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

Working Paper

Estimating Fixed Effects: Perfect Prediction and Bias in Binary Response Panel Models, with an Application to the Hospital Readmissions Reduction Program

IZA Discussion Papers, No. 11182

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Kunz, Johannes S.; Staub, Kevin E.; Winkelmann, Rainer (2017) : Estimating Fixed Effects: Perfect Prediction and Bias in Binary Response Panel Models, with an Application to the Hospital Readmissions Reduction Program, IZA Discussion Papers, No. 11182, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/174092

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11182

Estimating Fixed Effects: Perfect Prediction and Bias in Binary Response Panel Models, with an Application to the Hospital Readmissions Reduction Program

Johannes S. Kunz
Kevin E. Staub
Rainer Winkelmann

NOVEMBER 2017

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11182

Estimating Fixed Effects: Perfect Prediction and Bias in Binary Response Panel Models, with an Application to the Hospital Readmissions Reduction Program

Johannes S. Kunz
Monash University

Kevin E. Staub
University of Melbourne and IZA

Rainer Winkelmann
University of Zurich and IZA

NOVEMBER 2017

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Estimating Fixed Effects: Perfect Prediction and Bias in Binary Response Panel Models, with an Application to the Hospital Readmissions Reduction Program*

The maximum likelihood estimator for the regression coefficients, β, in a panel binary response model with fixed effects can be severely biased if *N* is large and *T* is small, a consequence of the incidental parameters problem. This has led to the development of conditional maximum likelihood estimators and, more recently, to estimators that remove the $O(T^{-1})$ bias in β^. We add to this literature in two important ways. First, we focus on estimation of the fixed effects proper, as these have become increasingly important in applied work. Second, we build on a bias-reduction approach originally developed by Kosmidis and Firth (2009) for cross-section data, and show that in contrast to other proposals, the new estimator ensures finiteness of the fixed effects even in the absence of within-unit variation in the outcome. Results from a simulation study document favourable small sample properties. In an application to hospital data on patient readmission rates under the 2010 Affordable Care Act, we find that hospital fixed effects are strongly correlated across different treatment categories and on average higher for privately owned hospitals.

JEL Classification: C23, C25, I18

Keywords: perfect prediction, bias reduction, penalised likelihood, logit, probit, Affordable Care Act

Corresponding author:
Kevin Staub
Department of Economics
The University of Melbourne
111 Barry Street
3010 VIC
Australia
E-mail: kevin.staub@unimelb.edu.au

* We thank participants of the Econometrics Workshop (Melbourne), International Panel Data Conference (Thessaloniki), Australian Health Economics Society conference (Sydney), as well as seminar participants at CHE Monash University and the University of Zurich, for helpful comments. We also thank Bob Breunig, Stefan Bruder, Denzil Fiebig, Joe Hirschberg, Andrew Jones, Maarten Lindeboom, Jenny Lye, Janina Nemitz, Carol Propper, Chris Skeels, Jenny Williams, and Tiemen Woutersen for helpful comments. 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 ssc under the name brglm.

1 Introduction

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

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

where $y_{it} \in \{0, 1\}$, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, $\boldsymbol{x}_{it}$ is a vector of covariates and $\boldsymbol{\beta}$ a conformable vector of coefficients. Typically, N is large and T is small. This model is very popular among empirical practitioners, since it does not require any assumption on the distribution of α_i regardless of whether the α_i's are exogenous (uncorrelated with $\boldsymbol{x}_{it}$) or endogenous.

As noted in the literature, the maximum likelihood estimator (MLE), $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = (\hat{\alpha}_1, \ldots, \hat{\alpha}_N, \hat{\boldsymbol{\beta}})$, that is obtained from maximising the log-likelihood function with respect to $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ has a number of deficiencies in this case. First, $\hat{\boldsymbol{\beta}}$ is inconsistent. This is an instance of the incidental parameters problem. Abrevaya (1997) has shown for the panel logit model with $T = 2$, that plim$\hat{\boldsymbol{\beta}}$ =2$\boldsymbol{\beta}$. 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 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" (e.g., Maddala, 1983).

The second and third problems would be of little relevance if the only objective was estimation of $\boldsymbol{\beta}$. However, this is rarely, if ever, the case. First, the $\hat{\alpha}_i$'s are needed to estimate marginal effects or obtain predicted probabilities. Second, the $\hat{\alpha}_i$'s can be of intrinsic interest per-se, for instance in order to classify or rank individuals by their propensity to experience the event.

A recent literature has reconsidered estimation of such models by giving up the fixed-T assumption (Hahn and Kuersteiner, 2002; Hahn and Newey, 2004; Fernández-Val, 2009; Dhaene and Jochmans, 2015).[1] The incidental parameters problem is then seen as a manifestation of first-order bias, and bias corrections can remove the $O(1/T)$ bias in $\hat{\boldsymbol{\beta}}$. In these papers, this is done by subtracting the first-order bias ex-post from the MLE. A related approach obtains an estimator free of first-order bias directly from a modified objective function. For instance, Bester and Hansen (2009) propose to add to the log-likelihood function a penalty term related to the discrepancy between Hessian and outer product of the score (HS estimator; the estimator was also developed independently in Arellano and Hahn 2016; see also Bartolucci et al. 2016 for a similar estimator).

[1] Asymptotics for the case that both T and N increase have been developed by Woutersen (2001) and Hahn and Newey (2004).

While these solutions address the incidental parameters problem, they do not solve the perfect prediction problem. This is obvious for bias-correction, as this approach requires estimation of the MLE. We show below that, in general, the HS estimator equally fails in the perfect prediction case in the context of a panel probit model. In this paper, we therefore explore an alternative bias-reduced (BR) estimator for panel probit and logit models that addresses the incidental parameters problem and works well for datasets with a high incidence of perfect prediction. The BR estimator we advocate is due to Kosmidis and Firth (2009) and it is based on the idea originally due to Firth (1993) of removing the first-order bias $b_1(\boldsymbol{\theta})$ of the MLE for the parameter $\boldsymbol{\theta}$ by using a modified score function $\tilde{s}(\boldsymbol{\theta}) = s(\boldsymbol{\theta}) - I(\boldsymbol{\theta})b_1(\boldsymbol{\theta})/N$, where $s(\boldsymbol{\theta})$ is the score, $I(\boldsymbol{\theta})$ is the Fisher information and N the sample size. This modification ensures that the root of $\tilde{s}(\boldsymbol{\theta})$, say $\hat{\boldsymbol{\theta}}$, is a bias-reduced estimator of $\boldsymbol{\theta}_0$.

Relatively simple expressions for $\tilde{s}(\boldsymbol{\theta})$ can be derived for the probit and logit models, as well as other linear exponential family models. In the logit case, this approach is equivalent to penalising the log-likelihood using Jeffreys invariant prior. Moreover, in this case, the BR estimator coincides with the expectation version (IE) of the HS estimator. But in the probit case (as well as other alternative binary response models) they do not. The BR estimator is not only immune to the perfect prediction problem, but it is also relatively easy to compute, as it can be obtained using an iteratively weighted least squares estimator (Kosmidis and Firth, 2009).

