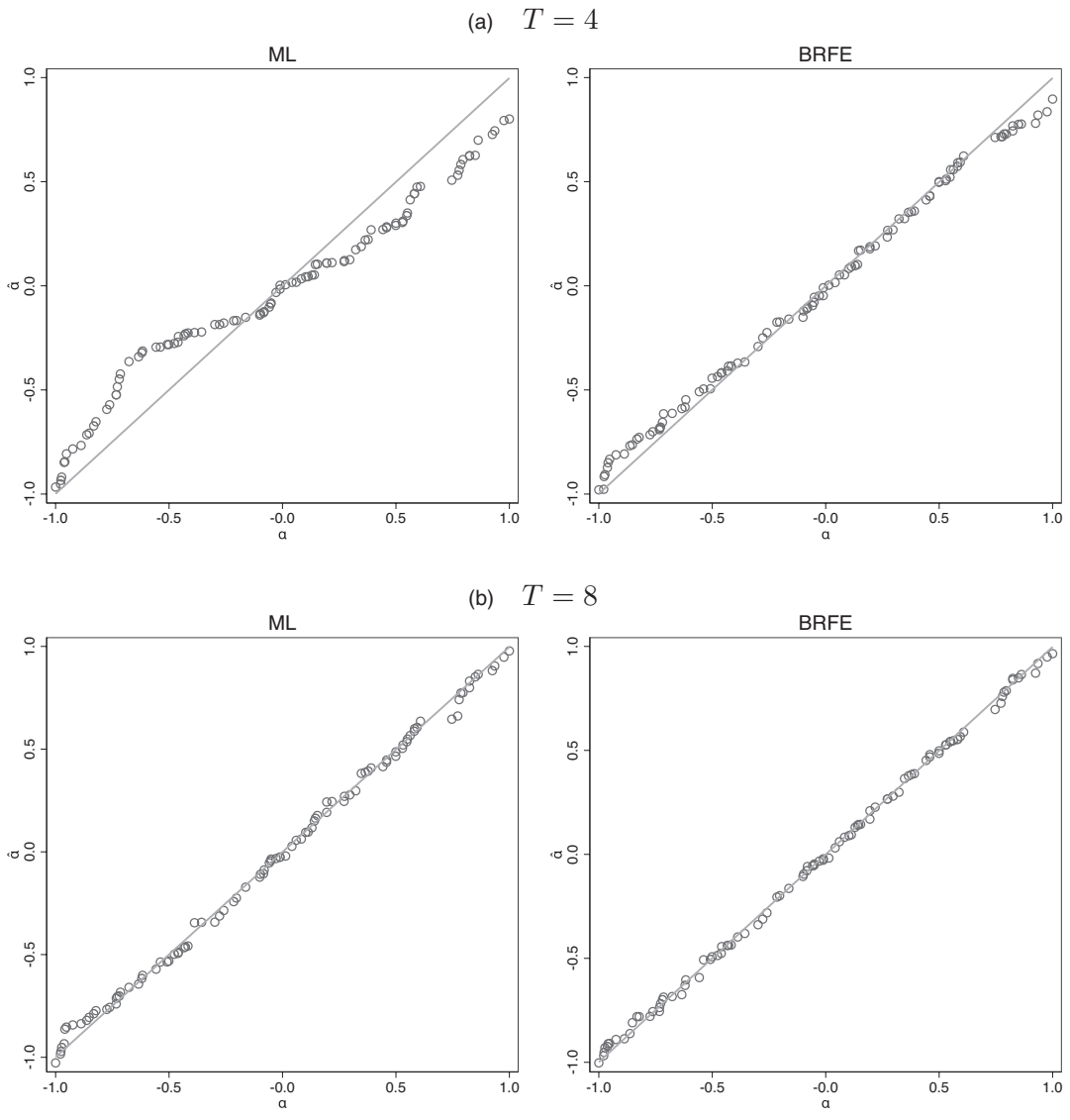

FIGURE 5 Mean of estimated individual effects by true α_i ($\alpha_i \sim$ *uniform*)
Notes: Graphs show average estimates of $\alpha_1, \ldots, \alpha_{100}$ over 500 replications against their true values.

the probability of a success is bounded, as well, for the average x_{it} between $\Phi(-0.5) = 30.9\%$ and $\Phi(0.7) = 75.8\%$ for the beta DGP. As a consequence, the beta simulations produce less concordance (overall 15 %) compared to the normal simulations (28 %), and avoiding the tails improves the fit substantially. Although to a lesser degree, this is true also for the uniform DGP. The main message from these simulation results is that the BRFE method gives approximately unbiased estimates of the individual effects regardless of sample size, whereas the ML method is unreliable for $T = 4$, while it works satisfactorily for the larger sample, where estimated means are mostly lined up along the 45-degree line. These conclusions also hold in similar DGPs but with correlation between the indvidual effects and the covariate, as illustrated in the Appendix for the case of $T = 4$, $\alpha_i \sim$ normal, in DGPs with both positive and negative correlation between x_{it} and α_i.

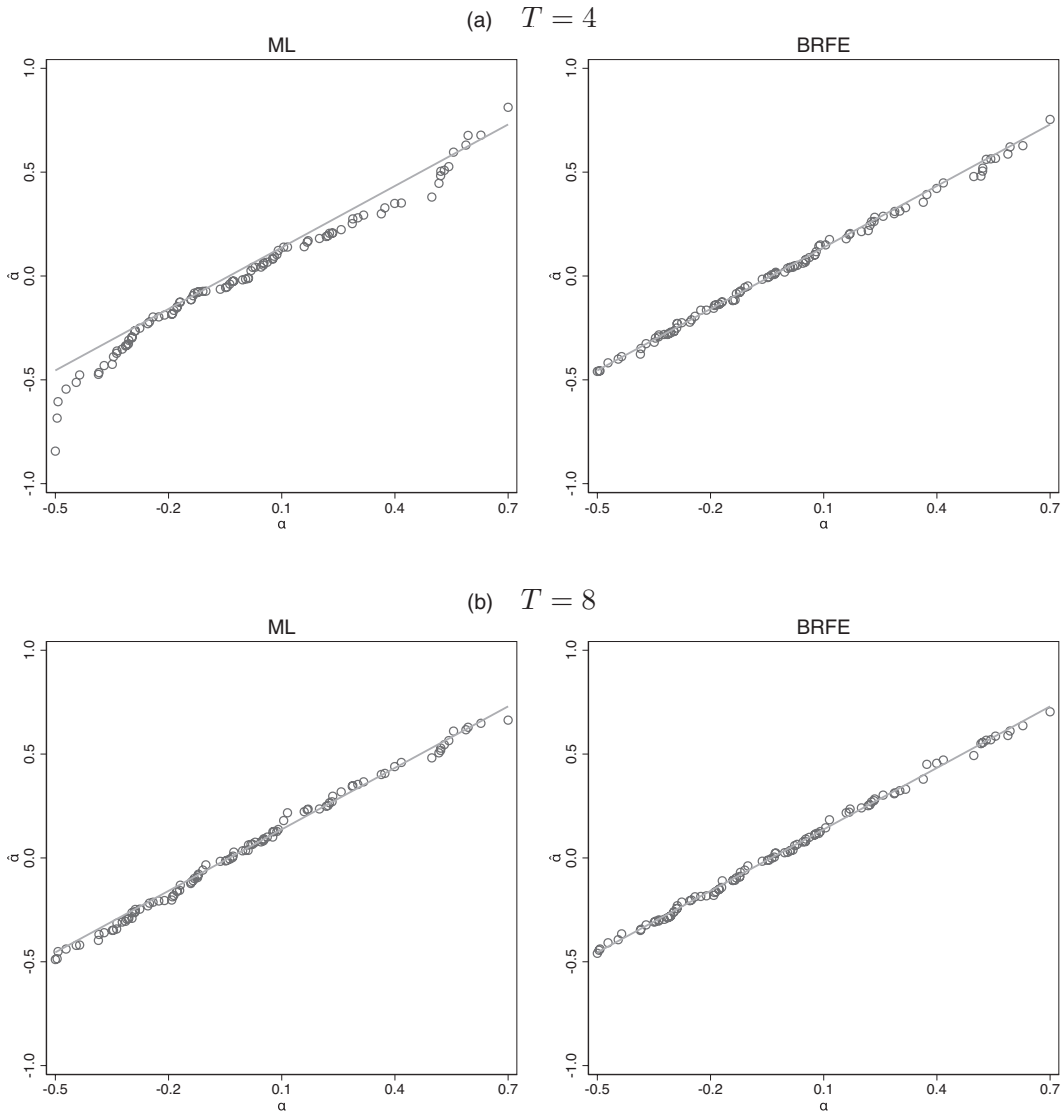

FIGURE 6 Mean of estimated individual effects by true α_i ($\alpha_i \sim beta$)
Notes: Graphs show average estimates of $\alpha_1, \ldots, \alpha_{100}$ over 500 replications against their true values.

