WP 18/23

Predicting fixed effects in panel probit models

Johannes S. Kunz; Kevin E. Staub and Rainer Winkelmann

August 2018

Predicting fixed effects in panel probit models

Johannes S. Kunz
Monash University

Kevin E. Staub
University of Melbourne, IZA

Rainer Winkelmann
University of Zurich, IZA

July 13, 2018

Preliminary

Abstract

We present a method to estimate and predict fixed effects in a panel probit model when N is large and T is small, and when there is a high proportion of individual units without variation in the binary response. Our approach builds on a bias-reduction method originally developed by Kosmidis and Firth (2009) for cross-section data. In contrast to other estimators, our approach ensures that predicted fixed effects are finite in all cases. Results from a simulation study document favorable properties in terms of bias and mean squared error. The estimator is applied to predict period-specific fixed effects for the extensive margin of health care utilization (any visit to a doctor during the previous three months), using German data for 2000-2014. We find a negative correlation between fixed effects and observed characteristics. Although there is some within-individual variation in fixed effects over sub-periods, the between-variation is four times as large.

Keywords: Perfect prediction; Bias reduction; modified score function;

JEL classification: I11; I18; C23; C25.

1 Introduction

This paper addresses the prediction of individual-specific fixed effects in a binary panel model when the individual dimension N is large and the time dimension T is small. Predictions can serve to rank individuals, or to group them into "high", "middle", "low" prevalence categories, or to use them in further correlation analyses. Examples from the literature where linear panel models are used to predict fixed effects include those of neigborhoods (Chetty and Hendren, 2015), teachers (Chetty, Friedman and Rockoff, 2014), workers and firms (Card, Heining and Kline, 2013), judges (Abrams, Bertrand and Mullainathan, 2012), and doctors and hospitals (Street et al., 2014), to name but a few.

Clearly, such predictions are noisy for short panels. In addition, if based on maximum likelihood (ML) estimation of nonlinear binary response models (in contrast to the linear model) two further problems arise. First, predictions have a small sample bias which also increases their mean squared error. And second, they may not even exist: in applications there are typically many units with identical responses in all time periods. For those units, ML predictions of fixed effects are $\pm\infty$, depending on whether all outcomes are zero or one. Since the model predicts the outcomes without error, this situation has been refered to in the literature as "perfect prediction" (see. e.g. Maddala, 1983), a somewhat misleading term in our context, because the predictions of the fixed effects do not exist then.

The goal of this paper is to study two estimation methods that remove the first-order, $O(1/T)$, bias of fixed-effects predictions in the probit model when the incidence of "perfect prediction" is high. These are: (i) BR, the bias-reduction estimator of Firth (1993) and Kosmidis and Firth (2009); (ii) HS, the penalised likelihood estimator (with a penalty based on the Hessian and score) of Bester and Hansen (2009). We chose these two approaches because they are relatively simple to implement, and because they avoid computation of the non-existing ML estimators.

The BR method has been developed specifically for the distribution class of linear exponential families, of which the probit model is a member. In the probit case, it shrinks the fixed effects predictions toward zero. It thereby also reduces the variance of the predictions and unambiguouly improves the mean squared error. The HS method was originally devised to remove incidental parameter bias for any non-linear objective function. To the best of our knowledge, these methods have not been considered to date with the specific aim of estimating and predicting fixed effects in binary response models.

Our main findings are: (i) We show analytically that the BR estimator always delivers finite estimates of the fixed effects. In a variety of simulation settings, the BR estimator turns out to be surprisingly good. The BR estimator of the common parameters (covariate slopes) is also bias-reducing. (ii) The HS estimator reduces first order bias as well. However, we show analytically for $T = 2$ and by simulations for $T > 2$ that it does not deliver finite estimates of the fixed effects for units with perfect prediction. This is not a contradiction, since the probability of perfect prediction vanishes exponentially as T increases, so the $O(1/T)$ bias does not arise due to perfect prediction. Overall then, there is a clear recommendation for the use of the BR estimator in applications where there is a high prevalence of perfect prediction.

The paper adds to, but is different from, a relative recent literature on panel binary probit and logit models, where the focus was on bias correction for β, in order to address the incidental parameters problem (see e.g. Lancaster, 2000, Woutersen, 2004, Arellano and Hahn, 2006, and Arellano and Honoré, 2001). There also has been some work on estimation of functionals of the distribution of the fixed effects. Hahn and Newey (2004), Fernandez-Val (2009) and Dhaene and Jochmans (2015) all consider estimation of average marginal effects in panel models.

In the next section, we formally introduce the problems of first-order bias and perfect prediction in the context of binary response fixed effects panel data models. We then present the BR estimator, which solves these problems, and the HS estimator, which does not. In Section 3, we set up Monte Carlo simulations for predicting the α_i's across a number of differently shaped distributions from which the true α_i's are drawn. The BR estimator performs well in these simulations, in terms of bias as well as mean squared error. The simulations also indicate that the BR estimator delivers reliable estimates of β in short panels.

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 fixed effects in a model, where the binary variable "any doctor visit during the last three months (yes/no)" is regressed on a number indicators of socio-economic status and health status. Section 5 concludes.

2 Econometric methods

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

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

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. Typically, N is large and T is small. This model does not make any assumption on the distribution of α_i, nor does it require the α_i's to be exogenous (uncorrelated with x_{it}).

As noted in the literature, the maximum likelihood estimator (MLE), $(\hat{\alpha}, \hat{\beta}) = (\hat{\alpha}_1, \ldots, \hat{\alpha}_N, \hat{\beta})$ has a number of deficiencies in this case. First, $\hat{\beta}$ is inconsistent. This is a manifestation of the incidental parameters problem. Abrevaya (1997) shows for the panel logit model with $T = 2$, that plim $\hat{\beta} = 2\beta$. 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, $\hat{\alpha}_i$ is technically inconsistent for fixed T and $N \to \infty$, and may have poor small sample properties for small T. Third, $\hat{\alpha}_i$ does not exist if $\sum_t y_{it} = 0$ or if $\sum_t y_{it} = T$. This is called the "perfect prediction problem" (Maddala, 1983).

We are here mostly concerned with the second and third issues, the small sample bias and the potential non-existence of $\hat{\alpha}_i$. Our main approach uses an estimator developed by Kosmidis and Firth (2009) (see also Firth, 1993) for cross-section data and adapts it to the estimation of fixed effects in a probit panel data model. We show that the resulting estimator is immune to the perfect prediction problem. It also is relatively easy to compute, as it can be obtained using an iteratively weighted least squares estimator (Kosmidis and Firth, 2009).

2.1 First-order bias

Non-linear ML estimators have a finite sample bias. Considering the T dimension, the bias can be split up into an $O(T^{-1})$ term, the first-order bias, and higher-order terms that converge in probability at a faster rate. 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:

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

where $\alpha_i = \tilde{\alpha}_i + \bar{x}_i'\gamma$ and Φ denotes the standard normal distribution function. In this case, $\bar{y}_i$ is a consistent estimator for $\mu_i = \Phi(\alpha_i)$. A standard Taylor expansion gives

$$\hat{\alpha}_i - \alpha_i \approx \frac{\partial \Phi^{-1}}{\partial \mu_i}(\bar{y}_i - \mu_i) + \frac{1}{2}\frac{\partial^2 \Phi^{-1}}{\partial \mu_i^2}(\bar{y}_i - \mu_i)^2$$

The second derivative is $\alpha_i/\phi(\alpha_i)^2$, $\mathrm{E}[(\bar{y}_i - \mu_i)^2] = \mu_i(1-\mu_i)/T = \Phi(\alpha_i)(1-\Phi(\alpha_i))/T$ and therefore

$$\mathrm{E}(\hat{\alpha}_i - \alpha_i) \approx \frac{1}{2T}\frac{\alpha_i \Phi(\alpha_i)(1-\Phi(\alpha_i))}{\phi(\alpha_i)^2} \tag{3}$$

The 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.

2.2 Perfect prediction

Perfect prediction in the general model (1) means that the first-order conditions for the ML estimator do not have a finite solution. This problem can arise with any ill-designed x-vector, but the concern here is with perfect prediction arising due to the presence of individual specific effects. The $K + N$ first-order conditions are:

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

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

where $\eta_{it} = \alpha_i + x_{it}'\beta$, and K is the number of regressors in x_{it}. Suppose that $y_{i1} = \ldots = y_{iT} = 0$ for some i. Then (5) 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 $y_{i1} = \ldots = y_{iT} = 1$ for some i, (5) 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, $\hat{\alpha}_i$ will tend to minus infinity, while it will tend to plus infinity in the second. Units i where observations are either all equal to zero or all equal to one are called *concordant.*

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 (4). Perfectly predicted observations do not contribute to the (concentrated) score, since

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

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

The problem of perfect prediction is most severe for small values of T: 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),

$$\Pr\left(\sum_{t=1}^{T} y_{it}=0\right)+\Pr\left(\sum_{t=1}^{T} y_{it}=T\right)=(1-\Phi(\alpha_i))^T+\Phi(\alpha_i)^T \qquad , \tag{8}$$

Hence, the probability of perfect prediction 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 perfect prediction.