Bias-reduced estimation is one way to address the perfect prediction problem, in particular, if individual heterogeneity is not treated as a nuisance parameter but rather as a model feature of intrinsic interest. There is a growing literature which focusses on such distributions of individual-specific heterogeneity net of the effects of some x_{it}, including neighborhood effects (Chetty and Hendren, 2015), teacher effects (Chetty, Friedman and Rockoff, 2014), worker and firm effects (Card, Heining and Kline, 2013), judges effects (Abrams, Bertrand and Mullainathan, 2012), and doctor and hospital effects (Street et al., 2014); see Abadie and Kasy (2016) for an excellent overview. To date, all this work has been confined to linear models, presumably for a lack of viable alternatives.

While linear models have the advantage that they do not suffer from the perfect prediction problem, the use of linear models for panel data with binary response variables in the setting of this paper—short panels with a high incidence of perfect prediction—is inadequate and can lead to severely distorted estimates. The high prevalence of "perfectly predicted" observations results in substantial shares of linear predictions outside the unit interval. In our simulations and in our application, OLS produced up to almost 50 per cent of such predictions. This can be a severe problem in itself; for instance, if predicted probabilities are needed as inputs into structural models. It also implies

that the estimates of probability effects based on OLS for values other than the mean might be substantially misleading, a point we return to in our simulation.

In the next section, we introduce the problem of perfect prediction in the context of binary response fixed effects panel data, the BR estimator, which solves it, and the HS estimator, which does not. In Section 3, we show in Monte Carlo simulations that the BR estimator has a superior performance in estimating the distribution of α_i across a number of starkly differently-shaped distributions. The simulations also indicate that the BR estimator performs well in terms of obtaining more reliable estimates of $\boldsymbol{\beta}$ in short panels.

Finally, in Section 4, we consider an application to the US Hospital Readmissions Reduction Program, a policy which came into effect in 2012 as part of the Affordable Care Act. Its broad aim was to increase the quality of health care offered to patients. It consisted in imposing negative incentive payments (monetary penalties) for hospitals which exceeded a threshold value of risk-standardised 30-day readmission rates in three specific health conditions. In the health policy literature, this measure is debated intensely, with a set of arguments focussing on the appropriateness of the risk adjustment, which by not including socio-economic characteristics of the patients might have disadvantaged certain hospitals unfairly. Using the BR estimator, we analyse hospital-specific unobserved heterogeneity in an unbalanced panel of about 3,000 hospitals over five years, and over the three penalised health conditions. Our results indicate that this time-invariant heterogeneity is an important determinant of penalty status for a given health condition, that the correlation in heterogeneity across conditions is positive, and that there are significant differences in this heterogeneity across for-profit and non-profit hospitals.

2 Econometric methods

Non-linear maximum likelihood estimators have a finite sample bias. Considering the T dimension, the bias can be split up in 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 maximum likelihood estimators is given in Cox and Snell (1971). For an illustration, consider a simple panel probit model with time-invariant regressors only:

$$\Pr(y_{it} = 1|\tilde{\alpha}_i, \bar{x}_i) = \Phi(\tilde{\alpha}_i + \bar{x}_i'\gamma), \qquad t = 1, \ldots, T.$$

Since $\tilde{\alpha}_i$ and γ are not separately identified, we substitute $\alpha_i = \tilde{\alpha}_i + \bar{x}_i'\gamma$, with first-order condition, for unit i,

$$\frac{\partial \log L}{\partial \alpha_i} = \sum_{y_{it}=1} \frac{\phi(\alpha_i)}{\Phi(\alpha_i)} - \sum_{y_{it}=0} \frac{\phi(\alpha_i)}{(1-\Phi(\alpha_i))} = 0,$$

where $\phi(\cdot)$ denotes the standard normal density function. Estimates of γ can then in principle be obtained in a second-step regression of $\hat{\alpha}_i$ on $\bar{x}$.[2]

In this example, it can be shown (cf. Appendix) that the first-order bias is given by

$$Bias = \frac{1}{T} \frac{\alpha_i \Phi(\alpha_i)(1-\Phi(\alpha_i))}{\phi^2(\alpha_i)}. \tag{2}$$

Therefore, the bias is positive if $\alpha_i > 0$, and hence $\Phi(\alpha_i) > 0.5$. It is negative for $\alpha_i < 0$. Moreover, the bias increases in the absolute value of α_i. As α_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. Since perfect prediction is more likely to occur for α_i which are large in absolute value, this is an indication of the close relationship between first-order bias and perfect prediction in this case.

2.1 Perfect prediction problem in the panel probit model

Perfect prediction in the general model (1) means that the first-order conditions for the maximum likelihood estimator do not have a finite solution. This problem can arise with any ill-designed x-vector, but it is particularly relevant, and easily detectable, in the context of the panel probit log-likelihood function with fixed effects. The first-order conditions are:

$$s^{ML}(\beta_k) = \frac{\partial \log L}{\partial \beta_k} = \sum_{i=1}^{N}\sum_{t=1}^{T} (y_{it} - \Phi(\eta_{it})) \frac{\phi(\eta_{it})}{\Phi(\eta_{it})(1-\Phi(\eta_{it}))} x_{k,it} = 0, \qquad k = 1, \ldots, K, \tag{3}$$

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

where $\eta_{it} = \alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta}$, and K is the number of regressors in $\boldsymbol{x}_{it}$.[3] Suppose that $y_{i1} = \ldots = y_{iT} = 0$ for some i. Then (4) simplifies to

$$-\sum_{t=1}^{T} \frac{\phi(\eta_{it})}{1-\Phi(\eta_{it})} = 0, \tag{5}$$

[2] An alternative approach would be to treat $\tilde{\alpha}_i$ as random effects and estimate γ directly by maximising the marginal log-likelihood function. This approach would require a distributional assumption, however, that may be invalid.

[3] We use notation for the balanced panel case for expositional simplicity, but this is immaterial for the argument. Our application in Section 4 uses an unbalanced panel.

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, (4) simplifies to

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

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 existence of the estimator for $\boldsymbol{\beta}$ is unaffected by perfect prediction. Assuming that there are some panel units with variation in y_{it} (i.e., some *discordant* units), $\boldsymbol{\beta}$ can be estimated using those observations only, based on (3). For perfectly predicted observations, the contributions to the (concentrated) score $y_{it} - \Phi(\hat{\alpha}_i(\boldsymbol{\beta}) + \boldsymbol{x}_{it}'\boldsymbol{\beta}) \approx 0$, so they do not contribute to estimation of $\hat{\boldsymbol{\beta}}$.

With perfect prediction, estimates $\hat{\alpha}_i$ for affected panel units i do not exist. Hence, we cannot make any inferences on quantities that depend on α_i. This problem will be most severe for small values of T. As T increases, and provided that $0 < \Pr(y_{it} = 1) < 1$, it becomes less and less likely to observe panel units with $\bar{y}_i = 0$ or $\bar{y}_i = 1$.

2.2 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 in so-called canonical parameterisation 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 canonical parameterisation is the logit model.