3.4 | Robustness check: replication of Heckman (1991) DGP and comparison to nonparametric bias correction

Figure A2 in the Appendix shows graphs for results obtained for the Heckman (1991) DGP; specifically, the implementation of that DGP as in Hahn and Newey (2004), Fernández-Val (2009) and Alexander and Breunig (2016). Compared to our baseline with normally distributed α_i, this DGP features a regressor that is serially correlated and standard normally distributed α_i. The figure shows results obtained for $T = 4$ and $T = 8$. The increased heterogeneity in α_i result in a visible deterioration of the ML estimates. BR is able to predict the true α_i quite accurately even in the shorter panel over most of the distribution. However, there are some larger biases for the units in the right tail of the

distribution. In the next robustness check below, we explore the sensitivity of ML and BRFE to the degree of heterogeneity in α_i more systematically.

We also used this DGP to assess the prediction of α_i using a different, nonparametric bias correction estimator, the split-panel jackknife estimator of Dhaene and Jochmans (2015). The results are presented alongside those of ML and BRFE in Figure A2. The estimates of these nonparametric estimator are clearly inferior to both those of ML and BRFE when it comes to prediction of α_i. The split-panel jackknife estimator relies on splitting the data in half along its time dimension and estimating the model in each half. This procedure exacerbates the problem of concordance, since even a unit which is discordant over the full T periods might be concordant in either the first or second $T/2$ time periods. Thus, such methods do not appear well-suited to predict individual effects in short panels.

3.5 | Sensitivity to the degree of heterogeneity across units

The simulations so far have focussed on recovering the shape of the distribution of the individual units, α_i. We now introduce an extra dimension on which to perform the simulation experiment: the simulated (true) values of α_i are multiplied by a scalar κ. This parameter controls the degree of heterogeneity across units. We vary $\kappa = 0.1, 0.2, \ldots, 1.5$. Thus, $\kappa = 1$ corresponds to the baseline results presented previously. For $\kappa = 0.1$ and $\kappa = 1.5$ the standard deviation of the individual effects is reduced to 10 % and increased to 150 % of that in the baseline results respectively. Figure 7 presents graphs with ML and BRFE estimates of the mean and standard deviation of the distribution of individual effects for each of the 15 values of κ, separately for $T = 4$ and $T = 8$ and each of the four distributions.

For the estimation of the mean of α_i, with $T = 4$ there is only a small bias in the BRFE estimates in the Bernoulli and uniform DGPs; with $T = 8$, the BRFE estimates are virtually unbiased. Bias in the estimates of the mean of α_i is more of a problem for ML with $T = 4$, and the bias increases with larger κ. However, most of this bias also disappears with $T = 8$. An exception is the Bernoulli distribution, where the ML bias visibly persists even for large κ.

The estimation of the standard deviation of α_i is more difficult. At $T = 4$, BRFE underestimates the standard deviation, and the degree of underestimation increases with κ. While for the Bernoulli and beta distributions the underestimation is minor, it is quite visible in the case of the normal and uniform distributions in conjunction with larger κ. For virtually all distributions and all values of κ, the performance of ML is substantially worse than that of BRFE. For small values of κ, ML overestimates the standard deviation; for large values, it underestimates it. Both estimators improve considerably with $T = 8$. Bias only remains for large κ in the normal DGP (and, for the ML estimator, the uniform DGP). Still, even in these cases, the reduction of the bias in BRFE relative to ML is sizeable.

In the Appendix, Figure A3 shows the performance of BRFE and ML in predicting the individual α_i in the DGPs with the largest standard deviation ($\kappa = 1.5$) in graphs identical to Figures 3–6. In these distributions, a substantial share of the probability mass of the empirical distribution of α_i may lie outside the [-1,1] interval; for the most extreme case (normal), this share amounts to about 40 %. The results from the figure indicate that even in this challenging DGP, BRFE provides nearly unbiased estimates for most of the α_i, but the estimation of the α_i in the tails of the distributions deteriorates. In contrast, the ML estimates of α_i are often biased along the whole distribution of α_i.

While we have stopped at a maximum of $\kappa = 1.5$, the results suggest that the performance of the estimation would worsen further for higher κ. However, with a small T, such as $T = 4$, if κ is much larger than what we have presented in the simulations, there will be almost no longitudinal variation left in the data, making it unlikely that such data would be used in practice for a longitudinal analysis. Researchers can check the standard deviation of the estimated distribution of individual effects.

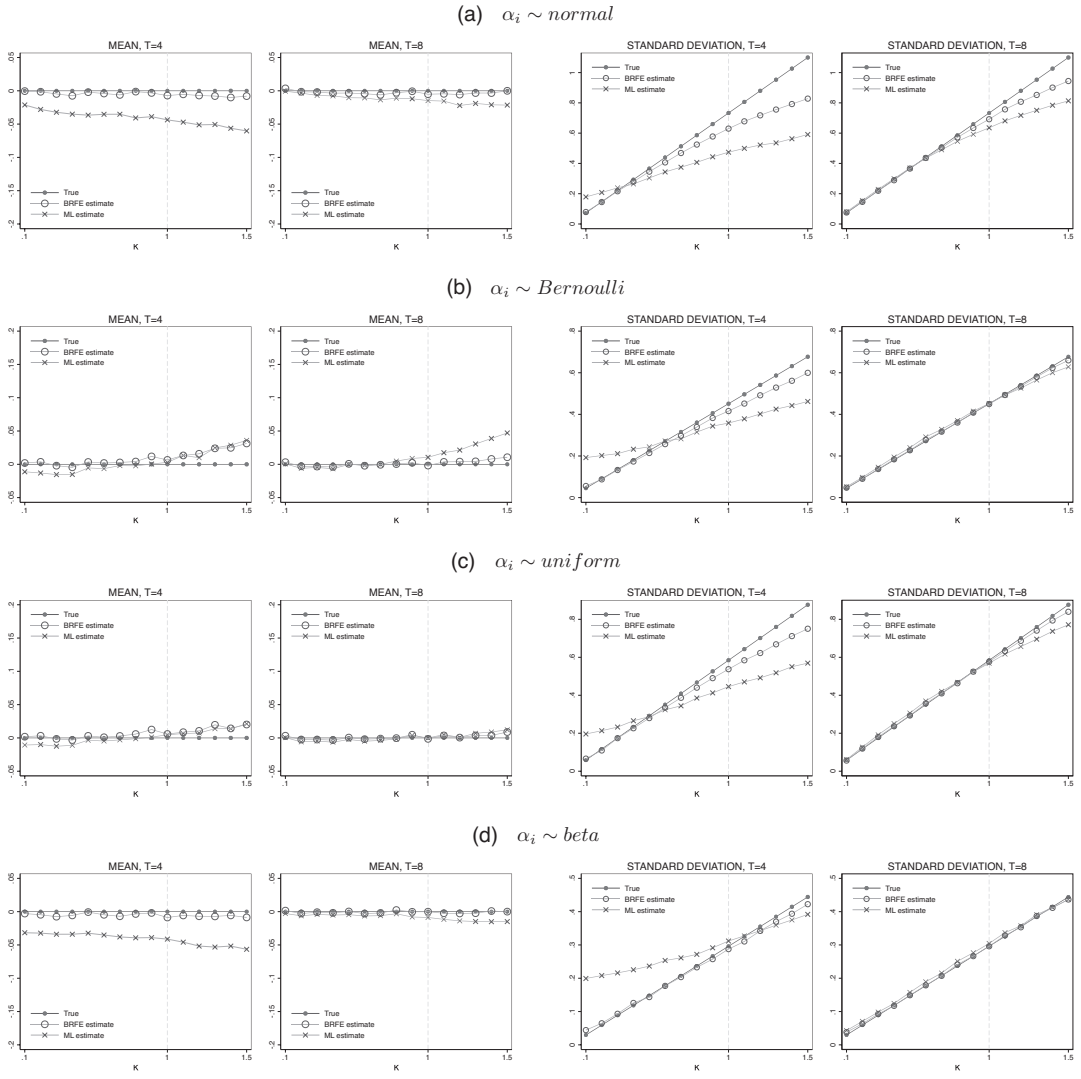

FIGURE 7 Estimated mean and estimated standard deviation of the distribution of individual effects (α_i) over different levels of heterogeneity (κ) of the individual effects

Notes: Graphs show average estimates of the mean of $\widehat{\alpha}_i - \overline{\alpha}_i^{ML}$ and $\widehat{\alpha}_i - \overline{\alpha}_i^{BRFE}$ (left-hand side panels, titled 'MEAN') and the standard deviation of $\widehat{\alpha}_i^{ML}$ and $\widehat{\alpha}_i^{BRFE}$ (right-hand side panels, titled 'STANDARD DEVIATION') over 500 Monte Carlo replications in data generating processes with four and eight time periods, and with κ=0.1,0.2, …, 1.5. In the replications corresponding to each point on the x-axis, all (true) α_i were multiplied by a fixed scalar κ. The point $\kappa = 1$ corresponds to the results reported in Table 1.