2.3 Bias reduction

Firth (1993) considered the first-order bias of maximum likelihood estimators in the context of linear exponential family models. He showed that for models with canonical link function, the first-order bias can be removed by maximising a modified log-likelihood function that includes a penalty term based on the log-determinant of the information matrix, equal to Jeffreys prior (see also Ehm, 1991). For binary response models, the logit model provides the canonical link.

For linear exponential family models with non-canonical link function – including, for example, the probit model – such a modified objective function does not exist. Instead, as shown by Kosmidis and Firth (2009) and Kosmidis (2007), it is possible to make an adjustment to the score function that achieves the same first-order bias reductions for the MLE. The adjusted score for the probit panel model is

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

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 $NT \times K$ matrix of the K regressors, and W is the $NT \times NT$ diagonal matrix with typical element $w_{it} = \phi(\eta_{it})^2/[\Phi(\eta_{it})(1-\Phi(\eta_{it}))]$. The β_k-terms of the score vector are adjusted accordingly. From (9), it can be seen that if we define

$$y^*_{it} = y_{it} - \frac{1}{2}h_{it}\eta_{it}\frac{\Phi(\eta_{it})(1-\Phi(\eta_{it}))}{\phi(\eta_{it})}, \tag{11}$$

then (9) is in the form of the standard MLE score $s^{ML}(\alpha_i)$, where y_{it} is replaced by the pseudo-response y^*_{it}.

It is therefore possible to solve the adjusted first-order conditions using an iteratively re-weighted least squares (IWLS) algorithm (McCullagh and Nelder, 1989, Kosmidis and Firth, 2009), which makes this approach attractive from a computational point of view. For implementation, pseudo-responses are constructed, at iteration s, using existing estimates from the previous iteration $s-1$ to replace the unknown quantities h_{it} and η_{it} with estimates $\hat{h}_{it}(\hat{\alpha}^{s-1}, \hat{\beta}^{s-1})$ and $\hat{\eta}_{it}(\hat{\alpha}^{s-1}, \hat{\beta}^{s-1})$. An implementation in Stata is available from the authors.

To examine whether the estimator based on the modified score (9) exists in the cases of perfect prediction, we consider the case where all observations of a unit i are equal to one, $\sum_t y_{it} = T$. Then, we can write (9) as

$$s^{BR}(\alpha_i) = \left(\sum_t \frac{\phi(\eta_{it})}{\Phi(\eta_{it})}\right) - \frac{\alpha_i}{2}\left(\sum_t h_{it}\right) - \frac{1}{2}\left(\sum_t h_{it}x'_{it}\beta\right) = g_1(\alpha_i) - \alpha_i g_2(\alpha_i) - g_3(\alpha_i). \tag{12}$$

When α_i becomes very 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, this implies that it has a finite solution. Similar arguments can be made to show that a solution exists for the other perfect prediction case, $\sum_t y_{it} = 0$, as well. This echos earlier results by Heinze and Schemper (2002) who showed that the Firth method for bias reduction solves the perfect prediction problem for the cross-sectional Logit model.

An interesting example is the case with no time varying regressors, i.e. the constants-only model. The first perfect prediction case with $\bar{y}_i = 1$ gives

$$\alpha_i = 2T \sum_t \frac{\phi(\alpha_i)}{\Phi(\alpha_i)}, \tag{13}$$

and the second case, $\bar{y}_i = 0$,

$$\alpha_i = -2T \sum_t \frac{\phi(\alpha_i)}{1 - \Phi(\alpha_i)}, \tag{14}$$

where we used the fact that with only a constant in the model, $\sum_t h_{it} = 1$. The two cases only differ in the sign. For $T = 2, 3, 4$ the estimates for α_i are about ± 1.06, ± 1.24, ± 1.37, respectively.

The associated predicted probabilities $\widehat{\Pr(y_{it} = 1)} = \Phi(\hat{\alpha}_i)$ are about 0.144, 0.107, and 0.086 when $\bar{y}_i = 0$ and $T = 2, 3, 4$. These estimates reflect the shrinkage away from the lower bound of 0, which are "built into" this estimator. The amount shrinkage decreases with increasing sample size. The MLE 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 or three periods is deemed impossible.

2.4 HS Estimator

An alternative penalised likelihood estimator has been proposed by Bester and Hansen (2009). We consider their "HS penalty" constructed using the sample Hessian and outer product of scores, since an approach based on analytical expectations is not feasible in our context. Although the HS estimator is very general and therefore applicable to a broad class of models, concrete implementations require programming of objective function-specific penalty terms. Here, we consider the case of the panel probit model with scalar fixed effects. In this case, the objective function is

$$Q^{HS}(\beta, \alpha_1, \ldots, \alpha_n) = \sum_{i=1}^{n} Q_i^{HS}$$

where

$$Q_i^{HS} = \sum_{t=1}^{T} [y_{it} \log(\Phi(\eta_{it})) + (1 - y_{it}) \log(1 - \Phi(\eta_{it}))] - \left(\frac{1}{2} \frac{\sum_t v_{it}^2}{\sum_t (-v_{it}^{\alpha})} + \frac{1}{2} \right). \tag{15}$$

and $\eta_{it} = \alpha_i + x_{it}'\beta$ as before.

The first sum on the right-hand side of (15) is the conventional log-likelihood contribution of unit i for a panel probit model. The remainder is a penalty term which depends on the relative discrepancy between the outer product of the score, given by equation (5), and the negative of the Hessian, $\sum_t \partial v_{it}/\partial \alpha_i = \sum_t v_{it}^{\alpha}$, both with respect to α_i. The adjustment is made only for the log-likelihood derivatives with respect to the fixed effects, as only those are estimated from a small number of T observations. It is not necessary to adjust the score for the common parameter β since N is large.

With perfect prediction, e.g. $y_{i1}, ...y_{iT} = 0$, we obtain

$$\begin{aligned} v_{it} &= -\frac{\phi(\eta_{it})}{1 - \Phi(\eta_{it})} = -\lambda_{it}, \\ v_{it}^{\alpha} &= -\lambda(\eta_{it})[\lambda(\eta_{it}) - \eta_{it}] = -\lambda_{it}^{\alpha}, \end{aligned}$$

where we use the shorthand notation $\lambda_{it} = \lambda(\eta_{it})$ to denote the inverse Mills ratio and $\lambda_{it}^{\alpha} = \partial\lambda(\eta_{it})/\partial\alpha_i$ its derivative. We can rewrite unit i's contribution to the penalised log-likelihood as

$$Q_i^{HS} = \sum_{t=1}^{T} \log(1 - \Phi(\eta_{it})) - \frac{\sum_t \lambda_{it}^2}{2 \sum_t \lambda_{it}^{\alpha}} + \frac{1}{2},$$

with associated score for α_i

$$s^{HS}(\alpha_i) = \sum_{t=1}^{T} -\lambda_{it} - \frac{\sum_t \lambda_{it}\lambda_{it}^{\alpha}}{\sum_t \lambda_{it}^{\alpha}} + \frac{1}{2} \frac{(\sum_t \lambda_{it}^2)(\sum_t \lambda_{it}^{\alpha\alpha})}{(\sum_t \lambda_{it}^{\alpha})^2}.$$

Since $\lambda_{it} > 0$, $0 < \lambda_{it}^{\alpha} < 1$, and $\lambda_{it}^{\alpha\alpha} = \partial\lambda_{it}^{\alpha}/\partial\alpha > 0$ (see, for instance, Heckman and Honoré, 1990, p. 1130), only the third term on the right-hand side provides a positive contribution to the score. However, this term may in general be too small to offset the negative contributions of the first two terms. As an illustration, consider the simple case where $\lambda_{it} \equiv \lambda_i$ for all t. The HS score thus simplifies to

$$\begin{aligned} s^{HS}(\alpha_i) &= -T\lambda_i - \frac{\lambda_i}{\lambda_i^{\alpha}} + \frac{\lambda_i^2 \lambda_i^{\alpha\alpha}}{2\lambda_i^{\alpha 2}} \\ &= -\left(T - \frac{1}{2}\right)\lambda_i - \lambda_i + \frac{\lambda_i^3(\lambda_i^{\alpha} - 1)}{2\lambda_i^{\alpha 2}} < 0, \end{aligned}$$

where the second equality used $\lambda_i^{\alpha\alpha} = 2\lambda_i\lambda_i^{\alpha} - \lambda_i^{\alpha}\eta_i - \lambda_i$ and therefore $\lambda_i^2\lambda_i^{\alpha\alpha} = \lambda_i\lambda_i^{\alpha 2} + \lambda_i^3(\lambda_i^{\alpha} - 1)$.

Thus, we see that there are cases where the HS estimator for α_i is not finite. We show in the Appendix that for $T = 2$ no finite value of $\hat{\alpha}_i$ may satisfy $s^{HS}(\alpha_i) = 0$ over a substantial region of $(x'_{i1}\beta, x'_{i2}\beta) \in R^2$. In our simulation study in the next section, we also considered several $T > 2$. We did not encounter a single case where the HS estimator $\hat{\alpha}_i$ existed for perfectly predicted cross-sectional units.