For linear exponential family models in non-canonical parameterisation—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

$$\begin{aligned} s^{BR}(\alpha_i) &= \sum_{t=1}^{T} \left[y_{it} - \Phi(\eta_{it}) - \frac{1}{2} h_{it} \eta_{it} \frac{\Phi(\eta_{it})(1 - \Phi(\eta_{it}))}{\phi(\eta_{it})} \right] \frac{\phi(\eta_{it})}{\Phi(\eta_{it})(1 - \Phi(\eta_{it}))} \\ &= s(\alpha_i) - \sum_{t=1}^{T} \frac{1}{2} h_{it} \eta_{it}, \end{aligned} \tag{7}$$

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

$$\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{1/2}, \tag{8}$$

with $\boldsymbol{X}$ the $NT \times K$ matrix of the K regressors, and $\boldsymbol{W}$ is the $NT \times NT$ diagonal matrix with typical element $w_{it} = \phi(\eta_{it})^2/[\Phi(\eta_{it})(1-\Phi(\eta_{it}))]$. From (7), it can be seen that if we redefined

$$y^*_{it} = y_{it} - \frac{1}{2}h_{it}\eta_{it}\frac{\Phi(\eta_{it})(1-\Phi(\eta_{it}))}{\phi(\eta_{it})}, \tag{9}$$

then (7) would be the standard MLE score $s^{ML}(\alpha_i)$, but for the pseudo-response y^*_{it}.

It is therefore possible to solve the 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 also from a computational point of view. For implementation, pseudo-responses are constructed, at the iteration s, using existing estimates from the previous iteration s–1 to replace the unknown quantities h_{it} and η_{it} with estimates $\hat{h}_{it}(\hat{\alpha}^{s-1}, \hat{\boldsymbol{\beta}}^{s-1})$ and $\hat{\eta}_{it}(\hat{\alpha}^{s-1}, \hat{\boldsymbol{\beta}}^{s-1})$.[4]

To examine whether the estimator based on the modified score (7) 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 (7) as

$$s^{BR}(\alpha_i) = \left(\sum_t \frac{\phi(\eta_{it})}{\Phi(\eta_{it})}\right) - \frac{\alpha_i}{2}\left(\sum_t h_{it}\right) - \frac{1}{2}\left(\sum_t h_{it}x'_{it}\beta\right) = g_1(\alpha_i) - \alpha_i g_2(\alpha_i) - g_3(\alpha_i). \tag{10}$$

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. It is evident that similar arguments can be made to show that a solution exists for the other perfect prediction case, $\sum_t y_{it} = 0$, as well.[5]

An interesting example is the case $\boldsymbol{\beta} = 0$, i.e. a constants-only model. The first perfect prediction

[4] An implementation in Stata is available from the authors. For an implementation in R, see Kosmidis et al. (2017).

[5] Heinze and Schemper (2002) noted that the Firth method for bias reduction solves the perfect prediction problem for the cross-sectional Logit model.

case gives

$$\alpha_i = 2T \sum_t \frac{\phi(\alpha_i)}{\Phi(\alpha_i)}, \tag{11}$$

and the second case gives

$$\alpha_i = -2T \sum_t \frac{\phi(\alpha_i)}{1-\Phi(\alpha_i)}, \tag{12}$$

where we used the fact that for this example $\sum_t h_{it} = 1$. The two cases only differ in the sign. For $T = 2, 3, 4$ this gives estimates for α_i of about ± 1.06, ± 1.24, ± 1.37. The associated probabilities for, e.g., the second case are about 0.144, 0.107, 0.086. These estimates reflect the shrinkage away from the bounds of 0 (and 1) associated with values of minus (and plus) infinity of α_i, which are "built into" this estimator. The shrinkage is asymptotically negligible.[6]

2.3 HS Estimator

An alternative estimator based on a penalised likelihood approach is the HS estimator of Bester and Hansen (2009). The HS estimator is very general and therefore applicable to a broad class of models; for instance, it has been used in applications with multiple fixed effects (Hospido, 2012; Carro and Traferri, 2014). Here, we consider the case of a probit model with a single fixed effect. For unit i, the objective function is

$$Q_i^{HS} = \sum_{t=1}^{T} [y_{it} \log(\Phi(\eta_{it}) + (1-y_{it}) \log(1-\Phi(\eta_{it}))] - \frac{1}{2} \frac{\sum_t v_{it}^2}{\sum_t -v_{it}^{\alpha}} + \frac{1}{2}. \tag{13}$$

The first term on the right-hand side is the conventional log-likelihood contribution of unit i associated with (1). The remainder is a penalty term which depends on the discrepancy between the outer product of the score and the (negative of the) Hessian, both with respect to α_i. The outer product of the score is given by $\sum_t v_{it}^2$, where v_{it} denotes the per-period score with respect to α_i, that is $s(\alpha_i) = \sum_t v_{it}$. The Hessian is given by $\sum_t v_{it}^{\alpha}$, where $v_{it}^{\alpha} = \partial v_{it}/\partial \alpha_i$.

For the perfect prediction case of $y_{i1}, ...y_{iT} = 0$, 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 used the shorthand notation $\lambda_{it} = \lambda(\eta_{it})$ to denote the inverse Mills ratio and $\lambda_{it}^{\alpha} =$

[6] The MLE solution, of course, is a probability of exactly zero in each of these cases, which, while unbiased, might be an unreasonable estimate for many applications: it means that an event is deemed impossible based on not having occurred in two or three periods.

$\partial\lambda(\eta_{it})/\partial\alpha_i$ its derivative. We can rewrite individual i's contribution to the penalised 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

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

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 Honore, 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 $\lambda_{it} = \lambda_{is} \equiv \lambda_i$ for all t, s. The HS score then simplifies to

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

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 $(\boldsymbol{x}_{i1}'\boldsymbol{\beta}, \boldsymbol{x}_{i2}'\boldsymbol{\beta}) \in \boldsymbol{R}^2$. In our simulation study in the next section, we also considered several $T > 2$. We did not find a case where the HS estimator for α_i existed in practice.

2.4 Other binary response models

In the Appendix, we discuss BR and HS estimators for the more general case of binary response panel models of the form $P(y_{it} = 1|\boldsymbol{x}_{it}, \alpha_i) = F(\boldsymbol{x}_{it}'\boldsymbol{\beta} + \alpha_i)$, with some known function $F(\cdot)$. For all the commonly used functions $F(\cdot)$ in the literature, such as logit, cloglog, Weibull, etc., the BR estimator ensures existence under perfect prediction. Of particular interest is logit, which is the canonical parametrisation for generalised linear models with a binary response variable. For this special case, the BR estimator has a penalised likelihood representation (Firth, 1993). Similarly, it is possible to modify the HS estimator in the logit case. The resulting estimator is called "IE" as it involves integrating expectations of the scores and their derivatives. The IE estimator has been shown to have a better small sample performance than the HS estimator (Bester and Hansen, 2009). We show that both estimators coincide in the logit case as they use the same penalty. The

only difference is that BR penalises both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, while IE only penalises $\boldsymbol{\alpha}$. However, as the penalty on $\boldsymbol{\beta}$ vanishes as NT increases, the difference is negligible for large N.