This is especially simple in probit models, where the standard deviation of the idiosyncratic errors is normalized to 1. If the standard deviation of α_i is high, care should be taken in interpreting and using BRFE estimates of individual α_i that are at either end of the quantile distribution. In the application we present in the next section, for instance, the estimated standard deviation of the individual effects ($T = 5$) is around 0.97 and 0.77 for the two samples used (males and females), and would correspond to a κ of around 1.3 and 1.0, respectively, under a normal distribution. Thus, even allowing for the fact that this estimated standard deviation of individual effects is somewhat underestimated, the application falls well within the range presented in the simulations.

4 | APPLICATION TO THE DETERMINANTS OF DOCTOR VISITS

To illustrate the benefits of being able to predict individual effects for every individual, we use data for a 2000–2014 subsample of the Socio-Economic Panel, a large representative household panel survey for Germany (SOEP, see Wagner et al., 2007). The dependent variable is an indicator variable, stating whether a visit to a physician took place (*anyvisit* = 1), or did not take place (*anyvisit* = 0), during the 3-month period prior to the annual interview. We apply the BRFE estimator to obtain predictions of the individual effects in the FE panel probit model. Our results relate to the previous literature on the demand for health services based on the number of doctor visits, a count variable (see, e.g. Cameron & Trivedi, 1986; Winkelmann, 2004). Specifically, we zoom in on the extensive margin decision, and correspondingly on the first step of a possible hurdle count data model (Mullahy, 1986; Pohlmeier & Ulrich, 1995, Kunz & Winkelmann, 2017). Our approach also relates more broadly to other strands of the health economics literature. For instance, Carro and Traferri (2014) use a bias-corrected estimator for a model of self-assessed health and considers the implied distribution of two sets of individual effects.

We focus on two advantages of our approach that are likely to be important to the study of persistence in health care utilization. First, we assess whether individual effects are in fact 'fixed', that is, capture *time-invariant* individual heterogeneity, such as genetic makeup or childhood experiences. One way to address this issue is to split the sample into different subperiods, for example 2000–2004, 2005–2009 and 2010–2014, to estimate separate individual effects for each period and then verify their stability over time. Second, we assess the predictive power of these doctoral visit individual effects onto other domains of future health care use, such as the propensity of future hospitalization. In this case, the idea is that individual effects obtained from current outcomes capture long-term health capital, which can be assessed by considering their correlation with future health outcomes.

Table 3 provides selected summary statistics, separately for men and women and the three time periods. We only include individuals who are aged between 20 and 65 at the time of interview, drop observations with missing values on any of the variables, and retain a balanced panel of 55,230 person-year observations, representing 1997 women and 1685 men. The use of a balanced sample is not necessary but it simplifies the analysis of changes in the distribution of individual effects. As a consequence of the way the sample is constructed, the average age increases by exactly 5 years for each 5-year sub-period. Hence, it is not surprising that the share of women and men with at least one visit increases over time, and self-assessed health (SAH) worsens. SAH is measured on a 5-point scale, where the best outcome [1] means 'Very good' and the worst outcome [5] means 'Bad' (the intermediate outcomes are 'Good', 'Satisfactory' and 'Poor' respectively). The main difference between males and females is the latter's overall higher prevalence of doctor visits.

The last row of the sub-panels of Table 4 states the proportion of concordant outcomes by gender and period. This proportion varies from a minimum of 29 % to a maximum of 46 %. It tends to increase over time for both men and women, which is due to the fact that overall utilization increases, making it more likely that individuals have outcome 'one' (i.e. at least one visit) in each year.

4.1 | Estimation results

Two models were estimated, separately by gender, as displayed in Table 5. First, a pooled cross-sectional probit model that does not include individual effects, estimated by ML (Columns 1 and 4). Second, a fixed effect probit model, estimated by ML (2,5) and BRFE (3,6). Both ML and the BRFE

TABLE 4 Selected descriptive statistics, by period

	2000–2004	2005–2009	2010–2014	All periods
	(1)	(2)	(3)	(4)
Men				
Any doctor, last 3 month (Yes/No)	0.559	0.595	0.638	0.597
Any hospital, last 3 month (Yes/No)	0.073	0.079	0.102	0.085
Age/10	3.954	4.454	4.954	4.454
Self-assessed health	2.385	2.555	2.671	2.537
Disability (Yes/No)	0.055	0.085	0.127	0.089
Years of education	12.478	12.567	12.595	12.546
Share of concordant obs.	0.291	0.312	0.400	0.335
Women				
Any doctor, last 3 month (Yes/No)	0.722	0.729	0.742	0.731
Any hospital, last 3 month (Yes/No)	0.122	0.101	0.103	0.109
Age/10	3.904	4.404	4.904	4.404
Self-assessed health	2.437	2.560	2.675	2.557
Disability (Yes/No)	0.044	0.067	0.104	0.072
Years of education	12.359	12.457	12.485	12.434
Share of concordant obs.	0.389	0.418	0.456	0.421

Source: SOEP v33, 2000–2014, own calculations.

estimate three sets of individual fixed effects: three indicator variables are included for each individual, one for each 5-year period.

One important takeaway is that the fixed effects ML probit coefficients, based on a model with dummies for each person without further adjustment, tend to be biased upward: they are always at least as large, in absolute values, as the BRFE probit estimates, and often substantially larger. This reflects the incidental parameters bias for a set-up where each of the individual effects needs to be estimated based on $T = 5$ observations only. Due to the substantial proportion of concordant observations, the effective ML sample size is reduced by 33% for men, and by 42% for women. The reduced sample size leads to estimated standard errors that are correspondingly higher for ML relative to BRFE.

Table 5 presents several sets of standard errors. The standard errors in parentheses are the conventional ones, based on the inverse of the information matrix evaluated at the value of the estimates. For the fixed effects ML probit, these standard errors are inconsistent due to the inconsistency of the estimates product of the incidental parameter problem (e.g. Hahn & Newey, 2004). The corresponding BRFE standard errors do not suffer from this problem given that the first-order bias in the estimation of α_i is removed. However, given that the BRFE estimator is a regular M-estimator, standard errors can also be based on the general Eicker-White–Huber asymptotic variance (the well-known sandwich estimator involving the inverse of the information matrix and the outer product of the score). Such standard errors can give a better representation of the accuracy of the estimation since they do not rely on the correct specification of the model. They also make it straightforward to account for potential correlation in the errors (clustering). In the table, we have included such robust standard errors in brackets, accounting for possible correlation within individual-effect units. Across estimators, the robust standard errors are larger than the conventional ones, although not very much. Finally, the table contains a third set of standard errors for the

TABLE 5 ML probit and BRFE probit results, all periods

	Men			Women		
	Pooled probit	Fixed effects probit		Pooled probit	Fixed effects probit	
		ML	BRFE		ML	BRFE
	(1)	(2)	(3)	(4)	(5)	(6)
Self-assessed health	0.40	0.49	0.36	0.36	0.45	0.31
	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)
	[0.01]	[0.03]	[0.02]	[0.01]	[0.02]	[0.02]
			{0.02}			{0.02}
Disability (Yes/No)	0.68	0.26	0.15	0.79	0.61	0.28
	(0.04)	(0.14)	(0.09)	(0.05)	(0.17)	(0.08)
	[0.06]	[0.16]	[0.10]	[0.06]	[0.20]	[0.08]
			{0.10}			{0.09}
Married (Yes/No)	0.05	0.13	0.10	0.02	0.14	0.10
	(0.02)	(0.08)	(0.06)	(0.02)	(0.08)	(0.05)
	[0.03]	[0.10]	[0.08]	[0.03]	[0.10]	[0.07]
			{0.06}			{0.07}
Age/10	−0.53	−0.94	−0.67	−0.65	−0.89	−0.62
	(0.08)	(0.40)	(0.31)	(0.08)	(0.38)	(0.26)
	[0.12]	[0.47]	[0.39]	[0.10]	[0.45]	[0.34]
			{0.36}			{0.30}
$Age^2/100$	0.06	0.13	0.09	0.07	0.08	0.06
	(0.01)	(0.04)	(0.03)	(0.01)	(0.04)	(0.03)
	[0.01]	[0.05]	[0.04]	[0.01]	[0.05]	[0.04]
			{0.04}			{0.03}
Years of education	0.02	−0.24	−0.17	0.01	−0.01	−0.01
	(0.00)	(0.07)	(0.05)	(0.00)	(0.05)	(0.04)
	[0.00]	[0.07]	[0.06]	[0.00]	[0.06]	[0.05]
			{0.05}			{0.04}
Fulltime work (Yes/No)	0.03	0.03	0.03	0.09	0.16	0.12
	(0.04)	(0.09)	(0.06)	(0.02)	(0.06)	(0.04)
	[0.05]	[0.10]	[0.08]	[0.03]	[0.07]	[0.06]
			{0.07}			{0.05}
Parttime work (Yes/No)	0.09	0.14	0.10	0.06	0.06	0.04
	(0.04)	(0.09)	(0.06)	(0.02)	(0.05)	(0.03)
	[0.05]	[0.09]	[0.07]	[0.03]	[0.06]	[0.04]
			{0.06}			{0.04}

(Continues)

TABLE 5 (Continued)

	Men			Women		
	Pooled probit	Fixed effects probit		Pooled probit	Fixed effects probit	
		ML	BRFE		ML	BRFE
	(1)	(2)	(3)	(4)	(5)	(6)
Log household income	0.06	−0.07	−0.05	0.05	−0.01	−0.01
	(0.02)	(0.05)	(0.04)	(0.02)	(0.05)	(0.03)
	[0.02]	[0.06]	[0.05]	[0.02]	[0.05]	[0.04]
			{0.04}			{0.04}
Number of observations	25,275	16,820	25,275	29,955	17,345	29,955
Number of individuals	1685	1194	1685	1997	1221	1997
Individual × 5-year-period fixed effects		✓	✓		✓	✓

Notes: Table contains main estimation results (coefficients and standard errors), regressing any doctor visit in last 3 month, separately for males (Columns 1–3) and females (Columns 4–6) using pooled ML probit (1,4), ML fixed effects probit (2,5) including three individual indicator variables for individual×period (2000–2004, 2005–2009, 2010–2014), and the BRFE probit estimator (3,6). Standard errors in parentheses, cluster-robust standard errors in brackets (clustered at the level of the fixed effects), bootstrap standard errors in braces (resampled from clusters at the level of the fixed effects, 200 replications).