3 Monte Carlo evidence

3.1 Experimental design

The primary aim of our Monte Carlo experiment is to investigate the suitability of the approaches discussed in the previous section for predicting fixed effects, in panel probit models with a small to moderate number of time periods and a high prevalence of perfect prediction.

– – – – – – – – – – Figure 1 about here – – – – – – – – – –

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 1. 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 correspond to a "random effects" model as the distribution of α_i does not depend on the regressor. This allows us to focus on biases purely related to small samples and the perfect prediction problem, whereas additional dependence on regressors would exacerbate or attenuate those biases.

Below, we report simulation results for $N = 100$ and $T \in \{2, 4, 8, 12\}$. For each of the four distributions from Figure 1, 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 $r = 1, \ldots, 500$ denotes Monte Carlo replications.

In each of the 500 replications, we keep track of the fraction of perfectly predicted, or concordant, observations, i.e., 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 individuals over the 500 replications amounts to 24 per cent. This fraction is somewhat lower for the beta (15 per cent) and Bernoulli (20 per cent) distributions, and higher for the normal distribution (28 per cent). Plots and summary statistics of our results are based on all finite estimates: since the maximum likelihood estimator and the HS penalised likelihood estimators of α_i do not exist for concordant observations, the effective replication sample size is below 500 in these cases. For example, for $T = 4$, the share of replications for a particular i with concordant observations ranges from 5.6 per cent to 91.6 per cent. For increasing T, the incidence of perfect prediction decreases.

3.2 Results

We start our presentation with some summary statistics of the discrepancy between the predicted fixed effects and their true values. Table 1 lists, for the three considered estimation methods and for each of the four distributions of α_i, the estimated mean and standard deviations of the $N = 100$ predictions, averaged over 500 replications. These can be benchmarked against the mean and standard deviation of the (once) simulated α_i's, for instance -0.030 and 0.451 in the case of the Bernoulli data generating process.

– – – – – – – – – – Table 1 about here – – – – – – – – – –

The best-performing estimator in terms of mean is BR. Note that for the ML and HS estimators, only finite fixed effects predictions can be used, i.e. perfectly predicted units have to be excluded. The consequences depend on the distribution from which the true α_i's are drawn. In the case of the beta and normal distributions, for instance, the averaged finite predictions of HS and ML tend to lie below the average of the true individual specific fixed effects. Regarding variation, we find that for short panels (in particular for $T = 2$), the BR estimator underestimates the true variance of the fixed effects somewhat. This is a result of the implied shrinkage, and it becomes very minor for $T = 8$ or $T = 12$.

Table 2 presents means and standard deviations of the estimated $\hat{\beta}$ across different distributions of α_i and different number 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 per cent for T equal to 2, 4, 8 and 12, respectively. The HS estimator reduces the bias, although not very effectively for small T. With T equal to 2, 4 and 8 the biases are still about 100, 20 and 5 per cent, respectively. In contrast, we find that BR removes much of the bias in $\hat{\beta}$. Even for $T = 2$, only a bias of about -10 per cent is left. At $T = 4$ the bias falls to between 0.6 per cent (Bernoulli) and 2.3 per cent (normal), and for larger T the bias is virtually zero.

− − − − − − − − − − Table 2 about here − − − − − − − − − −

3.3 Perfect prediction, bias, and mean squared error

In this section, we show how the prediction quality, measured in terms of bias and mean squared error (MSE), varies with the prevalence of perfect prediction. As noted before, the BR is unique among the three estimation approaches, in that it provides a finite prediction of fixed effects for all observation units, regardless of whether they are concordant (i.e. based on perfect prediction) or discordant. Formally, one could conclude that HS and ML have infinite bias and infinite MSE, whereas both are finite for the BR estimator. Hence, in this sense, it clearly dominates HS and ML for the purpose of predicting fixed effects.

In addition, and alternatively, one can compare bias and MSE for the subset of discordant observations, and this is what Figures 2-5 do, for $T = 4$ and $T = 12$, respectively, and for bias and MSE. The bias for each α_i is obtained as

$$\widehat{\text{Bias}}(\hat{\alpha}_i) = \left[\sum_{r=1}^{500} d_{ir}\right]^{-1} \sum_{r=1}^{500} d_{ir} \left(\hat{\alpha}_i^{(r)} - \alpha_i\right).$$

and the MSE is given by

$$\widehat{\mathrm{MSE}}(\hat{\alpha}_i) = \left[\sum_{r=1}^{500} d_{ir}\right]^{-1} \sum_{r=1}^{500} d_{ir} \left(\hat{\alpha}_i^{(r)} - \alpha_i\right)^2 .$$

To account for perfect prediction, concordant observations are given a weight of zero ($d_{ir} = 0$).

In principle, bias and MSE can be obtained for each α_i, $i = 1, \ldots, 100$. In order to highlight the consequences of perfect prediction in this context, we sorted the 100 values for bias and MSE by the propensity for perfect prediction, which is a function of the mean dependent variable of unit i across the 500 replications:

$$\bar{s}_i = \frac{1}{500} \frac{1}{T} \sum_{r=1}^{500} \sum_{t=1}^{T} y_{it}^{(r)}$$

Both $s_i = 1$ and $s_i = 0$ result in perfect prediction, and thus values of $\bar{s}_i$ close to these bounds indicate a high prevalence of perfect prediction across replications. In Figures 2-5, we group units into deciles according to their propensity to perfect prediction, and plot the mean bias and MSE in each decile. Circles indicate bias and MSE of the maximum likelihood estimator, squares that of HS and triangles that of BR. We also add results for the BR method that uses all observations (diamonds).

– – – – – – – – – – Figure 2 about here – – – – – – – – – –

Figure 2 shows the bias results for $T = 4$. Four observations stand out. First, the bias can be substantial, and reach, in some extreme cases (uniform and Bernoulli) up to a standard deviation of the underlying distribution of α_i. Second, for the subset of discordant pairs, the average predictions from BR, HS and ML are similar, although BR always has slightly less bias. Third, bias is an increasing function of the propensity for perfect prediction, and all three methods perform poorly if this propensity becomes large. Fourth, the BR approach using predictions for *all* units performs very well in all cases.

– – – – – – – – – – Figure 3 about here – – – – – – – – – –

Figure 3 repeats the bias analysis for $T = 12$. As expected, the bias becomes much smaller overall, which we account for by adjusting the y-scale in the panels. Patterns also appear more variable, albeit in the small scale, but the BR approach using all observations again outperforms the other three prediction approaches by far.

Figures 4 and 5 show the results for the MSE computations. Among the comparisons of fixed effect predictions for discordant units only, the BR approach always yields the lowest MSE, followed by HS. ML is the worst. The dominance of BR is owed to its shrinking property, which reduces variance without introducing much bias, as we have seen before. However, the real strengths of BR is that it gives predictions for all units, not only discordant ones. The MSE for all units would be infinity for both the ML and the HS estimators, yet, the MSE for the BR approach has a comparable fit to the discordant-only MSEs and decreases quickly from $T = 4$ to $T = 12$ (recall that the y-axis has been rescaled). Overall, this strengthens our conclusion that the BR estimator is the strongly preferred approach in the presence of perfect prediction.

− − − − − − − − − − Figures 4 and 5 about here − − − − − − − − − −

4 Application to the determinants of doctor visits in Germany

We apply the new estimator to predict fixed effects in a demand model for doctor visits in Germany. The analysis uses 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 did take place ($anyvisit = 1$), or did not take place ($anyvisit = 0$), during the three-months period prior to the annual interview. We express the probability $\Pr(anyvisit = 1)$ as a panel probit model with socio-economic determinants and fixed effects as explanatory variables, and apply the BR estimator to obtain predictions of the fixed effects. 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 and 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).

Our analysis sample was generated as follows: we restrict the sample to those aged 20-65 at the time of the interview, in order to allow for meaningful labor market effects of full-time ot part-time work, as well as earnings. We drop observations with missing values on marital status, disability status and self-assessed health. We retain a balanced panel of 55,230 person-year observations, representing 1997 women and 1685 men. Finally, we split the observation period into three five-year intervals, 2000-2004, 2005-2009, and 2010-2014, and estimate separate models for the three subperiods. We thereby obtain three separate predictions of fixed effects for each individual. One goal of the analysis is to use these predictions to assess their stability over time.

Table 3 provides some summary statistics, separately for men and women and for the three time periods. The share of men with at least one visit increases from 55.8% for the years 2000-2004 to 63.7% for the years 2010-2014. Since this is a balanced panel, the average age increases by exactly five years for each five year period by construction. The increasing age is associated with a worsening of self-assessed health (SAH) over time. SAH is measured on a five-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). Similar patterns for SAH and *anyvisit* are found for women. The main difference is their overall higher prevalence of doctor visits. Our panel analysis will allow us to estimate the effect of SAH (and other variables, including age, employment and income) on the probability of at least one doctor visit per three-months period, from within-subject variation in those characteristics.