3 Monte Carlo evidence

3.1 Experimental design

The primary aim of our Monte Carlo experiment is to investigate how well the approaches discussed in the previous section estimate the unobserved individual-specific heterogeneity as well as its distribution, in simple probit models with a small to moderate number of time periods and a high prevalence of perfect prediction.

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 on [-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

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

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 percent. This fraction is somewhat lower for the beta (15 percent) and Bernoulli (20 percent) distributions, respectively, and higher for the normal distribution (28

percent). 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 The distribution of unobserved heterogeneity

Figure 2 displays results for $T = 4$, with one row each for the different distributions of the true fixed effects and columns for the three different estimation methods. In each case, the average estimated fixed effects

$$\bar{\hat{\alpha}}_i = \left[\sum_{r=1}^{R} d_{ir}\right]^{-1} \sum_{r=1}^{R} d_{ir}\hat{\alpha}_i^{(r)},$$

is plotted against the true α_i Here, d_{ir} is an indicator variable indicating whether the estimator exists ($d_{ir} = 1$) or not ($d_{ir} = 0$). Points along the 45 degree line signify unbiasedness of the estimator.

Panel (A) displays results for uniformly distributed α_i. Both the Maximum Likelihood estimator (ML, left-hand-side graph) and the HS Penalised Likelihood estimator (HS, middle graph) miss the 45 degree line visibly and along the whole range of α_i. In contrast, the BR estimator (BR, right-hand-side graph) follows the 45 degree line closely for the most part, which indicates that it is capturing the shape of the true distribution of α_i well. However, some underestimation, in absolute value, is visible at the tails.

Panel (B) shows the same plots for the beta distribution. In the middle of the range of α_i, both ML and HS are closer to the 45 degree line. However the large discrepancies at the tails show that they fail to capture the skewed shape of this distribution. In contrast, BR estimates the shape of this distribution very well. Panel (C) contains similar plots for the normal distribution, and it suggests the same conclusions than for the previous panels. The bottom panel depicts results for the Bernoulli distribution. Here a plot like the ones in Panels (A)–(C) would only reveal the dispersion around the two mass points, so we show instead the true empirical cumulative distribution function of α_i against its estimated counterpart. Only the empirical cumulative distribution function (cdf) of the BR estimates resembles the true step function. The empirical cdfs of $\hat{\alpha}_i$ of the other estimators, on the other hand, do not capture the discreteness of the true cdf.

Thus, the results of Figure 2 illustrate the advantage of BR over the other approaches in estimating the distribution of the individual-specific heterogeneity. Results similar to these are also obtained

for all other values of T, where, as expected, the performance of all estimators improves with increasing T. For $T = 12$, the differences between estimators are very small. However, even there, BR produces visibly better estimates in the more difficult cases, such as the tails of the normal distribution or the discreteness of the Bernoulli distribution. In the Appendix, Figures E2 and E3 show the complete results for $T = 2$ and $T = 12$.

The primary motivation of our paper was the development of methods for estimating each α_i. In some applications, it might be sufficient to obtain statistics of the distribution of α_i, such as the first and second moments. Table 1 reports results for the three estimators considered. For each of the four distributions of α_i, the table lists the true mean and standard deviation of the $\alpha_i's$, as well as the estimated mean and standard deviations, averaged over the 500 replications. Here, too, the best-performing estimator in terms of mean is BR. It is interesting to observe that for short panels (in particular for $T = 2$), the BR estimators underestimates the true variance of the fixed effects. This is a result of shrinkage that was already visible in Figure 2, where we saw that BR does not work perfectly in the tails of the distribution. Reassuringly however, the problem becomes already quite minor for $T = 4$. Moreover, the other estimators tend to perform worse than BR on that dimension as well.

Figure 3 returns to the type of plots of Figure 2 and documents the improvement in the estimation of the distribution of α_i as T increases for the BR estimator. For the most part, it seems that four time periods are enough to provide a good approximation to the true distribution, but eight or twelve periods might be needed to estimate the tails of the distributions without any distortion.

3.3 Mean squared error

Lower small sample bias may not be a desirable property if it comes at the expense of an increased variance of the estimator. Figure 4 therefore focusses on the simulation root mean squared error (RMSE) for each α_i, where

$$\widehat{\text{RMSE}}(\hat{\alpha}_i) = \sqrt{\frac{1}{500}\sum_{r=1}^{500}\left(\hat{\alpha}_i^{(r)} - \alpha_i\right)^2}.$$

To highlight the consequences of perfect prediction in this context, we sorted the total of 300 RMSEs (for $i = 1, \ldots, 100$ and for each of the three estimators) by the mean dependent variable of unit i across the 500 replications:

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

The average $\bar{s}_i$ gives an indication of the severity of the perfect prediction for individual i. 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 Figure 4, circles show the RMSE of the maximum likelihood estimator, squares that of HS and diamonds that of BR, and the lines are corresponding kernel-weighted local polynomial regressions. As expected, the RMSE is highest for those α_i whose average s_i is close to zero or one, as can be seen from the U-shape of the nonparametric fit through the RMSE of $\hat{\alpha}_i$ for each of the three estimators. As T increases, the RMSE falls for all estimators, which we account for by adjusting the y-scale in the panels. In all cases, the RMSE of $\hat{\alpha}_i$ estimated by BR is always substantially smaller than the RMSE of ML and HS. Its U-shape is also much flatter, indicating its robustness to perfect prediction. Indeed, for $T = 8$ and $T = 12$, there seems not to be any difference in RMSE of BR for different values of s_i.

3.4 Estimation of β

Table 2 presents means and standard deviations of the estimated $\hat{\beta}$ across different distributions of α_i and different numbers of time periods. The true value is 1. The corresponding entries in the table confirm the incidental parameters bias of the ML estimator. 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 for β. Already 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.

Figure 5 plots, for each estimator and distribution of α_i, the sampling distribution of $\hat{\beta}$, based on the 500 replications and using a kernel density approximation, for $T = 4$. All three sampling distributions of $\hat{\beta}$ are reasonably close to normal. However, only BR is centred around the true value of $\beta = 1$, and it also is the distribution with the smallest dispersion.

Finally, we investigate estimation of some quantities which involve both β and all the α_i. We focussed on average predicted probabilities. Differences between such probabilities for different values of x_i give average effects, which are often of direct interest in empirical studies. Table 3 gives average predicted probabilities at the observed values of x_i (Columns "Mean") as well as at the first, fifth and ninth decile of the distribution of x_i (Columns "D_1", "D_5" and "D_9") for $T = 2$ and $T = 4$. The true average predicted probabilities are contained in Rows "True",

while the remaining rows show the estimates relative to these true values. We see that while all estimators give reasonable estimates for the estimates at the observed values of x_i, ML and HS often result in misleading estimates at other values of x_i. BR not only gives the best estimates at the observed values of x_i, but it is also often able to improve substantially in the other cases as well. Corresponding results for $T = 8$ and $T = 12$ are given in Table E1 in the appendix. With more time periods the bias for all estimators is reduced and differences in the ranking of their performance are less stable. Thus, BR seems to be useful for estimation of such derived quantities, and especially so for small T where it outperforms the other estimators. We also estimated the model by OLS. For the case of α_i following a Bernoulli distribution with $T = 2$ and $T = 4$, OLS gave more than 50 per cent of its predicted probabilities out of the unit interval. Indeed, both $D_1 < 0$ and $D_9 > 1$. Results were similar for the other distributions of α_i.