Source: SOEP v33, 2000–2014, own calculations.

BRFE estimates, in braces. These are bootstrap standard errors (resampled from clusters at the level of the individual-effect units), which we present to evaluate how well the asymptotic BRFE standard errors (in parentheses and brackets) compare to standard errors based on the observed sample size. The bootstrap standard errors in the table are of a magnitude similar to the asymptotic ones, most often being either equal or slightly smaller to the robust standard errors. Thus, the use of robust (asymptotic) standard errors seems to be an advisable conservative choice in practice for the proposed estimator.

In terms of substantive results (based on the BRFE probit estimates), better self-assessed health reduced the probability of any doctor visit for both men and women. We calculate a simple approximate upper bound for the effect size: for a man at the margin, that is, with a 50/50 chance of seeing a doctor, a one-point worsening in SAH increases the probability of seeing a doctor by about $0.36 \times 0.4 = 14.4$ percentage points. Married women are more likely to see a doctor than non-married ones, and having more education reduces the probability only for men.

4.2 | Analysing individual effects

The BRFE probit estimator directly provides predictions for the individual-specific effects. These are available for further analyses, for example, for ranking of individuals by their underlying propensity of experiencing the event. Figure 8 depicts a few examples of the distributions of individual effects. Panel A compares the distribution of individual effects from the first period (2000–2004) between males and females. Panel B compares the distribution of individual effects of males over time (2000–2004 vs. 2010–2014) and C, analogously, the ones of females.

The shapes of the distributions differ markedly. For males, the time constant propensity net of covariates has a larger mean, is more variable, and appears to be unimodal. In contrast, the females'

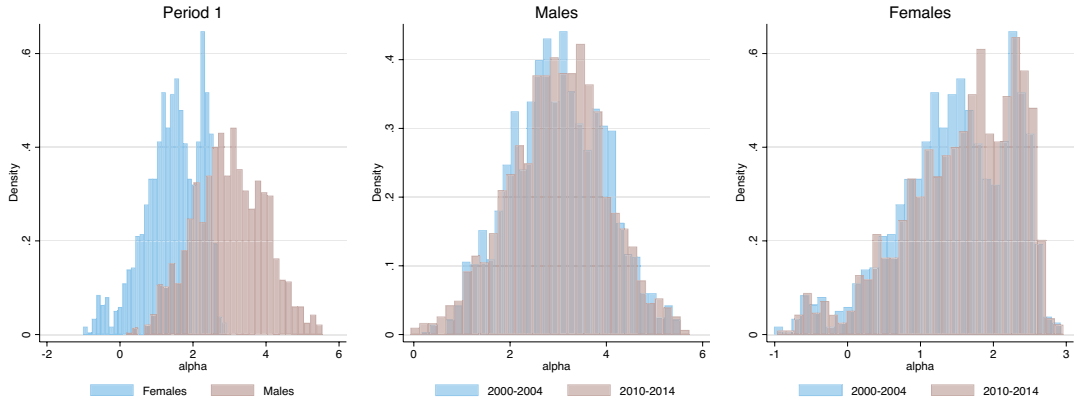

FIGURE 8 Histograms of predicted individual effects

Notes: Predicted individual effects based on separate regressions for male and females presented in Table 3 for the BRFE probit estimator. Regressions are pooled across years and include three sets of individual×period fixed effects: period 1 (2000–2004), period 2 (2005–2009), period 3 (2010–2014). The left panel compares the distributions of individual effects in the first period of males and females; the middle panel, the distributions of males in period 1 and period 3; the right panel, those of females. *Source*: SOEP v33, 2000–2014, own calculations.

FIGURE 9 Scatterplots of period-specific individual effects

Notes: Plot depicts correlation of the individual×period effects based on BRFE probit results of Table 3, from first period to second (blue) and third period (red). Lines denote the linear fit. Left panel males, right panel females. *Source*: SOEP v33, 2000–2014, own calculations.

individual effects distribution is shifted to the left and exhibits three modal points, which could indicate a finite mixture type of distribution. This shows the benefits of the fixed effects approach, which does not restrict the shape of the unobserved heterogeneity, relative to, for example, random effects models with parametric distributional assumptions.

To test that the BRFE is indeed able to predict these individual effects well, we performed Monte Carlo simulations based on resampling from the actual SOEP data and estimating the full empirical model. That is, the variables, the distribution of the individual effects, and the value of the β parameters are those estimated with real data. We limited the simulation to the first period (2000–2004) for both women and men, and the results are shown in Appendix Figure A4. The figure confirms that the estimator is able to precisely recover the true values in an empirically relevant setting.

Since for each of the $N = 1685$ males and 1997 females we have obtained three distinct predictions, the total variance of the $3N$ effects (5055 for males and 5991 for females) can be decomposed into a within-person and a between-person component. The individual-specific component is fairly stable over time: Of its total variance (0.956 for males and 0.601 for females) variation within individuals contributes not more than 26 and 36 % respectively.

The bivariate scatter plots in Figure 9 provide another perspective on stability over time. We find a substantial positive correlation in estimated individual effects for the same person between adjacent sample periods (black dots and linear fit). Yet, the fit is far from the 45-degree line, which would imply complete stability. When comparing the first to the last period, 10 years later (lightly coloured dots and linear fit), the correlation is again lower (indicated by the flatter linear fit), although only slightly so. For women, the 5-year fixed effects change somewhat more over the longer time frame, implying a slightly lower stability over time.

In Figure 10, we assess the predictive power of the first-period individual effects for another measure of heath care utilization: The 10-year-ahead probability to visit any hospital (in the last 3 month before the annual survey elicitation). To assess this, we group the first-period individual effects (2000–2004) into 10 percentile bins and plot them against the unconditional mean hospital visit probability in the last period of our sample (2010–2014). The percentile ranks of the time-constant doctor-visit individual effects are clearly predictive of the overall probability to visit any hospital 10 years later,

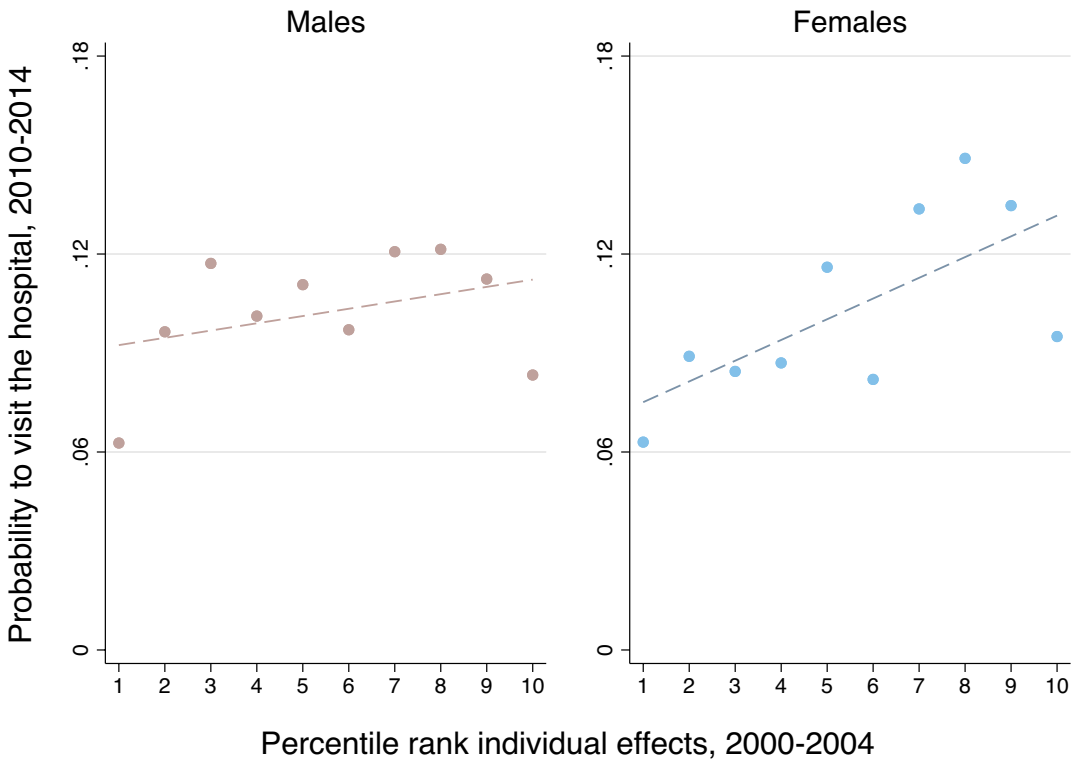

FIGURE 10 Predicting future hospital visits

Notes: Figure plots the average probability to visit a hospital (in the last 3 month) in 2010–2014 against bins of percentile rank in the individual unobserved time constant component of doctor visits (individual effects) 10 years earlier (2000–2004), based on the BRFE probit results of Table 3. Left panel males, right panel females, dashed lines depict the linear fit.
Source: SOEP v33, 2000–2014, own calculations.

and this predictive ability is stronger for females than for males. Taken together, these results show the persistence in unobserved heterogeneity not only over a long period of an individual's life cycle, but also the spillover into other (more costly) domains of health care use, findings which might need to be better accounted for in general models of health care demand.