− − − − − − − − − − Table 3 about here − − − − − − − − − −

The penultimate row of the sub-panels of Table 3 states the proportion of perfectly predicted outcomes by gender and period. This proportion varies from a minimum of 29 percent to a maximum of 46 percent. There is an interesting pattern here, as the proportion with perfect prediction increases over time for both men and women, and is overall higher for women. Before, we noted the same pattern for the proportion of any visits. This is not a coincidence: in our application, most perfectly predicted observations satisfy $\bar{y}_i = 1$, i.e. they go in the direction of any utilization. The opposite case, $\bar{y}_i = 0$, is relatively less common. Hence, all factors that increase $\Pr(anyvisit_{it} = 1)$ will also tend to increase the prevalence of perfect prediction. We see in the last row that the proportion of individuals with at least one visit is always smaller in the non-perfect prediction subsample than in the overall sample. This is interesting, because it means that dropping perfectly predicted observations, as is required for the brute-force estimation of the panel probit model with fixed effects, suffers from an endogenous selection problem, on top of the incidental parameters bias. The BR estimator keeps all observations and therefore avoids this kind of issue.

4.1 Estimation results

A total of six models were estimated, one for each five-year period, separately by gender. Since the focus of this paper is on predicting fixed effects, we display in Table 4 only a subset of the regression results, based on the period 2010-2014 (the others are available on request). The main takeaway is that the brute force probit coefficients, based on a model with dummies for each person without further adjustment, tend to be biased upward: this reflects a possible combination of the incidental parameters bias and a selection bias, since about 40 percent of observations are concordant and need to be dropped (3370 out of 8425 for men, 4555 out of 9985 for women). The reduced sample size also leads to estimated standard errors that are correspondingly higher for the brute-force probit model relative to the BR probit model.

− − − − − − − − − − Table 4 about here − − − − − − − − − −

We note that a worse self-assessed health (which is an increase in the SAH variable as coded here) increases the probability of any visit for both men and women. For women, statistically significant effects can be found as well for *disability* and *fulltime work*.

4.2 Predicting fixed effects

Once a BR probit model has been estimated, predictions for the individual specific fixed effects are directly available for further analysis, e.g. for ranking of individuals by their underlying propensity. Figure 5 gives an example for the distribution of the fixed effects for men in the 2010-2014 period. The histogramm appears approximately normal distributed, although this of course does not need to be that way.

− − − − − − − − − − Figure 5 about here − − − − − − − − − −

One use of the predictions is to study their stability over time. For each person (e.g. for $N = 1684$ men), we obtain $T = 3$ distinct predictions. The total variance for the $NT = 5052$ fixed effects is given by 1.046; When we decompose this variance, we find that the between variance is 0.762, with the remaining 0.284 reflecting the within variance. Hence, there is some instability of the fixed effects over time; however, it only contributes about 25 percent to the overall variance. Hence, the fixed effects capture mostly between person differences in the propensity of health care utilization.

Another way to look at stability over time is in terms of bivariate scatter plot, see Figure 6. Again, we find a substantial positive correlation in estimated fixed effects for the same person between adjacent sample periods.

− − − − − − − − − − Figure 6 about here − − − − − − − − − −

Finally, one can use predictions to explore any time-invariant relationships between the observed and unobserved components of the demand for health services. For the observed components (or propensity), define $\hat{\eta}_i = \bar{x}_i\hat{\beta}$, where $\bar{x}_i$ are the average characteristics and $\hat{\beta}$ is the coefficient vector from the BR probit regression. For men in 2010-2014, the correlation between $\hat{\alpha}_i$ and $\hat{\eta}_i$ equals -0.35; Hence, individual difference in observed factors tend to be associated with unbservables that move in the other direction. Ignoring this correlation (such as in a pooled probit model) would understate the importance of either of the two.

5 Conclusions

This paper studied bias-reduction approaches to address perfect prediction problems in fixed-T panel probit models for binary responses with fixed effects, and applied them to study the determinants of health care utilization in Germany. We advocated an estimator based on Kosmidis and Firth (2009), which has not been adapted to the context of panel data so far, and for which we showed that it always produces finite predictions of all fixed effects.

Perfect prediction is a problem which is very common in applications, especially in short and very short panels. In the data of our application—a balanced panel covering a five-year period—about 40 percent of the observations were concordant and would have led to infinite estimates for the corresponding fixed effects had we used conventional panel data model estimators or bias-corrected estimators. While the incidence of the type of perfect prediction we discussed in this paper lessens with increasing T, a substantial incidence of perfect prediction can persist even in longer panels if the outcome is a rare event.

We focussed on the probit model as it is a common choice in empirical work, but the advocated approach is applicable to a number of other binary response models as well. More broadly, the estimator can be extended to other nonlinear fixed effects panel models which suffer from perfect prediction, such as models for ordered and count data.

References

Abrams, David S., Marianne Bertrand and Sendhil Mullainathan. 2012. Do Judges Vary in Their Treatment of Race? Journal of Legal Studies 41(2):347-383.

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

Arellano, Manuel and Jinyong Hahn. 2016. A likelihood-Based Approximate Solution to the Incidental Parameter Problem in Dynamic Nonlinear Models with Multiple Effects. Global Economic Review 45(3):251-274.

Arellano, Manuel and Bo Honoré. 2001. Panel Data Models: Some Recent Developments. in: Handbook of Econometrics, Chapter 53.

Bester, Alan C. and Christian Hansen. 2009. A penalty function approach to bias reduction in nonlinear panel models with fixed effects. Journal of Business & Economic Statistics 27(2):131-148.

Bloom, Nicholas, Carol Propper, Stephan Seiler and John Van Reenen. 2015. The impact of competition on management quality: evidence from public hospitals. Review of Economic Studies 82(2):457-489.

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.

Card, David, Jrg Heining and Patrick Kline. 2013. Workplace heterogeneity and the rise of West German wage inequality. Quarterly Journal of Economics 128(3):967-1015.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. 2014. Measuring the impacts of teach- ers II: Teacher value-added and student outcomes in adulthood. American Economic Review 104(9):2633-2679.

Chetty, Raj and Nathaniel Hendren. 2015. The impacts of neighborhoods on intergenerational mobility: Childhood exposure effects and county-level estimates. NBER Working Paper No. 23002

Cox, David R. and Emily J. Snell. 1971. On test statistics calculated from residuals. Biometrika 58(3):589-594.

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

Ehm, Werner. 1991. Statistical problems with many parameters: Critical quantities for approximate normality and posterior density based inference. Habilitationsschrift, University of Heidelberg .

Fernndez-Val, Ivn. 2009. Fixed effects estimation of structural parameters and marginal effects in panel probit models. Journal of Econometrics 150(1):71-85.

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

Greene, William 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, Jinyong and Whitney Newey. 2004. Jackknife and analytical bias reduction for nonlinear panel models. Econometrica 72(4):1295-1319.

Heckman, James J. and Bo E. Honoré. 1990. The empirical content of the Roy model. Econometrica 58(5):1121-1149.

Heinze, Georg and Michael Schemper. 2002. A solution to the problem of separation in logistic regression. Statistics in Medicine 21(16):2409-2419.

Kosmidis, Ioannis. 2007. Bias Reduction in Exponential Family Nonlinear Models. Doctoral thesis, The University of Warwick.

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

Lancaster, Tony. 2000. The incidental parameter problem since 1948. Journal of Econometrics 95 (2), 391-413

Maddala, Gangadharrao S. 1983. Qualitative and limited dependent variable models in econometrics. Cambridge: Cambridge University Press.

McCullagh, Peter. and John A. Nelder. 1989. Generalized Linear Models. 2nd ed. London, UK: Chapman and Hall/CRC.

Mullahy J. 1986. Specification and testing in some modified count data models. Journal of Econometrics 33: 341-365.

Street, Andrew, Nils Gutacker, Chris Bojke, Nancy Devlin and Silvio Daidone. 2014. Variations in outcome and costs among NHS providers for common surgical procedures: econometric analyses of routinely collected data. Health Services and Delivery Research 2(1).

Wagner, G.G., J.R. Frick and J. Schupp. 2007. The German Socio-Economic Panel Study (SOEP) scope, evolution and enhancements, Schmollers Jahrbuch, 127 , 139-169.

Winkelmann, Rainer. 2004. Co-payments for prescription drugs and the demand for doctor visitsevidence from a natural experiment. Health Economics 13, 1081-1089.

Woutersen, Tiemen. 2004. Bayesian Analysis of Misspecif ied Models with Fixed Effects, 2004, in Advances in Econometrics: Maximum Likelihood Estimation of Misspecified Models: Twenty

Years Late . Edited by T. B. Fomby and R. Hill, Emerald Group Publishing, UK.