4 Application to hospital readmissions

Hospital readmissions have been identified as a major driver of health care costs. If patients are discharged too early after hospitalisation, readmission rates will be higher than would otherwise be the case (e.g., Heggestad, 2002). For the US, the aggregate costs of "excessive" readmissions have been estimated to be in the order of $1 billion per year (Jencks, Williams and Coleman, 2009). While costly at the aggregate level, an early discharge or not offering sufficient post-discharge care can be rational from the point of view of an individual hospital when reimbursements are based on diagnosis-related groups [DRG] rather than actual costs.

In an attempt to have hospitals internalise the costs of readmissions, the 2010 Affordable Care Act [ACA] established a financial penalty for hospitals whose Medicare readmission rates exceed a certain threshold in three common emergency conditions. In the financial year 2013, the Hospital Readmission Reduction Program [HRRP], a part of the ACA, started to reduce Medicare reimbursements for high-readmission hospitals. In the following years, aggregate readmission rates fell. McIlvennan, Eapen and Allen (2015) report a drop in the overall 30-day readmission rate for all causes from around 19 percent in 2010 to under 18 percent in 2013.

Our empirical analysis is based on administrative data for the years 2012–2016. We use separate panel probit models to estimate the determinants of a penalty for each emergency condition using the BR estimator. A first question that we can explore with our approach concerns risk adjustment. The imposition of a penalty under the HRRP depended on the hospital's actual readmission rate during a reference period as well as on a threshold value. The threshold is calculated as the average readmission rate of hospitals with a comparable case mix as defined by age, gender and

co-morbidities. With perfect risk adjustment, controls for relevant socio-demographic and health characteristics of a hospital's patients should be orthogonal to that hospital's propensity of being fined. This is a prediction about the $\boldsymbol{\beta}$ vector that can be tested.

Importantly, our approach also delivers estimates of hospital fixed effects for each emergency condition. This matters in an application such as this one, where the fixed effects are not just nuisance parameters that do not have any significance beyond avoiding omitted variable bias when estimating $\boldsymbol{\beta}$. There are a number of reasons why the fixed effects are of intrinsic interest. First, fixed effects provide a ranking of hospitals by their propensity of receiving readmission penalties. Similar rankings are common in the literature as indicators of hospital quality (Joynt and Jha 2013; Herrin et al. 2015; for a recent review, see Fischer et al. 2014), although this requires the assumption of absence of confounders. Second, having estimates of the fixed effects for each emergency condition allows us to study correlation patterns in rankings across conditions. Positive correlation patterns would suggest the presence of common, hospital-specific underlying causes of penalties, such as in the overall quality assurance systems. In contrast, a negative correlation might indicate specialisation in the treatment of a particular condition, or "competing risks", where a hospital which invests in reducing readmissions by targeting one condition increases readmissions in another.

Third, fixed effects can reflect heterogeneity in strategic decision making by hospitals. While penalties affect the tradeoff between treatment cost and readmission probability, it is by no means clear whether none, or very few, readmissions are optimal from the point of view of a hospital. If the costs of avoiding readmission are very high, optimising behaviour will tolerate some penalties up to the point where marginal costs are equalised. Systematic differences in the penalty likelihood between hospitals can be indicative of differences in the trade-off, or the way it is evaluated, by these hospitals. Of particular interest, for instance, is whether for-profit hospitals systematically differ in this regard from non-profit hospitals. This can be tested by regressing the estimated fixed effects on hospitals' time invariant characteristics, including hospital type (for-profit/non-profit).

4.1 Hospital Readmission Reduction Program

The Hospital Readmission Reduction Program became first effective for the financial year 2013. The penalty consists of reduced rates of Medicare reimbursements for those hospitals whose past readmission rates following emergency conditions among Medicare patients are "too high" during a three-year reference period. Initially, the reductions amounted up to one per cent of total Medicare reimbursements. They were increased to three per cent later on, resulting in aggregate penalties of about three hundred million dollars in 2013 and over half a billion in 2017 (Boccuti and Casillas,

2017). The program applies to three emergency conditions: heart attack or *acute myocardial infarction* (AMI), *congestive heart failure* (HF), and *pneumonia* (PN).

To determine the number of readmissions per hospital, a 30-days window after release was applied, and any admission to the same or any other hospital for the same or any other condition was counted (all-cause). The hospital specific rate, 30-day readmissions divided by the number of discharges, was then compared to the average rate of hospitals with a similar case-mix. Risk adjustment was based on age, sex, and co-morbidities. The risk adjustment did not take into account differences in socio-economic characteristics of the case-mix, nor of the communities' patient pool. Hospitals with above average readmission rates were subject to a penalty.

To assess the effects of the policy, we use administrative Hospital Compare Data for information on penalties announced in each July of the years 2012–2016.[7] Reporting is delayed by one year, so the data relate to the three-year aggregates of readmissions during the years 2011–2015. For each of the 3,135 included hospitals and each of the three emergency conditions, we know whether or not a penalty was issued.

To these primary data, we add geographic and hospital specific information (urban/rural, teaching status of hospital, number of beds) from the corresponding final rule impact files, an approach which broadly follows Gu et al. (2014). Based on hospital referral regions (HRR) provided by the Dartmouth Atlas of Health Care, we merge the number of ambulatory-care-sensitive conditions (ACSC), measuring accessibility of local primary health care (Gu et al., 2014), and the number of hospitals in the region, a local competition measure (Chandra et al., 2016). Lastly, we use the Federal Information Processing Standard (FIPS) to add county-wide community characteristics, such as the poverty rate and the median household income, which have been discussed as determinants of readmission rates outside the control of the hospital (cf. Herrin et al., 2015). Detailed variable descriptions, as well as descriptive statistics for our sample by condition and penalty status, are reported in Table F1 in the Appendix.

4.2 Results

Let $y_{it}^c \in \{0,1\}$ denote the imposition of a penalty for hospital i in condition c at time t. For each condition c, we specify a probit model of the form

$$\Pr(y_{it}^c = 1 | \alpha_i^c, \boldsymbol{x}_{it}^c) = \Phi\left(\alpha_i^c + \boldsymbol{x}_{it}^{c\prime}\boldsymbol{\beta}^c\right), \tag{14}$$

[7] The data can be found at `https://data.medicare.gov/data/hospital-compare`. All employed datasets are public use files; more detail on the data construction is presented in Appendix F.1.

where α_i^c is a hospital and condition specific fixed effect, $\boldsymbol{x}_{it}^c$ a covariate vector including time fixed effects and county-level variables, and $\boldsymbol{\beta}^c$ a conformable vector of condition-specific regression parameters.