Finally, we use predictions to explore any time-invariant relationships between the observed and unobserved components of the demand for health services. For men, the correlation between $\widehat{\alpha}_i$ and $\overline{x}_i'\widehat{\beta}$ (where $\overline{x}_i$ are the average characteristics) equals -0.403, -0.414, and -0.409 in the first, second, and third period respectively; Hence, individual differences in observed factors tend to be associated with unobservables that move in the opposite direction. Ignoring this correlation (such as in the pooled probit model) would understate the importance of either of the two. Interestingly, the correlation among females, in contrast, is almost non-existent (0.018, 0.031 and 0.062 respectively), implying a potentially less problematic assessment of the cross-sectional variation.

5 | CONCLUSIONS

This paper discusses the problem of estimating a panel probit model with individual-specific constants treated as fixed effects. A useful estimator should address the incidental parameters problem, provide finite estimates for all individual effects, and ideally estimate them with small bias. We show that a specific first-order bias reduction method, based on a modified joint score function of the structural parameters and the individual effects as in Kosmidis and Firth (2009), addresses all three of the above issues.

An extended simulation study confirms that the theoretical properties of this estimator materialize already in quite small samples (e.g. $T = 4$ and $N = 100$). The incidental parameter bias disappears, and the estimated individual effects are nearly unbiased for the true heterogeneity parameters. While other first-order bias correction methods have been proposed to overcome the incidental parameters problem (e.g. Hahn & Newey, 2004, or Bester & Hansen, 2009), these do not remove the concordance problem that is so common in applications, especially in short panels. In the data of our application—balanced panels covering 5-year periods—about 40 % of the observations were concordant and would have led to infinite estimates for the corresponding individual effects had we used the conventional fixed effects ML estimators or other bias-corrected estimators.

One could argue that the technical problem of binary response models leading to infinite estimates of α_i is not necessarily a problem; that some α_i might actually be equal to positive or negative infinity. However, for many applications, having infinite individual effects is clearly unattractive. From a theoretical point of view, an infinite α_i implies that no unobserved or observed shock could possibly produce a change in the outcome, a very strong assumption. In our application, this would imply a person that has not visited a doctor in 5 years will under no circumstance ever visit a doctor. From a practical point of view, discarding the information contained in time-varying covariates seems inefficient and leads to a clumping of predictions with the same values. Moreover, as seen in the simulations and application, the exclusion of a potentially large share of the data from the estimation is also likely to lead to larger standard errors for the parameter β.

Of course one can focus on predicted probabilities instead of directly on the individual effects. In that case, the infinite individual effects of concordant units translate to fitted probabilities of zero and one. Obviously, this does not solve the problem of a clump in the predictions stemming from ignoring the information in the covariates. Again, from an empirical point of view, making extreme predictions that assign probability of one and zero based on a small number of observations might often be undesirable. Finally, obtaining predicted probabilities of zero and one can be problematic when used in further analyses that require probabilities to be strictly bounded away from these values; examples

include using the predictions for inverse probability weighting (such as propensity score and marginal treatment effects methods) or as inputs in some structural dynamic discrete choice models. This last problem also affects FE linear probability models estimated by OLS. While they obtain finite estimates for the individual effects of concordant units, linear probability models do not restrict the range of predictions to the unit interval, and concordant units frequently result in predictions greater than one or smaller than zero. The advocated BRFE probit does not suffer from any of these problems. It produces consistent and meaningful fitted probabilities which lie strictly within the (0,1) interval.

Because of its advantages over other approaches across a number of domains, coupled with the simplicity of implementing it, we believe that the BRFE probit estimator has the essential features of a new workhorse for the estimation of panel probit models. We focused in our discussion on the probit model as it is a common choice in empirical work, but, as shown in the appendix, the advocated approach is applicable to a number of other binary response models as well, such as logit or cloglog. More broadly, the estimator can be extended to other non-linear fixed effects panel models which suffer from perfect prediction, such as models for ordered and count data. Finally, since bias reduction eliminates any first-order bias, including, potentially, the bias introduced by specifications where there are lags of the dependent variable, the BRFE approach can also be directly applied to dynamic panel data models with individual effects and lagged dependent variables (see Buchmueller et al., 2021, for an example).

ACKNOWLEDGEMENTS

We thank the editor, the associate editor, referees, Alex Ballantyne, Bob Breunig, Stefan Bruder, Joseph Doyle, Denzil Fiebig, Joe Hirschberg, Andrew Jones, Maarten Lindeboom, Jenny Lye, Janina Nemitz, Dennis Petrie, Carol Propper, Chris Skeels, Jenny Williams, Frank Windmeijer, Tiemen Woutersen, Valentin Zelenyuk and participants in the European Workshop on Econometrics and Health Economics in Groningen, the conference of the Australian Health Economics Society in Sydney, the meeting of the International Health Economics Association in Basel and the Australia New Zealand Econometric Study Group meeting in Melbourne for helpful comments. An earlier version of the paper circulated under the title 'Estimating fixed effects: Perfect prediction and bias in binary response panel models'. Staub acknowledges funding from the Australian Research Council through grant DE170100644. STATA code for the bias-reduced estimators described in this paper is available from the authors' websites or through GitHub under the name https://github.com/JohannesSKunz/brfeglm.

ORCID

Johannes S. Kunz http://orcid.org/0000-0003-3289-5147
Kevin E. Staub http://orcid.org/0000-0002-1546-1761
Rainer Winkelmann http://orcid.org/0000-0002-8686-5560

REFERENCES

Abrams, D.S., Bertrand, M. & Mullainathan, S. (2012) Do judges vary in their treatment of race?. *Journal of Legal Studies*, 41(2), 347–383.

Abrevaya, J. (1997) The equivalence of two estimators of the fixed-effects logit model. *Economics Letters*, 55(1), 41–43.

Alexander, B. & Breunig, R. (2016) A Monte Carlo study of bias corrections for panel probit models. *Journal of Statistical Computation and Simulation*, 86(1), 74–90.

Arellano, M. & Hahn, J. (2016) A likelihood-based approximate solution to the incidental parameter problem in dynamic non-linear models with multiple effects. *Global Economic Review*, 45(3), 251–274.

Autor, D., Duggan, M. & Gruber, J. (2014) Moral hazard and claims deterrence in private disability insurance. *American Economic Journal: Applied Economics*, 6(4), 110–41.

Bartolucci, F. & Nigro, V. (2010) A dynamic model for binary panel data with unobserved heterogeneity admitting a $\sqrt{n}$-consistent conditional estimator. *Econometrica*, 78(2), 719–733.

Bartolucci, F., Bellio, R., Salvan, A. & Sartori, N. (2016) Modified profile likelihood for fixed-effects panel data models. *Econometric Reviews*, 35(7), 1271–1289.

Bester, A.C. & Hansen, C. (2009) A penalty function approach to bias reduction in non-linear panel models with fixed effects. *Journal of Business & Economic Statistics*, 27(2), 131–148.

Browning, M. & Carro, J.M. (2014) Dynamic binary outcome models with maximal heterogeneity. *Journal of Econometrics*, 178(2), 805–823.

Buchmueller, T., Cheng, T.C., Pham, N.T.A. & Staub, K.E. (2021) The effect of income-based mandates on the demand for private hospital insurance and its dynamics. *Journal of Health Economics*, 75, 102403.

Cameron, A.C. & Trivedi, P.K. (1986) Econometric models based on count data: comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, 1, 29–53.

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

Carro, J.M. & Traferri, A. (2014) State dependence and heterogeneity in health using a bias?corrected fixed?effects estimator. *Journal of Applied Econometrics*, 29(2), 181–207.

Chernozhukov, V., Fernández-Val, I., Hahn, J. & Newey, W. (2013) Average and quantile effects in nonseparable panel models. *Econometrica*, 81(2), 535–580

Chetty, R. & Hendren, N. (2018) The impact of neighborhoods on intergenerational mobility II: county-level estimates. *Quarterly Journal of Economics*, 113(3), 1163–1228.