Figures and Tables

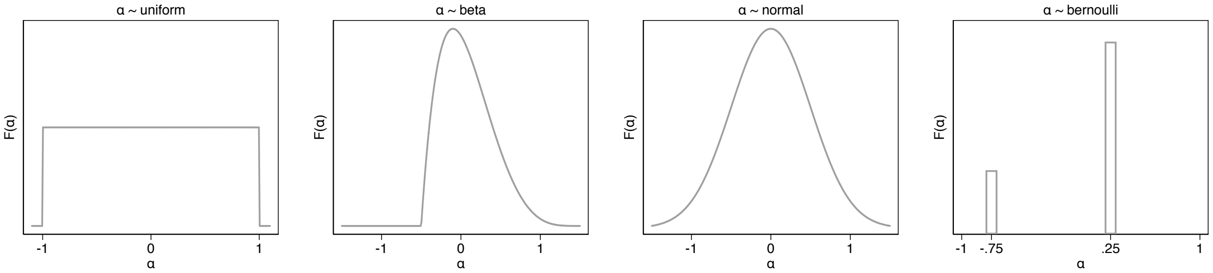
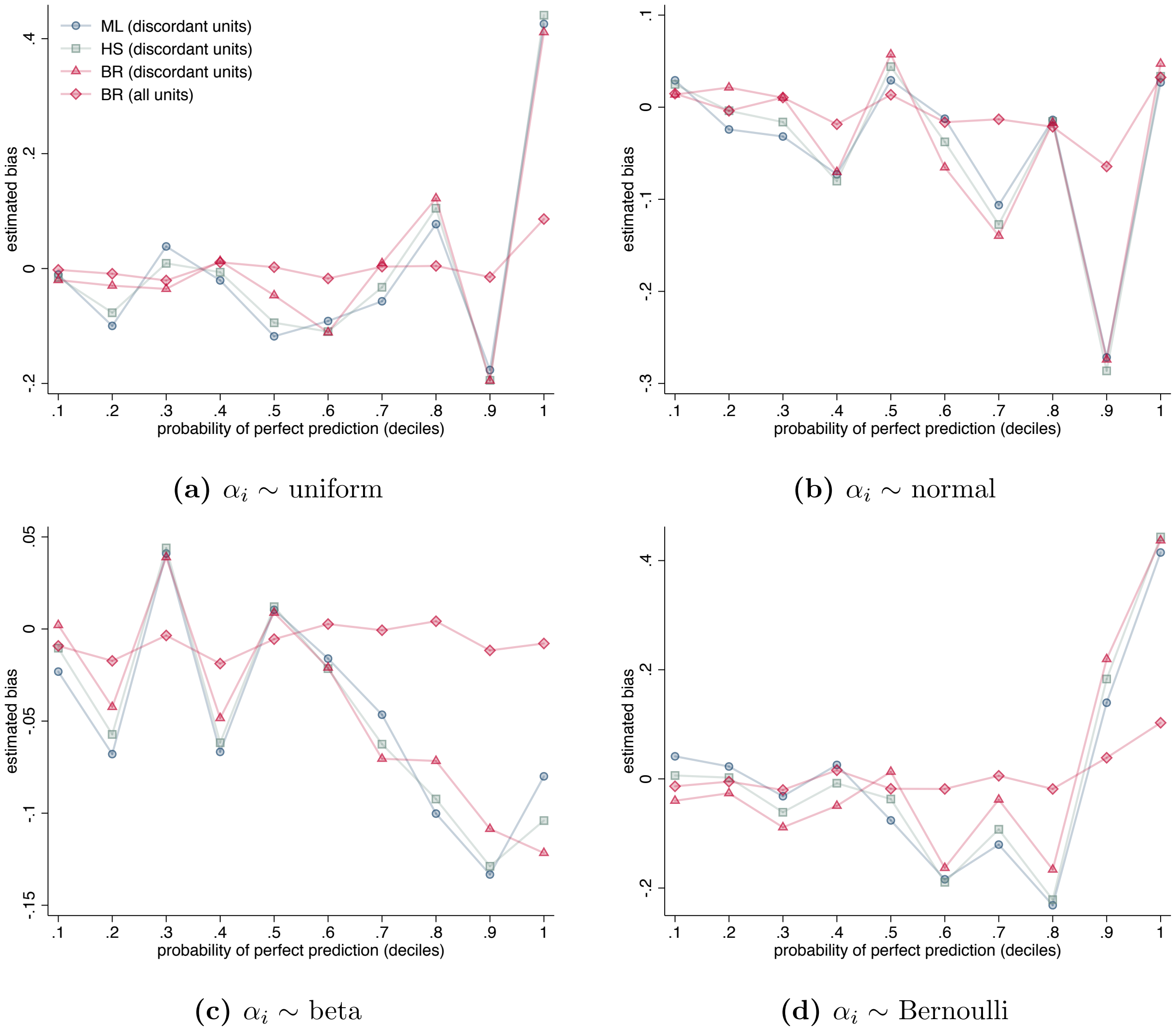


Figure 1: Distributions of α_i

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.