Estimation of these three panel probit models for $c = \{AMI, HF, PN\}$, each with dummy variables for every hospital, by maximum likelihood would be subject to the problems discussed above: (1) $\hat{\alpha}_i^c$'s are non-existent in case of concordant observations (hospitals which were either always or never fined in the sample period), (2) few time periods, here five, mean that the $\hat{\alpha}_i^c$'s suffer from small sample bias, and (3) this also biases estimates of the common parameter, $\hat{\boldsymbol{\beta}}^c$.

In our application, problem (1) is of particular relevance, since roughly 50 percent of hospitals do not change their penalty status during the sample period. One natural reason for the high level of persistence in the dependent variable over time is that the three-years reference windows determining the penalty status do overlap.[8] Table 4 shows the estimation results for the standard as well as the bias-reduced (BR) probit models for each condition. A comparison of the standard probit and the BR probit results highlights the incidental parameters bias of the standard estimation method: almost all standard probit coefficient estimates are substantially larger than the ones using the BR estimator, illustrating that the standard probit estimator should not be used in this setting.

Turning to the BR probit results, we find that having a large number of condition-specific discharges increases the penalty-propensity significantly across conditions. The size of the emergency department, as measured by the number of discharges in the respective non-c conditions, does not affect the condition-specific penalty probabilities equally across diagnostic conditions.

Next, we use hospital referral regions (HRR) to assess the effects of local health care provisions, first by the number of discharges with ambulatory care sensitive conditions (Gu et al., 2014) and second by the competition in the local health care market (Bloom et al., 2015; Chandra et al., 2016; Gobillon and Milcent, 2017). We find that the number of discharges with ACSC increases the probability of being fined across the three conditions. ACSC measure potentially preventable medical problems, such as hypertension, which with proper medication and management of care should be treatable outside of a hospital. The average marginal effect of increasing ACSC by 1 standard deviation—roughly 15 discharges per 1,000 enrolees—on the penalty-probability amounts to 7.1 percentage points for heart attack DRGs, 1.5 for heart failure, and 2.1 for pneumonia.[9]

These effects are both economically and statistically relevant determinants of the penalty propensity. In contrast, the entry of competitors in the HRR significantly decreases readmission risk only

[8] We account for this overlap when computing standard errors by clustering at the hospital level.

[9] Marginal effects at the average can easily be obtained by multiplying the coefficient with $\overline{\phi(\boldsymbol{x}'\boldsymbol{\beta})}$, which is reported at the bottom of the table. E.g., for heart attack, $\overline{\phi(\boldsymbol{x}'\boldsymbol{\beta})} \times \beta \times sd_x = 0.34 \times 0.014 \times 15 = 0.071$.

for heart attack patients, but not for those experiencing a heart failure or pneumonia. The exit of competitors in the local market is negatively, yet insignificantly, associated with penalty risk. Consequently, we do not find evidence that competitive forces or disruptions in the local hospital market impact the readmission risk in a meaningful way. One simple explanation is that these forces take some time to materialise and the aggregation to three-year reference periods makes them difficult to observe in this setting. In contrast to the cross-sectional evidence presented by Herrin et al. (2015), we do not find much evidence that socioeconomic community characteristics influence the penalty risk (over and above those captured by the other factors). Among the economic county characteristics, only median household income is significantly related to the readmission risk of heart attack diagnosed patients.

Figure 6 compares the estimated hospital-specific fixed effects across diagnostic conditions: A positive correlation indicates that a hospital that performs poorly in one emergency condition is likely to perform poorly in another condition as well. Such a situation might be due the hospital's management choice of a low value of care across (emergency) conditions due to common, hospital-specific marginal costs of providing high quality care. Provision of care for different conditions would then be complements rather than substitutes.

We find that the correlation is stronger between heart failure and pneumonia (with a regression slope of 0.67), than with respect to heart attacks (0.43 between AMI and HF; and 0.39 between AMI and PN). The grey line displays the regression ignoring observations (the grey points) which are concordant in at least one of the two conditions—and which thus would not be estimable using other estimation approaches. Relative to the regression using all observations, slopes are biased downwards for AMI-by-HF and AMI-by-PN correlations and upwards for HF-by-PN. All distributions exhibit a long tail in the negative domain, meaning that there are some hospitals that have a much lower penalty propensity (across years and conditions) than could be expected based on their observed characteristics.

The positive correlations in Figure 6 seem to indicate the presence of a common hospital-specific component in the α_i^c. On the other hand, the fact that there is considerable dispersion around the regression lines in the figure suggests that condition-specific components might be important as well. The model we have in mind is

$$\alpha_i^c = \alpha_i + \tau_i^c,$$

where α_i is the common component and τ_i^c the condition-specific component of the time-invariant effect α_i^c. Because each α_i^c is estimated based on a small number of years and thus potentially with low precision, the dispersion around the regression line might only reflect sampling error rather than

the presence of τ_i^c. To formally test for the presence of condition-specific heterogeneity we consider the null hypothesis $H_0 : (N-1)^{-1/2}\sqrt{\sum_i(\alpha_i^c - \alpha_i^{c'})^2} = 0$. By differencing across two conditions c and c' we remove the common component. We then tests for any remaining variation, which by definition must be due to condition-specific components, using a Monte Carlo permutation test (see Abrams, Bertrand and Mullainathan, 2012).

The permutation distributions of the test statistics are shown in Figure 7 for each of the three paired conditions. In each panel, the sample test statistic and the 95 percentile of the distribution are indicated by vertical lines. We do not find evidence, at the 5 per cent significance level, of condition-specific differences between heart infarct and heart failure (left panel), nor between heart infarct and pneumonia (middle panel). Only for heart failure and pneumonia (right panel) do we see clear evidence for condition-specific heterogeneity.

In a final step, we regress the hospital-specific fixed effects on other time-invariant characteristics. For instance, to test whether for-profit hospitals behave differently than other hospitals, one can run the following regression

$$\alpha_i^c = \gamma^c \text{for-profit}_i + \boldsymbol{z}_i^{c'}\boldsymbol{\delta}^c + u_i^c, \tag{15}$$

the estimated fixed effects are used as dependent variable, and γ^c measures the difference between for-profit and non-profit hospitals. Here, $\boldsymbol{z}_i^c$ is a condition- and hospital-specific covariate vector, $\boldsymbol{\delta}^c$ its corresponding coefficient vector, and u_i^c an error term. Table 5 presents the estimation results.

We find that for-profit hospitals indeed have a significantly larger time-constant penalty propensity across the three emergency conditions. This suggests that for-profit hospitals chose a different level of care resulting in higher readmissions, potentially due to cost-benefit considerations by the hospital's owners and managers. These differences change only minimally when accounting for other time invariant hospital characteristics identified in the literature on the readmission determinants (e.g., Gu et al., 2014).

5 Conclusions

This paper studied the use of 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 excessive readmission rates among Medicare patients in the US. We advocated an estimator based on Kosmidis and Firth (2009), which had not been adapted to the context of panel

data so far, and for which we showed that it always produces finite estimates of all fixed effects. This feature is essential if the interest lies in the distribution of the unobserved heterogeneity, and it also facilitates the estimation of derived quantities which depend on the fixed effects. In our simulations, the estimator performed better than either the MLE or HS estimator.