Chetty, R., Friedman, J.N. & Rockoff, J.E. (2014) Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9), 2633–2679.

Cox, D.R. & Snell, E.J. (1971) On test statistics calculated from residuals. *Biometrika*, 58(3), 589–594.

Dhaene, G. & Jochmans, K. (2015) Split-panel jackknife estimation of fixed-effect models. *Review of Economic Studies*, 82(3), 991–1030.

Ehm, W. (1991) Statistical problems with many parameters: critical quantities for approximate normality and posterior density based inference. Habilitationsschrift, University of Heidelberg.

Fernández-Val, I. (2009) Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics*, 150(1), 71–85.

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

Greene, W.H. (2004) The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *Econometrics Journal*, 7(1), 98–119.

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

Hampel, F.R., Ron Chetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986) *Robust statistics: the approach based on influence functions*. Wiley Series in Probability and Statistics. Vol. (196). Hoboken: John Wiley & Sons.

Heckman J.J. (1991) The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. In: Manski, C.F., McFadden, D.L. (Eds.) *Structural analysis of discrete data and econometric applications*. Cambridge, MA: MIT Press. 179195.

Heinze, G. & Schemper, M. (2002) A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16), 2409–2419.

Honoré, B.E. & Tamer, E. (2006) Bounds on parameters in dynamic discrete choice models. *Econometrica*, 74(3), 611–629

Kosmidis, I. (2007) Bias reduction in exponential family non-linear models. PhD Thesis, University of Warwick.

Kosmidis, I. & Firth, D. (2009) Bias reduction in exponential family non-linear models. *Biometrika*, 96(4), 793–804.

Kunz, J.S. & Winkelmann, R. (2017) An econometric model of health care demand with non-linear pricing. *Health Economics*, 26(6), 691–702.

Lunardon, N. (2018) On bias reduction and incidental parameters. *Biometrika*, 105, 233–238.

Maddala, G.S. (1983) *Qualitative and limited dependent variable models in econometrics*. Cambridge: Cambridge University Press.

Mullahy, J. (1986) Specification and testing in some modified count data models. *Journal of Econometrics*, 33, 341–365.

Pohlmeier, W. & Ulrich, V. (1995) An econometric model of the two-part decision-making process in the demand for health care. *Journal of Human Resources*, 30(2), 339–361.

Sartori, N. (2003) Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika*, 90, 533–49.

Sun, Y. & Dhaene, G. (2019) xtspj: A command for split-panel jackknife estimation. *The Stata Journal*, 19(2), 335–374.

Wagner, G.G., Frick, J.R. & Schupp, J. (2007) The German socio-economic panel study (SOEP): scope, evolution and enhancements. *Schmollers Jahrbuch*, 127, 139–169.
Winkelmann, R. (2004) Co-payments for prescription drugs and the demand for doctor visits: evidence from a natural experiment. *Health Economics*, 13, 1081–1089.

How to cite this article: Kunz JS, Staub KE, Winkelmann R. Predicting individual effects in fixed effects panel probit models. *J R Stat Soc Series A*. 2021;00:1–37. https://doi.org/10.1111/rssa.12722

APPENDIX A

ADDITIONAL MONTE CARLO SIMULATION RESULTS

TABLE A1 MC simulation: replication of Alexander and Breunig (2016), Tables 1–2

		Own simulations		Simulations reported in Alexander and Breunig (2016)						
	Statistic	**BRFE**	**ML**	**ML**	**Jackknife**	**HN-B**	**HN-M**	**FV**	**SPJ**	**MPL**
$T = 4$	Mean of $\widehat{\beta}$	0.895	1.400	1.402	0.755	1.105	1.211	1.054	0.580	1.148
	Median of $\widehat{\beta}$	0.881	1.354	1.376	0.760	1.084	1.199	1.041	0.631	1.135
	SD of $\widehat{\beta}$	0.275	0.446	0.404	0.253	0.315	0.340	0.284	0.823	0.316
	RMSE of $\widehat{\beta}$	0.294	0.599	0.570	0.352	0.332	0.400	0.289	0.924	0.349
$T = 8$	Mean of $\widehat{\beta}$	0.976	1.178	1.188	0.957	1.058	1.098	1.023	0.942	1.059
	Median of $\widehat{\beta}$	0.967	1.166	1.191	0.958	1.060	1.100	1.025	0.944	1.061
	SD of $\widehat{\beta}$	0.147	0.181	0.146	0.114	0.129	0.127	0.120	0.251	0.126
	RMSE of $\widehat{\beta}$	0.149	0.254	0.238	0.122	0.142	0.160	0.123	0.257	0.139
$T = 4$	Average ME(x)	0.909	0.968	0.985	1.016	0.975	1.055	0.928	1.050	0.863
	Median ME(x)	0.874	0.960	0.986	1.021	0.978	1.059	0.930	1.054	0.868
	SD of ME(x)	0.126	0.257	0.253	0.285	0.257	0.269	0.234	0.407	0.221
	RMSE of ME(x)	0.156	0.259	0.253	0.285	0.259	0.274	0.245	0.410	0.260
$T = 8$	Average ME(x)	0.981	0.989	0.998	1.007	1.007	1.035	0.983	1.002	0.951
	Median ME(x)	1.053	1.150	0.994	1.004	1.003	1.031	0.979	0.993	0.948
	SD of ME(x)	0.124	0.189	0.106	0.109	0.109	0.106	0.104	0.178	0.101
	RMSE of ME(x)	0.125	0.190	0.106	0.110	0.109	0.112	0.106	0.178	0.112

Notes: Cells in the columns labelled 'own simulations' contain statistics calculated over 500 replications of the same DGP as in Alexander and Breunig (2016), Tables 1 (upper panel, statistics for $\widehat{\beta}$) and 2 (lower panel, statistics for marginal effects of x_{it} —indicated as ME(x)). The remaining columns are reproduced from Alexander and Breunig (2016) Tables 1–2). The true value of β is 1. In the lower panel, statistics refer to ratios of estimated marginal effects to the true marginal effect. SD and RMSE stand for standard deviation and root mean squared error. 'Jackknife', HN-B and HN-M are estimators proposed in Hahn and Newey (2004), FV is the estimator in Fernández-Val (2009), SPJ is the estimator in Dhaene and Jochmans (2015), and MPL is the estimator in Bartolucci et al (2014). The DGP is $y_{it} = 1(x_{it}\beta + \alpha_i + \varepsilon_{it})$, $x_{it} = 0.1t + 0.5x_{it-1} + u_{it}$, $u_{it} \sim Uniform(-0.5, 0.5)$, $x_{i0} = u_{i0}$, $\alpha_i \sim N(0,1)$, $\varepsilon \sim N(0,1)$.