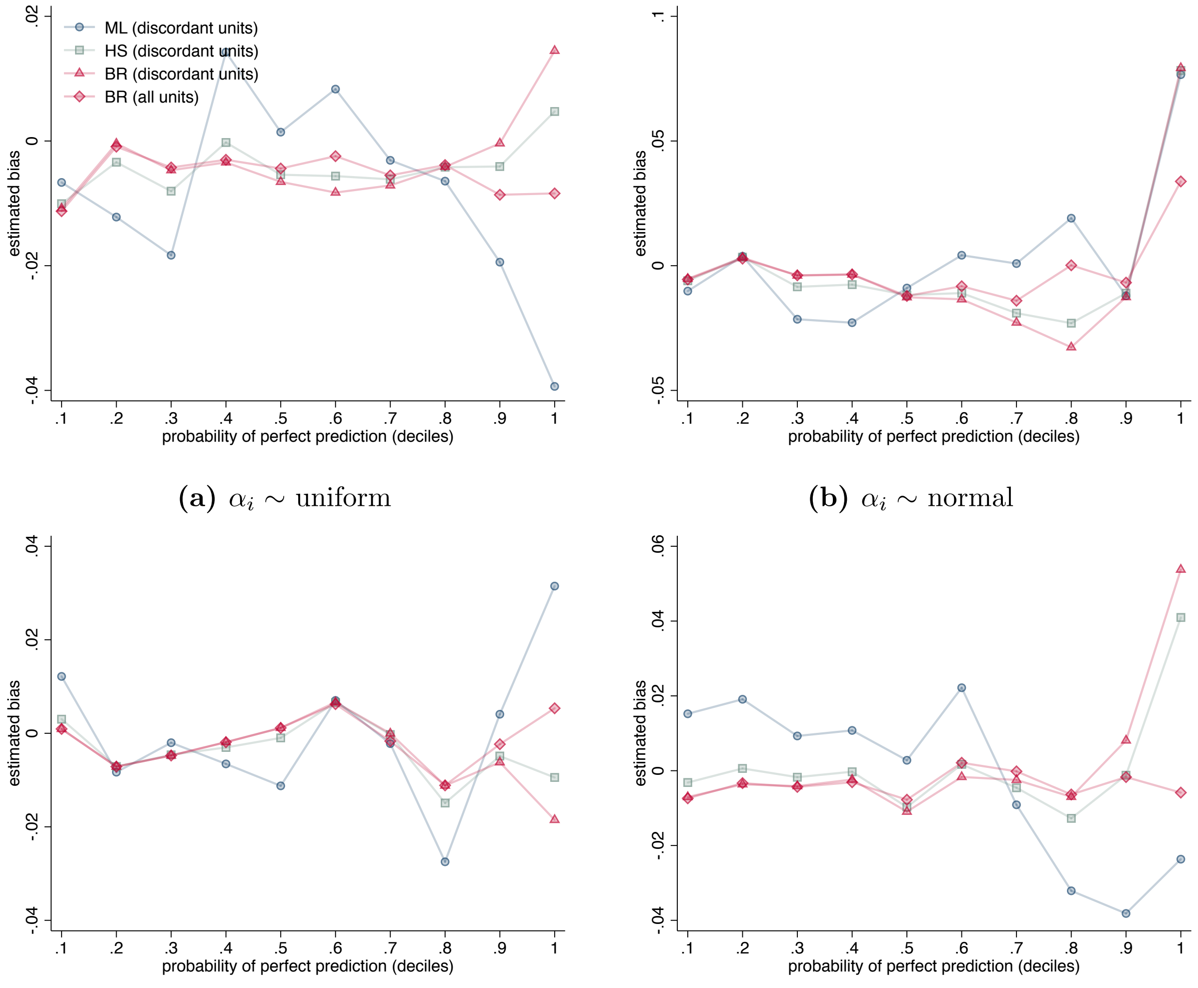
(a) $\alpha_i \sim$ uniform

(b) $\alpha_i \sim$ normal

(c) $\alpha_i \sim$ beta

(d) $\alpha_i \sim$ Bernoulli

Figure 2: Bias and Incidence of Perfect Prediction, $T = 4$



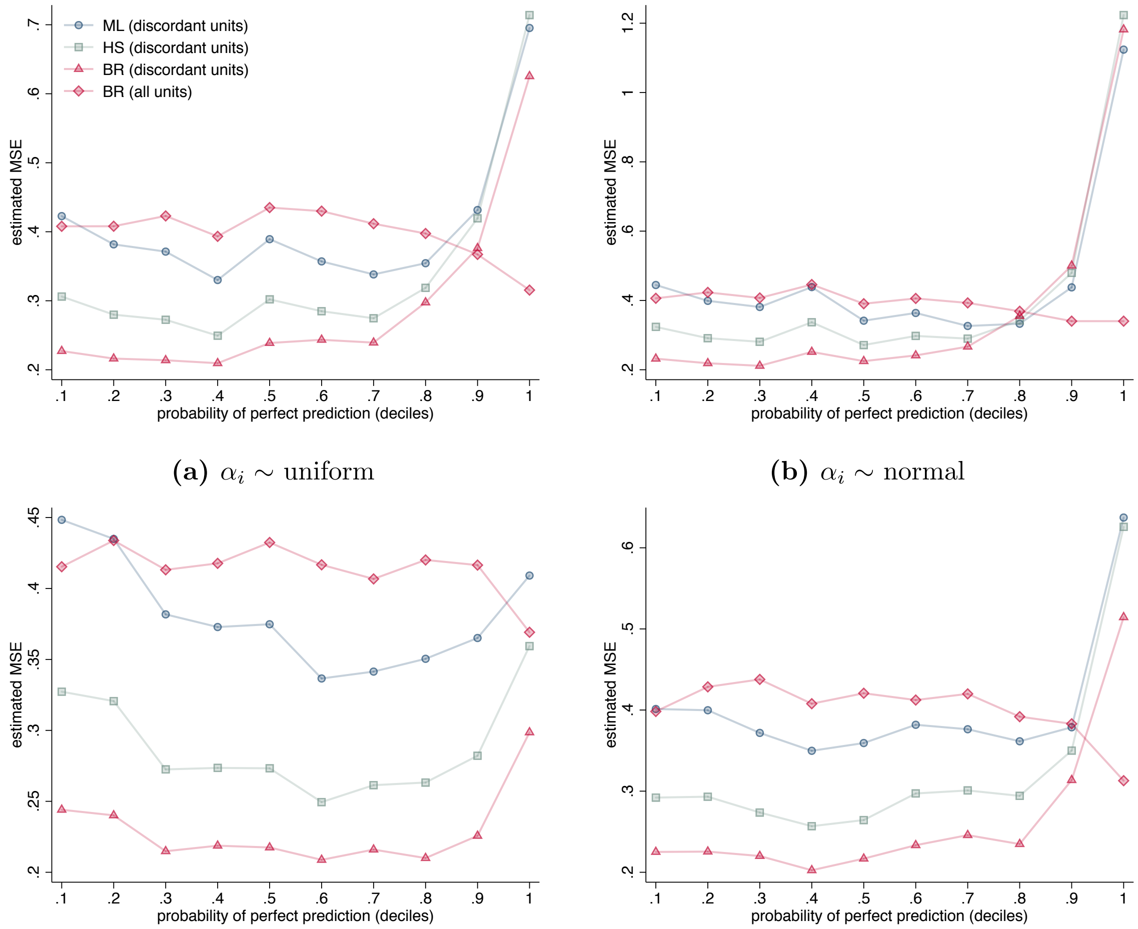
(a) $\alpha_i \sim$ uniform

(b) $\alpha_i \sim$ normal

(c) $\alpha_i \sim$ beta

(d) $\alpha_i \sim$ Bernoulli

Figure 3: Bias and Incidence of Perfect Prediction, $T = 12$



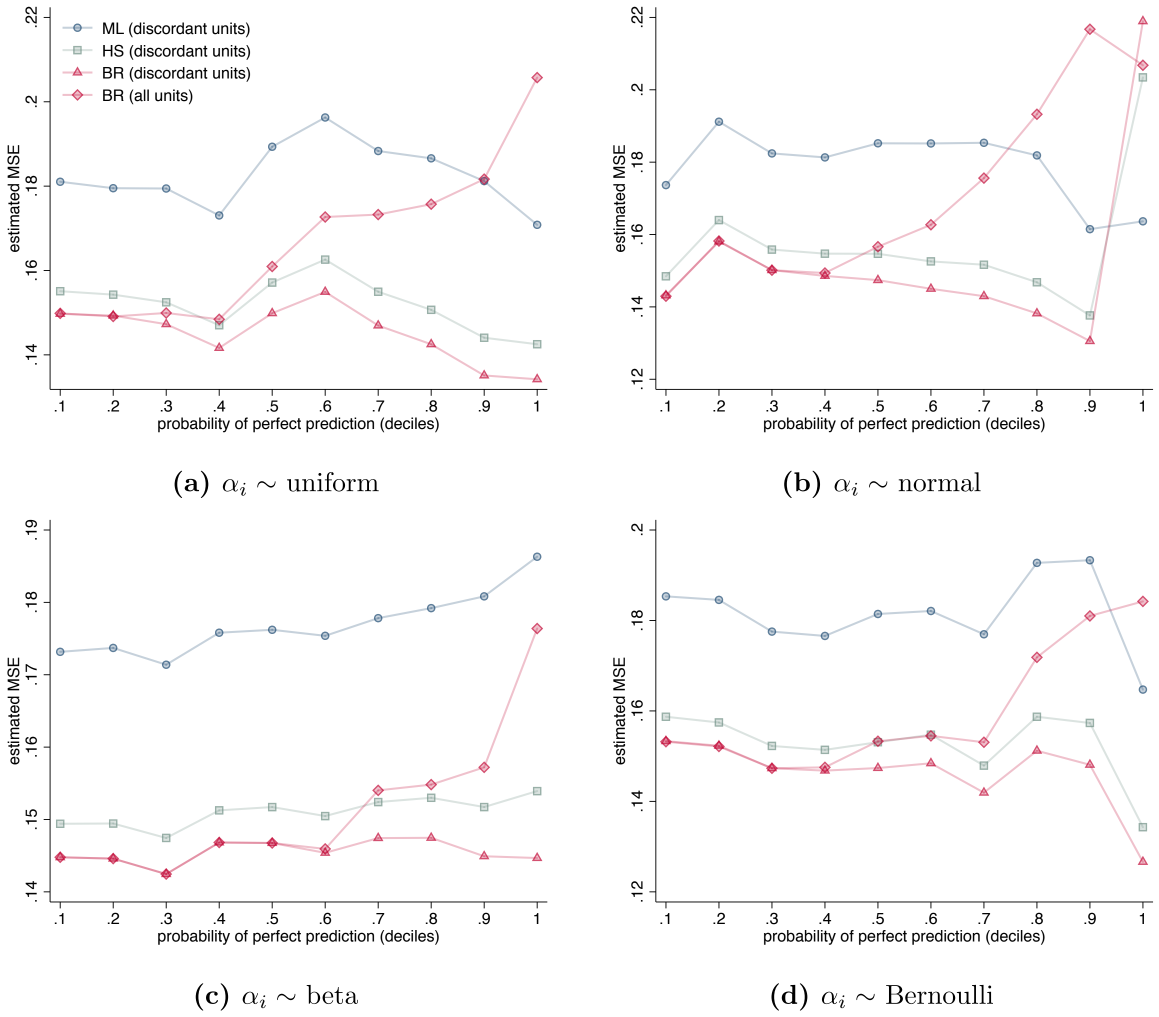
(a) $\alpha_i \sim$ uniform

(b) $\alpha_i \sim$ normal

(c) $\alpha_i \sim$ beta

(d) $\alpha_i \sim$ Bernoulli

Figure 4: Mean Squared Error (MSE) and Incidence of Perfect Prediction, $T = 4$



(a) $\alpha_i \sim$ uniform

(b) $\alpha_i \sim$ normal

(c) $\alpha_i \sim$ beta

(d) $\alpha_i \sim$ Bernoulli

Figure 5: Mean Squared Error (MSE) and Incidence of Perfect Prediction, $T = 12$

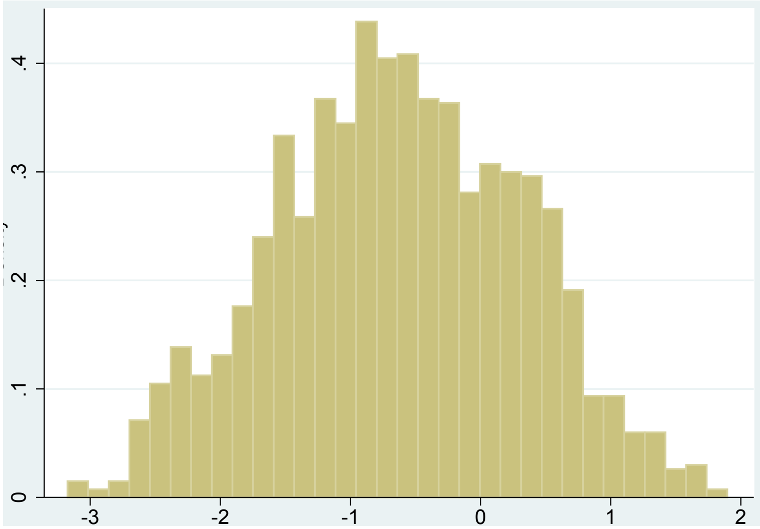
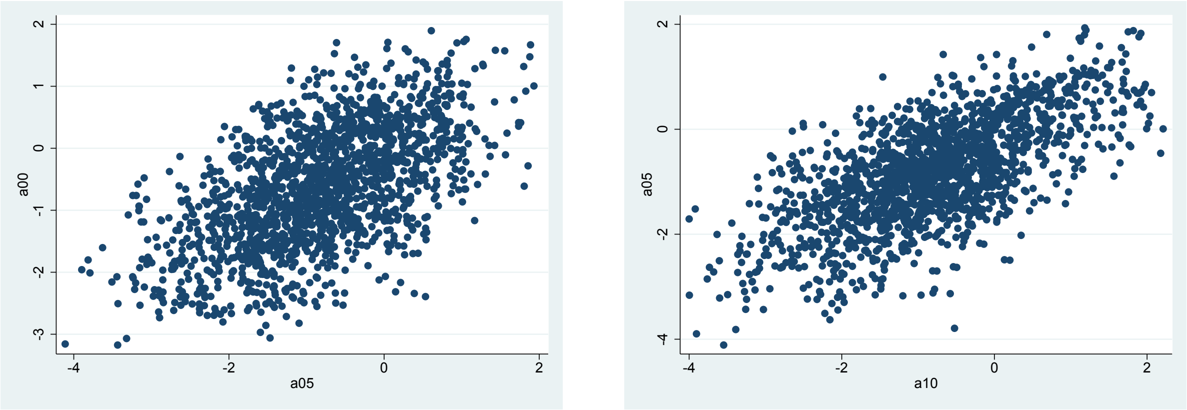


Figure 6: Histogram of predictions of fixed effects, Male sample.