Perfect prediction is a problem which is very common in applications, especially in short and very short panels. In the data of our application—an unbalanced panel covering a five-year period—about half 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.

Our simulations showed that estimators which fail to obtain estimates for all fixed effects can give severely distorted estimates of the shape of the distribution of the fixed effects and of moments such as mean and variance. Using the advocated BR estimator is a simple and effective way of reducing such distortions. In our empirical application, we illustrated several ways in which estimates of the fixed effects can be used to answer economic questions. For instance, we plotted the joint distribution of the estimated fixed effects from different models to see whether care for different health conditions behaved as substitutes or complements, and we regressed the estimated fixed effects on time-invariant regressors to answer questions about differences in strategic behaviour by hospital ownership type.

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

Abadie, Alberto and Maximilian Kasy. 2016. "The risk of machine learning." *Harvard University OpenScholar Working Paper No. 383316.*

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.

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

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.

Bartolucci, Francesco, Ruggero Bellio, Alberto Salvan and Nicola Sartori. 2016. "Modified profile likelihood for fixed-effects panel data models." *Econometric Reviews* 35(7):1271–1289.

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.

Boccuti, Cristina and Giselle Casillas. 2017. "Aiming for fewer hospital U-turns: the Medicare hospital readmission reduction program." *Policy Brief, March 2017 (update), The Henry J. Kaiser Family Foundation.*

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

Carro, Jesús M. and Alejandra Traferri. 2014. "State Dependence And Heterogeneity In Health Using A Bias-Corrected Fixed-Effects Estimator." *Journal of Applied Econometrics* 29(2):181–207.

Chandra, Amitabh, Amy Finkelstein, Adam Sacarny and Chad Syverson. 2016. "Health Care Exceptionalism? Performance and Allocation in the US Health Care Sector." *American Economic Review* 106(8):2110–2144.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. 2014. "Measuring the impacts of teachers 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. 1968. "A general definition of residuals." *Journal of the Royal Statistical Society. Series B (Methodological)* 30(2):248–275.

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.*

Fernández-Val, Iván. 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.

Fischer, Claudia, Hester F. Lingsma, Perla J. Marang-van de Mheen, Dionne S. Kringos, Niek S. Klazinga and Ewout W. Steyerberg. 2014. "Is the readmission rate a valid quality indicator? A review of the evidence." *PloS One* 9(11):e112282.

Gobillon, Laurent and Carine Milcent. 2017. "Competition and hospital quality: Evidence from a French natural experiment." *IZA Discussion Paper No. 10476.*

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.

Gu, Qian, Lane Koenig, Jennifer Faerberg, Caroline Rossi Steinberg, Christopher Vaz and Mary P. Wheatley. 2014. "The Medicare Hospital Readmissions Reduction Program: potential unintended consequences for hospitals serving vulnerable populations." *Health Services Research* 49(3):818–837.

Hahn, Jinyong and Guido Kuersteiner. 2002. "Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large." *Econometrica* 70(4):1639–1657.

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. Honore. 1990. "The empirical content of the Roy model." *Econometrica* 58(5):1121–1149.

Heggestad, Torhild. 2002. "Do Hospital Length of Stay and Staffing Ratio Affect Elderly Patients' Risk of Readmission? A Nation-wide Study of Norwegian Hospitals." *Health Services Research* 37(3):647–665.

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

Herrin, Jeph, Justin St. Andre, Kevin Kenward, Maulik S. Joshi, Anne-Marie J. Audet and Stephen C. Hines. 2015. "Community factors and hospital readmission rates." *Health Services Research* 50(1):20–39.

Hospido, Laura. 2012. "Modelling heterogeneity and dynamics in the volatility of individual wages." *Journal of Applied Econometrics* 27(3):386–414.

Jencks, Stephen F., Mark V. Williams and Eric A. Coleman. 2009. "Rehospitalizations among patients in the Medicare fee-for-service program." *New England Journal of Medicine* 360(14):1418–1428.

Joynt, Karen E. and Ashish K. Jha. 2013. "Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program." *Journal of the American Medical Association: Research Letters* 309(4):342–343.

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.

Kosmidis, Ioannis, Kjell Konis, Euloge C. K. Pagui and Nicola Sartori. 2017. "brglm2: Bias Reduction in Generalized Linear Models.".
URL: *https://cran.r-project.org/web/packages/brglm2/index.html*

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.

McIlvennan, Colleen K., Zubin J. Eapen and Larry A. Allen. 2015. "Hospital readmissions reduction program." *Circulation* 131(20):1796–1803.

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).

Woutersen, Tiemen. 2001. "Robustness against incidental parameters and mixing distributions." *Research Report, Department of Economics, University of Western Ontario*.

Figures and Tables

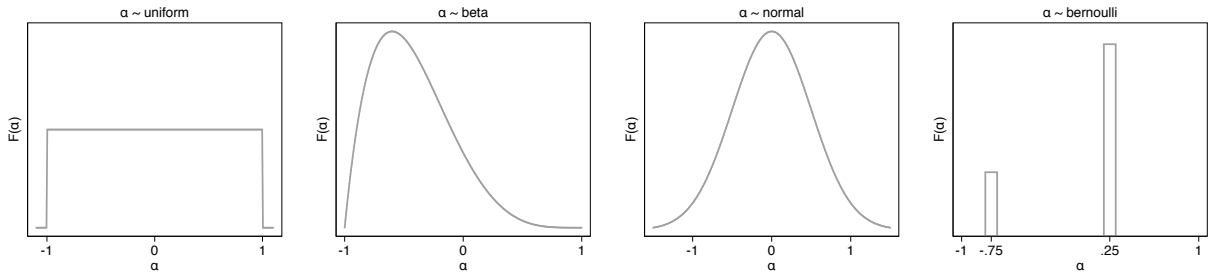


Figure 1: Distributions of α_i

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.