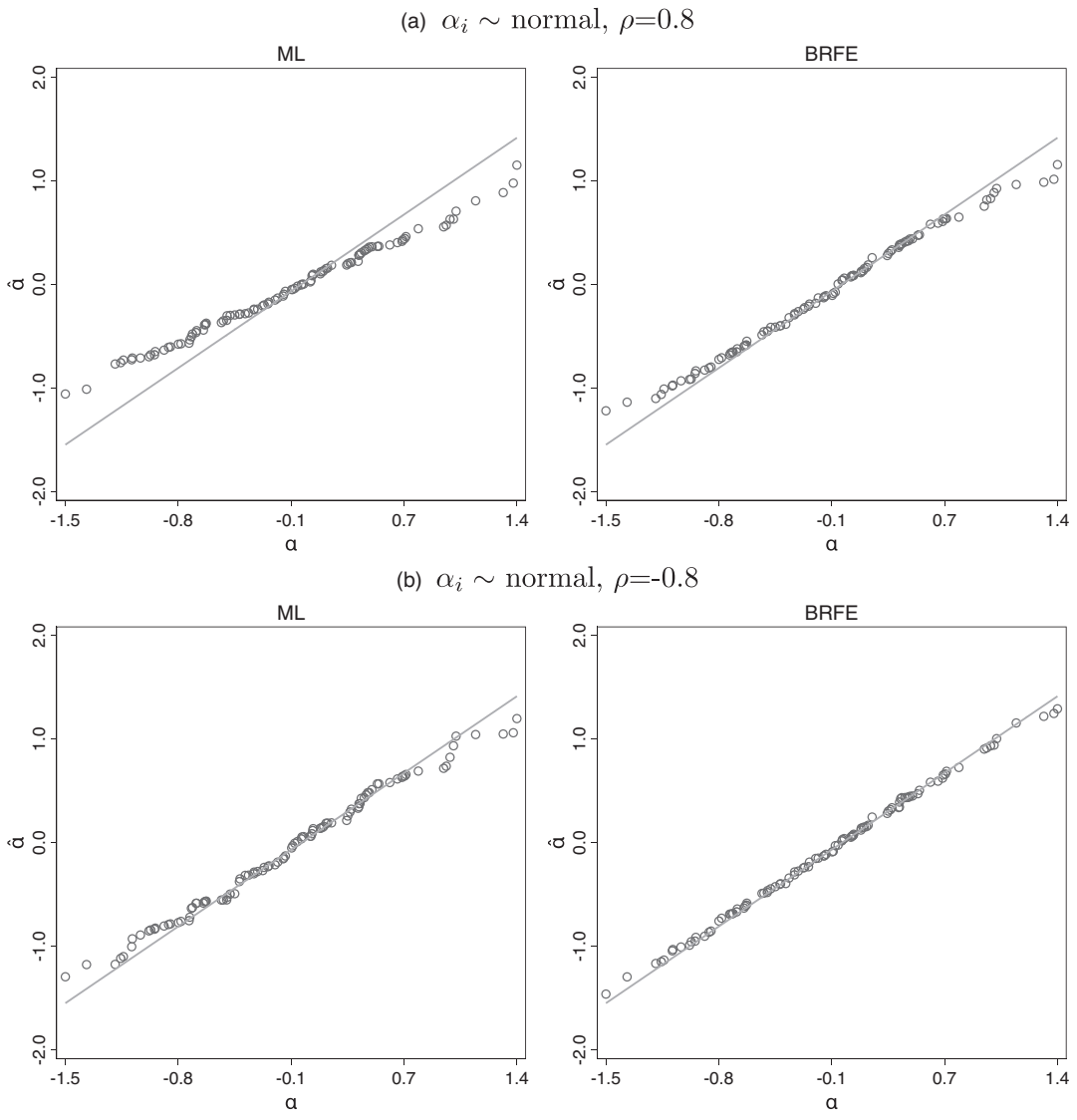

FIGURE A1 Appendix: Estimated versus true distributions of α_i when α_i and x_{it} are correlated, N=100, T=4

Notes: Graphs show average estimates of $\alpha_1, \ldots, \alpha_{100}$ over 500 replications against their true values. The hundred average $\widehat{\alpha}_i$ estimated over 500 replications. The DGP has a regressor with the same marginal distribution ($x_{it} \sim U(-1, 1)$) and an alpha with the same marginal distribution ($\alpha_i \sim N(0, 0.5)$) as in the simulations of Section 3. The dependence between x_{it} and α_i is induced by a Gaussian copula with parameter ρ. In Panel A, $\rho = 0.8$, and in Panel B $\rho = -0.8$, leading to a correlation between x_{it} and α_i of about 0.40 and -0.40 respectively. The case of $\rho = 0$ corresponds to Panel (A) of Figure 3 shown in Section 3.

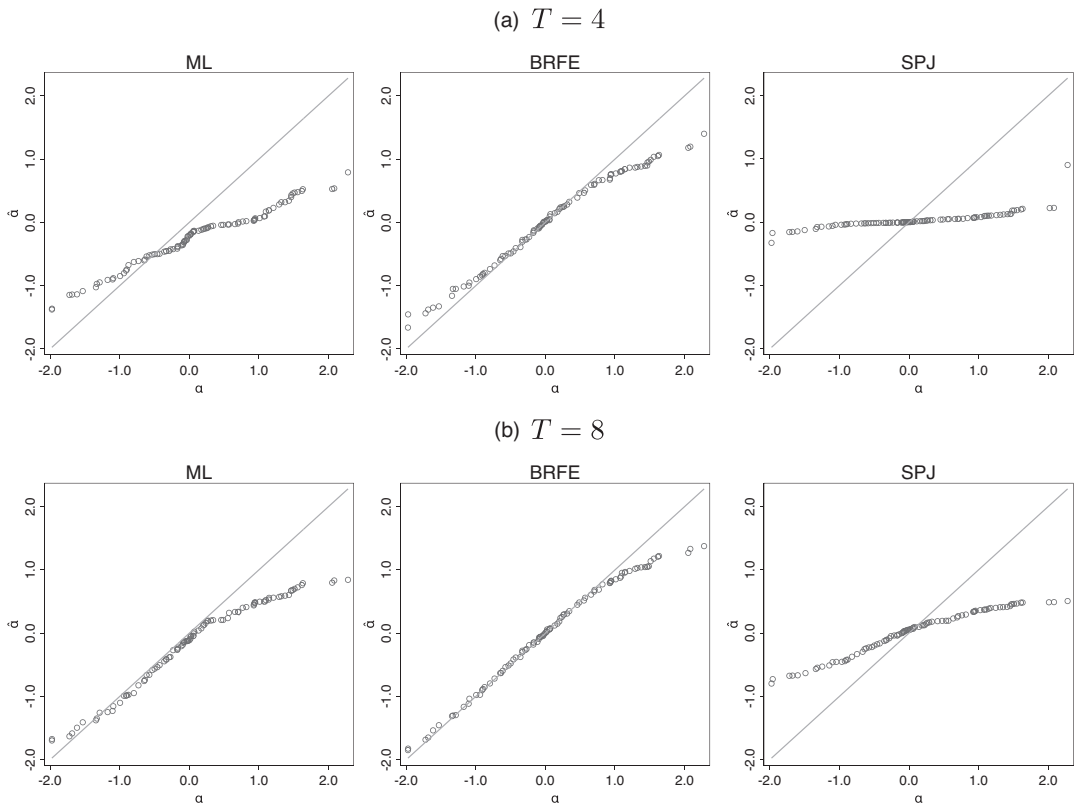

FIGURE A2 Estimated versus true distributions of α_i for the baseline DGP employed by Heckman (1991), Hahn and Newey (2004), Fernández-Val (2009), and Alexander and Breunig (2016).

Notes: Graphs show average estimates of α_i over 500 replications against their true values. SPJ denotes the split-panel jackknife estimator proposed in Dhaene and Jochmans (2015) implemented through the STATA command by Sun and Dhaene (2019). Of the 100 individual effects, the following two outliers were excluded from the graphs for better visibility: The average estimates for $\alpha_{13} = -2.80$ were $\widehat{\alpha}_{13}^{ML} = -1.92$ and $\widehat{\alpha}_{13}^{BRFE} = -2.26$; for SPJ, no finite estimates were obtained for α_{13}. The average estimates for $\alpha_{49} = -2.81$ were $\widehat{\alpha}_{49}^{ML} = -1.82$, $\widehat{\alpha}_{49}^{BRFE} = -2.16$, and $\widehat{\alpha}_{49}^{SPJ} = -0.79$.

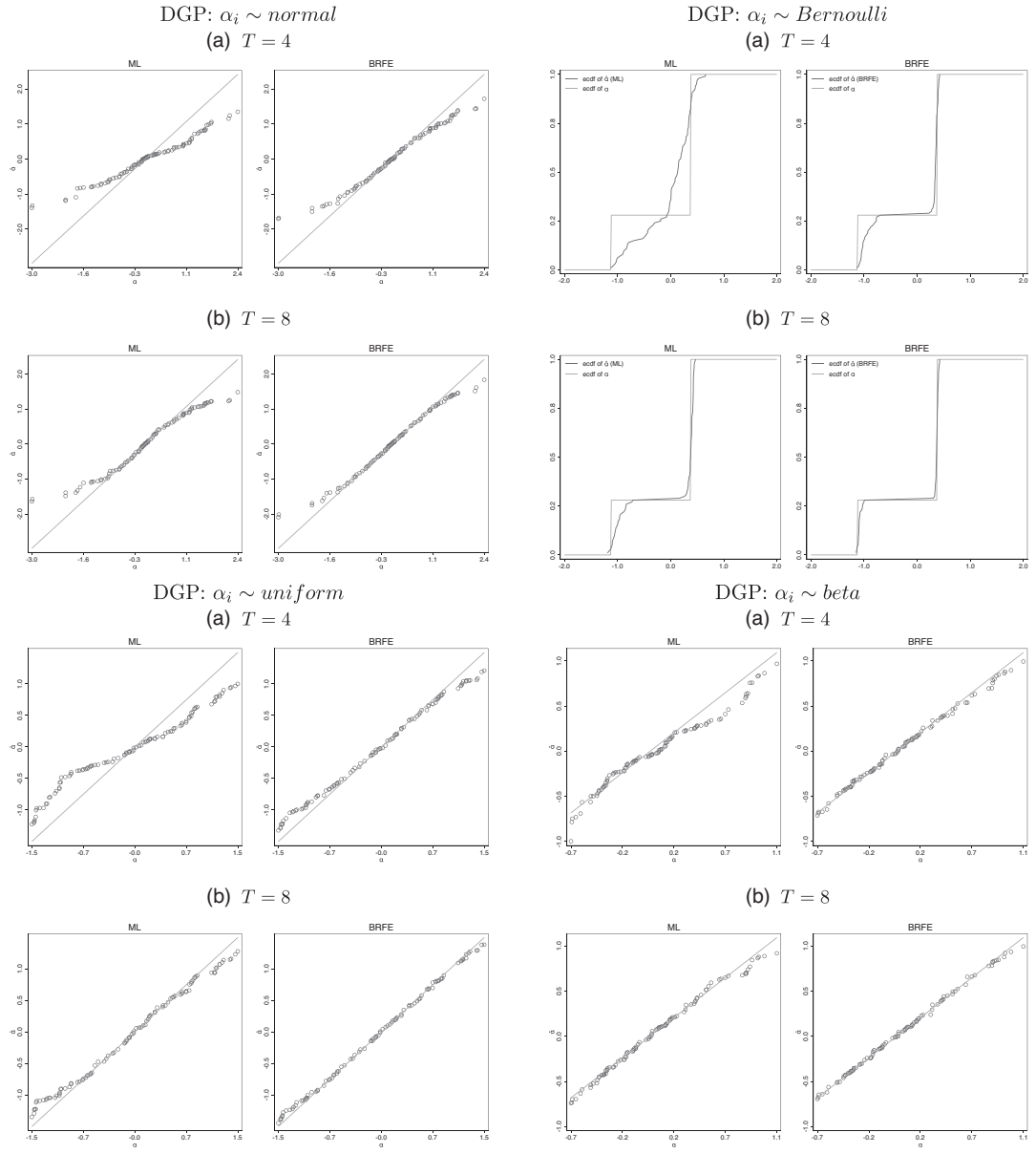

FIGURE A3 DGP with $\kappa = 1.5$: Mean of estimated individual effects by true α_i

Notes: Graphs show average estimates of α_1, …, α_{100} over 500 replications against their true values.