Source: German Socio-Economic Panel 2010-2014, BR probit estimates.



2000-2004 against 2005-2009

2005-2009 against 2010-2014

Figure 7: Scatterplots of period-specific fixed effects

Table 1: MC Simulation: Estimates of $E(\alpha_i)$ [Mean] and $SD(\alpha_i)$ [SD]; $N = 100$, 500 replications

	$T = 2$		$T = 4$		$T = 8$		$T = 12$	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\alpha_i \sim Bernoulli$								
True	*-0.030*	*0.451*						
ML	0.025	0.778	-0.030	0.361	-0.019	0.452	-0.032	0.482
HS	0.024	0.752	-0.027	0.317	-0.017	0.405	-0.029	0.445
BR	-0.018	0.353	-0.023	0.420	-0.031	0.446	-0.034	0.453
$\alpha_i \sim Uniform$								
True	*-0.045*	*0.585*						
ML	0.025	0.819	-0.048	0.445	-0.047	0.573	-0.053	0.620
HS	0.024	0.794	-0.042	0.388	-0.043	0.512	-0.049	0.572
BR	-0.044	0.431	-0.041	0.540	-0.047	0.576	-0.050	0.588
$\alpha_i \sim Beta$								
True	*0.034*	*0.296*						
ML	-0.147	0.836	-0.014	0.317	0.025	0.305	0.034	0.318
HS	-0.142	0.809	-0.014	0.281	0.022	0.274	0.031	0.295
BR	0.007	0.229	0.027	0.290	0.032	0.295	0.033	0.296
$\alpha_i \sim Normal$								
True	*0.045*	*0.733*						
ML	-0.138	0.831	-0.000	0.473	0.028	0.638	0.047	0.710
HS	-0.133	0.803	-0.002	0.410	0.024	0.569	0.043	0.650
BR	0.009	0.502	0.038	0.629	0.040	0.696	0.043	0.715

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

Table 2: MC SIMULATION: MEAN AND STANDARD DEVIATION [SD] OF $\hat{\beta}$ ($\beta = 1$, 500 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
HS	2.038	0.658	1.246	0.228	1.055	0.111	1.022	0.083
BR	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
HS	2.138	0.730	1.268	0.241	1.063	0.110	1.028	0.084
BR	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
HS	2.009	0.699	1.212	0.204	1.047	0.113	1.018	0.081
BR	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
HS	2.124	0.967	1.253	0.234	1.063	0.114	1.030	0.083
BR	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 three estimators, ML, HS, and BR. The true value of β is 1.

Table 3: Descriptive statistics, by period

	2000-2004	2005-2009	2010-2014
Men			
Any visit	0.558	0.594	0.637
Age	39.5	44.5	49.5
SAH	2.38	2.55	2.67
Perfect prediction	0.291	0.312	0.400
Any visit|pp=0	0.515	0.523	0.537
Women			
Any visit	0.721	0.728	0.742
Age	39.0	44.0	49.0
SAH	2.43	2.55	2.67
Perfect prediction	0.388	0.418	0.456
Any visit|pp=0	0.603	0.593	0.593

Table 4: Probit and BR Probit results, 2010-2014

	Men		*Women*	
	BR Probit	Probit	BR Probit	Probit
Age	0.005	0.001	-0.009	-0.001
	(0.062)	(0.088)	(0.052)	(0.080)
Age squared/100	0.009	0.022	0.006	-0.003
	(0.063)	(0.090)	(0.053)	(0.083)
Fulltime work	-0.125	-0.178	0.242	0.371
	(0.171)	(0.266)	(0.115)	(0.187)
Parttime work	0.056	0.077	0.155	0.245
	(0.163)	(0.257)	(0.104)	(0.167)
Married	-0.008	-0.009	0.002	0.029
	(0.117)	(0.164)	(0.104)	(0.162)
Disability	0.165	0.343	0.376	0.976
	(0.120)	(0.218)	(0.113)	(0.275)
Self-assessed health	0.311	0.453	0.258	0.385
	(0.028)	(0.042)	(0.023)	(0.038)
log household income	-0.013	-0.019	0.014	0.011
	(0.071)	(0.106)	(0.051)	(0.080)
log earnings	0.003	0.001	-0.011	-0.017
	(0.018)	(0.028)	(0.012)	(0.019)
Observations	8425	5055	9985	5430
Fixed effects	yes	yes	yes	yes

Appendix A: Panel probit HS estimator for T=2

For $T = 2$, the score for α_i corresponding to the HS estimator is

$$\begin{aligned} s^{HS}(\alpha_i) &= -(\lambda_1+\lambda_2) - \frac{\lambda_1\lambda_1^\alpha + \lambda_2\lambda_2^\alpha}{\lambda_1^\alpha + \lambda_2^\alpha} + \frac{1}{2}\frac{(\lambda_1^2+\lambda_2^2)(\lambda_1^{\alpha\alpha}+\lambda_1^{\alpha\alpha})}{(\lambda_1^\alpha+\lambda_2^\alpha)^2} \\ &= \frac{-2(\lambda_1+\lambda_2)(\lambda_1^\alpha+\lambda_2^\alpha)^2 - 2(\lambda_1\lambda_1^\alpha+\lambda_2\lambda_2^\alpha)(\lambda_1^\alpha+\lambda_2^\alpha) + (\lambda_1^2+\lambda_2^2)(\lambda_1^{\alpha\alpha}+\lambda_1^{\alpha\alpha})}{(\lambda_1^\alpha+\lambda_2^\alpha)^2}, \end{aligned}$$

where we have suppressed the dependence of the notation on i; that is, $\lambda_{i1} = \lambda_1$, etc. Since the denominator is positive for any $(\eta_{i1}, \eta_{i2}) \in R^2$, we only need focus on the numerator, $s^{HS}_{num}(\alpha_i)$:

$$s^{HS}_{num}(\alpha_i) = -4\lambda_1\lambda_1^{\alpha 2} - 4\lambda_2\lambda_2^{\alpha 2} - 2\lambda_1\lambda_2^{\alpha 2} - 2\lambda_2\lambda_1^{\alpha 2} - 6\lambda_1\lambda_1^\alpha\lambda_2^\alpha - 6\lambda_2\lambda_1^\alpha\lambda_2^\alpha + \lambda_1^2\lambda_1^{\alpha\alpha} + \lambda_2^2\lambda_2^{\alpha\alpha} + \lambda_2^2\lambda_1^{\alpha\alpha} + \lambda_1^2\lambda_2^{\alpha\alpha}.$$

For the last four terms, we use

$$\begin{aligned} \lambda_s^2\lambda_t^{\alpha\alpha} &= \lambda_s^2(2\lambda_t\lambda_t^\alpha - \lambda_t^\alpha\eta_t - \lambda_t) \\ &= \lambda_s^2[\lambda_t^\alpha(\lambda_t - \eta_t) + \lambda_t(\lambda_t^\alpha - 1)], \end{aligned}$$

which for $t = s$ simplifies further to

$$\lambda_t^2\lambda_t^{\alpha\alpha} = \lambda_t\lambda_t^{\alpha 2} + \lambda_t^3(\lambda_t^\alpha - 1).$$

Inserting these expressions for the four last terms and rearranging, we obtain

$$\begin{aligned} & s^{HS}_{num}(\alpha_i) = \\ & \quad -3\lambda_1\lambda_1^{\alpha 2} - 3\lambda_2\lambda_2^{\alpha 2} - 2\lambda_1\lambda_2^{\alpha 2} - 2\lambda_2\lambda_1^{\alpha 2} - 6\lambda_1\lambda_1^\alpha\lambda_2^\alpha - 6\lambda_2\lambda_1^\alpha\lambda_2^\alpha + \lambda_1^3(\lambda_1^\alpha - 1) + \lambda_2^3(\lambda_2^\alpha - 1) \\ & \quad + \underline{\lambda_2^2\lambda_1^\alpha(\lambda_1 - \eta_1)} + \lambda_2^2\lambda_1(\lambda_1^\alpha - 1) + \underline{\lambda_1^2\lambda_2^\alpha(\lambda_2 - \eta_2)} + \lambda_1^2\lambda_2(\lambda_2^\alpha - 1). \end{aligned}$$

Other than the two terms underlined with a solid line, all terms are strictly negative. We are interested in the case $\eta_1 \neq \eta_2$; the case $\eta_1 = \eta_2$ was discussed in Section 2.4. Without loss of generality, assume $\eta_1 > \eta_2$. This implies $\lambda_1 > \lambda_2$ and $\lambda_1^\alpha > \lambda_2^\alpha$.

Then, the sum of the first term and the first underlined positive term is negative:

$$-3\lambda_1\lambda_1^{\alpha 2} + \lambda_2^2\lambda_1^\alpha(\lambda_1 - \eta_1) < -3\lambda_1\lambda_1^{\alpha 2} + \lambda_1^2\lambda_1^\alpha(\lambda_1 - \eta_1) = -2\lambda_1\lambda_1^{\alpha 2} < 0,$$

where the first inequality used $\lambda_1^2 > \lambda_2^2$. Thus,

$$\begin{aligned} & s^{HS}_{num}(\alpha_i) < \\ & \quad -2\lambda_1\lambda_1^{\alpha 2} - 3\lambda_2\lambda_2^{\alpha 2} - 2\lambda_1\lambda_2^{\alpha 2} - 2\lambda_2\lambda_1^{\alpha 2} - 6\lambda_1\lambda_1^\alpha\lambda_2^\alpha - 6\lambda_2\lambda_1^\alpha\lambda_2^\alpha + \lambda_1^3(\lambda_1^\alpha - 1) + \lambda_2^3(\lambda_2^\alpha - 1) \\ & \quad + \lambda_2^2\lambda_1(\lambda_1^\alpha - 1) + \underline{\lambda_1^2\lambda_2^\alpha(\lambda_2 - \eta_2)} + \lambda_1^2\lambda_2(\lambda_2^\alpha - 1), \end{aligned}$$

where the underlined term is the only positive one.

We now consider a case where $\hat{\alpha}_i$ does not exist. Suppose $1 < \eta_1 - \eta_2 \leq 6$; that is, $1 < x'_{i1}\beta - x'_{i2}\beta \leq 6$. We only consider the limit case, as the results for differences smaller than 6 follow immediately using the same arguments.[1] Without loss of generality, $\eta_1 = \alpha_i$, $\eta_2 = \alpha_i - 6$. We consider only the first, the fifth and the underlined positive term from the right-hand-side of the previous inequality:

$$\begin{aligned} -2\lambda_1\lambda_1^{\alpha 2} - 6\lambda_1\lambda_1^{\alpha}\lambda_2^{\alpha} + \lambda_1^2\lambda_2^{\alpha}(\lambda_2 - \eta_2) &= -2\lambda_1^2\lambda_1^{\alpha}(\lambda_1 - \eta_1) - 6\lambda_1^2\lambda_2^{\alpha}(\lambda_1 - \eta_1) + \lambda_1^2\lambda_2^{\alpha}(\lambda_2 - \eta_2) \\ &= \frac{1}{2}\lambda_1^2\left[\lambda_2^{\alpha}(\lambda_2 - \eta_2) - 4\lambda_1^{\alpha}(\lambda_1 - \eta_1)\right] \\ &\quad + \frac{1}{2}\lambda_1^2\lambda_2^{\alpha}\left[\lambda_2 - \eta_2 - 12(\lambda_1 - \eta_1)\right] \\ &= \frac{1}{2}\lambda_1^2\left[g(\eta_2) - 4g(\eta_1)\right] + \frac{1}{2}\lambda_1^2\lambda_2^{\alpha}\left[h(\eta_2) - 12h(\eta_1)\right], \end{aligned}$$

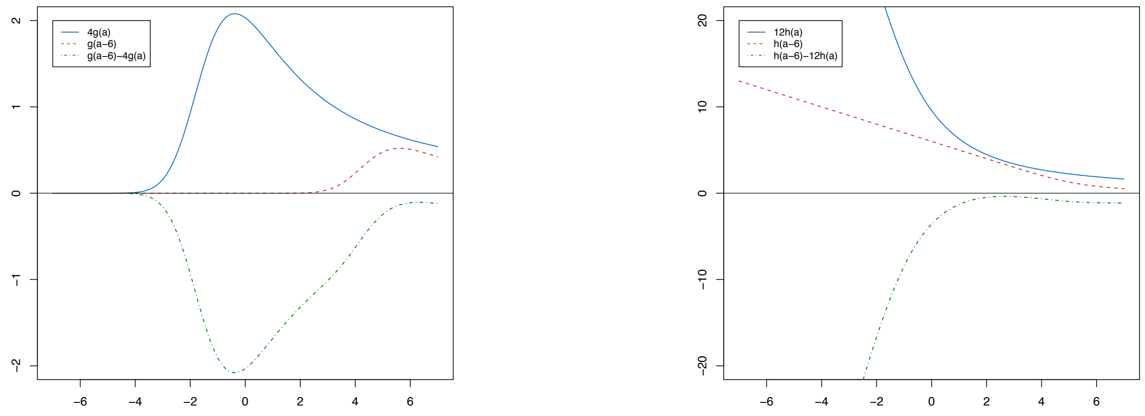
where we defined the functions $g(\eta) = \lambda^{\alpha}(\lambda - \eta)$ and $h(\eta) = \lambda - \eta$. The factors multiplying the two terms in brackets on the right-hand-side of the last equality are positive for all α_i, so it suffices to show that the terms in brackets are negative for any finite value of α_i to prove that $s^{HS}(\alpha_i)$ is negative for all $\alpha_i \in R$ and thus that for $x'_{i1}\beta - x'_{i2}\beta = 6$ the estimator $\hat{\alpha}_i$ does not exist.

Figure 8 plots the two terms in brackets. Each one is negative over the entire plotted range (green dash-dotted lines). This holds in general as well. Consider the first term in brackets, $g(\eta_2) - g(\eta_1)$. It is straightforward to show that the function $g(\eta)$ is positive for all η and has a unique global maximum of about 0.52 at about $\eta = -0.3826$. Then, since $4g(-0.3826 + 6) \approx 0.6547 > 0.52$, the first term in brackets is strictly negative for all α_i. (In the left panel of Figure 8, this can be seen as the solid blue line, which depicts $4g(\alpha_i)$, clearly passes over the maximum of the red dashed line, which depicts $g(\alpha_i - 6)$.)

Now consider the second term in brackets, $h(\eta_2) - 12h(\eta_1) = h(\alpha_i - 6) - 12h(\alpha_i)$. As $\alpha_i \to -\infty$, the slopes of the two components tend to $h'(\alpha_i - 6) \to -1$ and $-12h'(\alpha_i) \to -12$. For $\alpha_i \to +\infty$, we have $h'(\alpha_i - 6) \to 0$ and $-12h'(\alpha_i) \to 0$. The slope $h'(\eta) = \lambda^{\alpha} - 1$ is monotonically increasing with exactly one inflection point, $h''(\eta^*) = 0$, at about $\eta^* \approx -1.002$. Thus, to show that $12h(\alpha_i) > h(\alpha_i - 6)$ for all α_i, we just need to show that this holds at $\alpha = \alpha^{\circ}$ where $12h'(\alpha^{\circ}) = h'(\alpha^{\circ} - 6)$ (i.e., where the slopes of the two components are the same), and at $\alpha = \alpha^*$ where $h''(\alpha^* - 6) = 0$ (i.e., at the inflection point of the positive component). Here, $\alpha^{\circ} \approx 2.468$, at which point $h(\alpha^{\circ} - 6) - 12h(\alpha^{\circ}) \approx -0.367$; and $\alpha^* \approx 4.998$, at which $h(\alpha^{\circ} - 6) - 12h(\alpha^{\circ}) \approx -0.950$. (In the right panel of Figure 8, this can be seen as the solid blue line, which depicts $12h(\alpha_i)$, always lies higher than the dashed red line, which depicts $h(\alpha_i)$.)

[1] This example is meant as an illustration. The interval $1 < x'_{i1}\beta - x'_{i2}\beta \leq 6$ is not a tight bound for the interval in which $\hat{\alpha}_i$ does not exist. However, differences such as $x'_{i1}\beta - x'_{i2}\beta = 6$ already represent quite extreme change in the covariates of a unit i; in this case, corresponding to a change of 6 standard deviations in the distribution of the error term.

Figure 8: Appendix: Some terms in $s^{HS}(\alpha)$



Panel I: $g(\eta_2) - 4g(\eta_1)$

Panel II: $h(\eta_2) - 12h(\eta_1)$

Notes: For both panels, $\eta_1 = \alpha$, $\eta_2 = \alpha - 6$. Panel I: $g(\eta) \equiv \lambda^{\alpha}(\lambda - \eta)$. Panel II: $h(\eta) \equiv \lambda - \eta$. The blue solid lines show the absolute value of the negative term, the red dashed lines the positive term of the functions. The dash-dotted green line depicts the sum of the negative and positive terms.