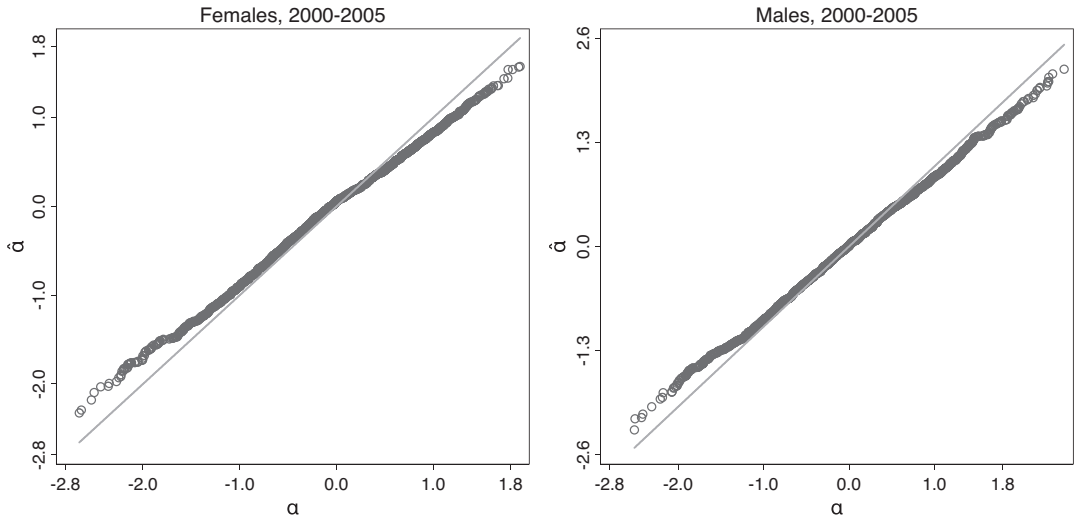

FIGURE A4 Estimated versus true distributions of α_i for a DGP based on the GSOEP data used in the empirical application.

Notes: Graphs show average BRFE estimates of α_i over 500 replications against their true values. 'True' values where obtained by performing BRFE probit of the probability of visiting the doctor with the same specification as in the empirical application.

APPENDIX B

DETAILS ON THE BRFE ESTIMATOR FOR GENERAL BINARY RESPONSE PANEL MODELS

B.1 Modified score for α_i

For a general binary response fixed effects model with

$$P(y_{it} = 1 \,|\, x_{it}, \alpha_i) = E(y_{it} = 1 \,|\, x_{it}, \alpha_i) = G(\eta_{it}) = G(\alpha_i + x_{it}'\beta) \quad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

where G:IR→(0,1) is a known smooth and strictly increasing distribution function, the modified score of the bias-reduced estimator for the parameter α_i is

$$s^{BRFE}(\alpha_i) = s^{ML}(\alpha_i) + \frac{1}{2}\sum_{t=1}^{T} h_{it}\frac{g_{it}'}{g_{it}},$$

where $g_{it} = g(\eta_{it})$ and $g_{it}' = g'(\eta_{it})$ are the first and second derivative of $G_{it} = G(\eta_{it})$ with respect to α_i, and h_{it} is the it-th diagonal elements of the $NT \times NT$ projection matrix

$$H = W^{1/2}X(X'WX)^{-1}X'W^{1/2},$$

with X the $NT\times(K+N)$ matrix of the K regressors and N panel unit indicator vectors, and W is the $NT\times NT$ diagonal matrix with typical element

$$w_{it} = \frac{g_{it}^2}{G_{it}(1-G_{it})}.$$

The expressions for probit, $G_{it} = \Phi(\eta_{it})$, are given in Section 2. For logit, $G_{it} = \Lambda_{it} = \Lambda(\eta_{it}) = \exp(\eta_{it})/(1+\exp(\eta_{it}))$, and so

$$s^{BRFE}(\alpha_i) = \sum_{t=1}^{T} y_{it} - \Lambda_{it} + h_{it}\left(\frac{1}{2} - \Lambda_{it}\right),$$

with the corresponding h_{it} being based on W with typical element $w_{it} = \Lambda_{it}(1-\Lambda_{it})$.

B.2 IRLS estimation

The BRFE estimator can be obtained by iteratively reweighted least squares. In iteration $s+1$, estimates are obtained by solving the the weighted least squares first-order conditions

$$\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\%y_{it}^{s} - \widehat{\eta}_{it}^{s+1}\right)\widehat{w}_{it}^{s} = 0,$$

where $\widehat{\eta}_{it}^{s+1} = \widehat{\alpha}_i^{s+1} + x_{it}'\widehat{\beta}^{s+1}$ contains the updated estimates, and $\tilde{y}_{it}^{s}$ and $\widehat{w}_{it}^{s}$ are constructed using iteration-s estimates of η_{it}. The expression for w_{it} was given above, and $\tilde{y}_{it}$ is defined as

$$\tilde{y}_{it} = \eta_{it} + \frac{(y_{it}^{*} - G_{it})}{g_{it}}, \quad \text{with } y_{it}^{*} = y_{it} + \frac{1}{2}h_{it}\frac{g_{it}'}{w_{it}}.$$

For instance, for the probit model, $y_{it}^{*} = y_{it} - h_{it}\eta_{it}\Phi_{it}(1-\Phi_{it})/(2\phi_{it})$; while for the logit model, $y_{it}^{*} = y_{it} + h_{it}(0.5 - \Lambda_{it})$. (And ML estimates are obtained for $y_{it}^{*} = y_{it}$.)

B.3 Existence under concordance

In the cases of concordant panel units, $\sum_t y_{it} = 0$ and $\sum_t y_{it} = T$, a finite ML estimator for the individual effect α_i does not exist. We consider the case $\sum_t y_{it} = T$ for the BRFE estimator:

$$s^{BRFE}(\alpha_i) = \sum_{t=1}^{T}\frac{g_{it}}{G_{it}} + \frac{1}{2}h_{it}\frac{g_{it}'}{g_{it}}.$$

Since G_{it} is a smooth, strictly increasing distribution function, $\ln g_{it}$ is globally concave: its first derivative, g_{it}'/g_{it}, has a unique root with g_{it}'/g_{it} being positive for small values of η_{it} ($\eta_{it}\to-\infty$) and negative for large values of η_{it} ($\eta_{it}\to\infty$). This implies that the second term on the right-hand side of the equation, $\sum_t \frac{1}{2}h_{it}g_{it}'/g_{it}$, is positive for small values of α_i ($\alpha_i\to-\infty$) and negative for large values of α_i ($\alpha_i\to\infty$), as $h_{it}\in(0,\,1]$. Meanwhile, the first term, $\sum_t g_{it}/G_{it}$, tends to zero for small α_i ($\alpha_i\to-\infty$) and to a positive

constant or positive infinity for large α_i ($\alpha_i \to \infty$). Therefore, because $s^{BRFE}(\alpha_i)$ is continuous, there must exist a $\widehat{\alpha}_i^{BRFE}$ such that $s^{BRFE}(\widehat{\alpha}_i^{BRFE}) = 0$.

A detailed example was given in Section 2 for $G_{it} = \Phi_{it}$. For logit, $G_{it} = \Lambda_{it}$, and

$$s^{BRFE}(\alpha_i) = \sum_{t=1}^{T} (1 - \Lambda_{it}) + \frac{1}{2} h_{it} \left(1 - 2\Lambda_{it}\right).$$

As $\alpha_i \to -\infty$, the first term $\sum_t 1 - \Lambda_{it}$ tends to T and the second to $\sum_t h_{it}/2 > 0$; thus, $lim_{\alpha \to -\infty} s^{BRFE}(\alpha_i) > 0$. As $\alpha_i \to \infty$, the first term $\sum_t 1 - \Lambda_{it}$ tends to 0 and the second to $-\sum_t h_{it}/2 < 0$; thus, $lim_{\alpha \to -\infty} s^{BRFE}(\alpha_i) < 0$. Therefore, $s^{BRFE}(\widehat{\alpha}_i^{BRFE}) = 0$ exists.

Existence for the case $\sum_t y_{it} = 0$ can be examined using the same arguments.