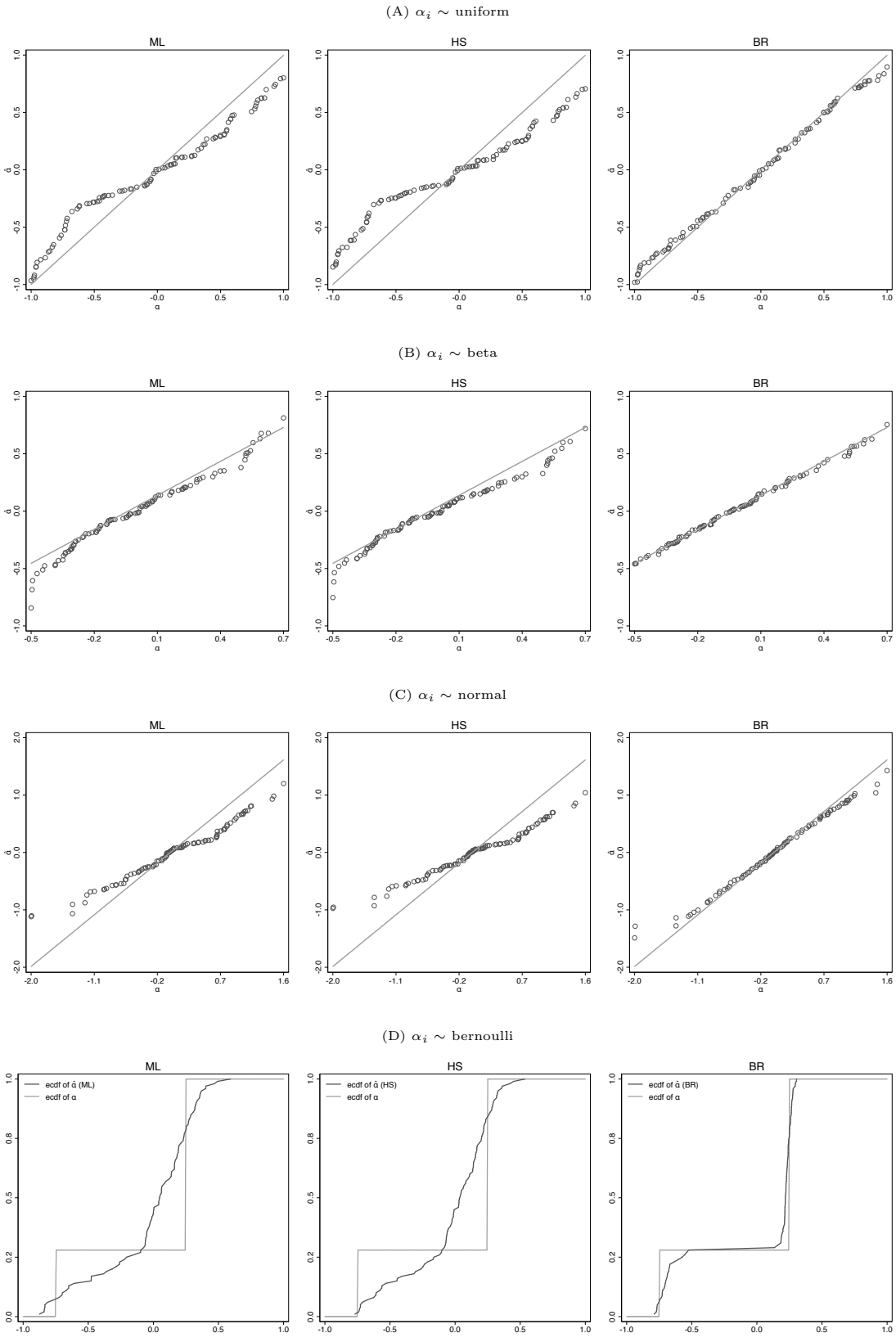


Figure 2: Estimated versus true distributions of α_i, N=100, T=4

Notes: Graphs in panels (A), (B), (C) show average estimates of $\alpha_1, \ldots, \alpha_{100}$ over 500 replications against their true values. Graphs in panel (D) show the empirical cdf of the hundred true α_i against the empirical cdf of the hundred average $\hat{\alpha}_i$ estimated over 500 replications.

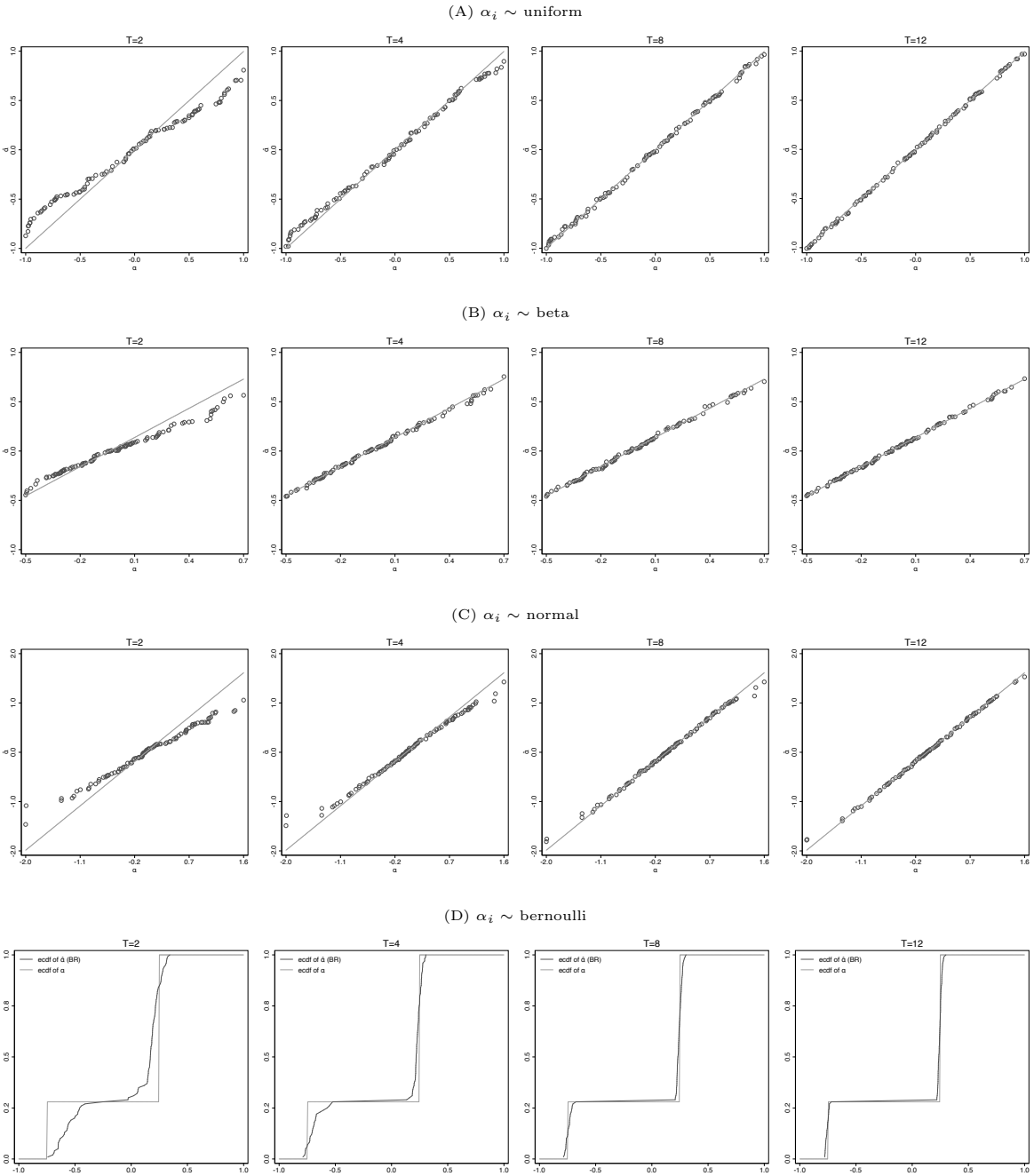


Figure 3: Estimated versus true distributions of α_i, N=100, BR Likelihood Estimator

Notes: Graphs in panels (A), (B), (C) show average estimates of $\alpha_1, \ldots, \alpha_{100}$ over 500 replications against their true values. Graphs in panel (D) show the empirical cdf of the hundred true α_i against the empirical cdf of the hundred average $\hat{\alpha}_i$ estimated over 500 replications.

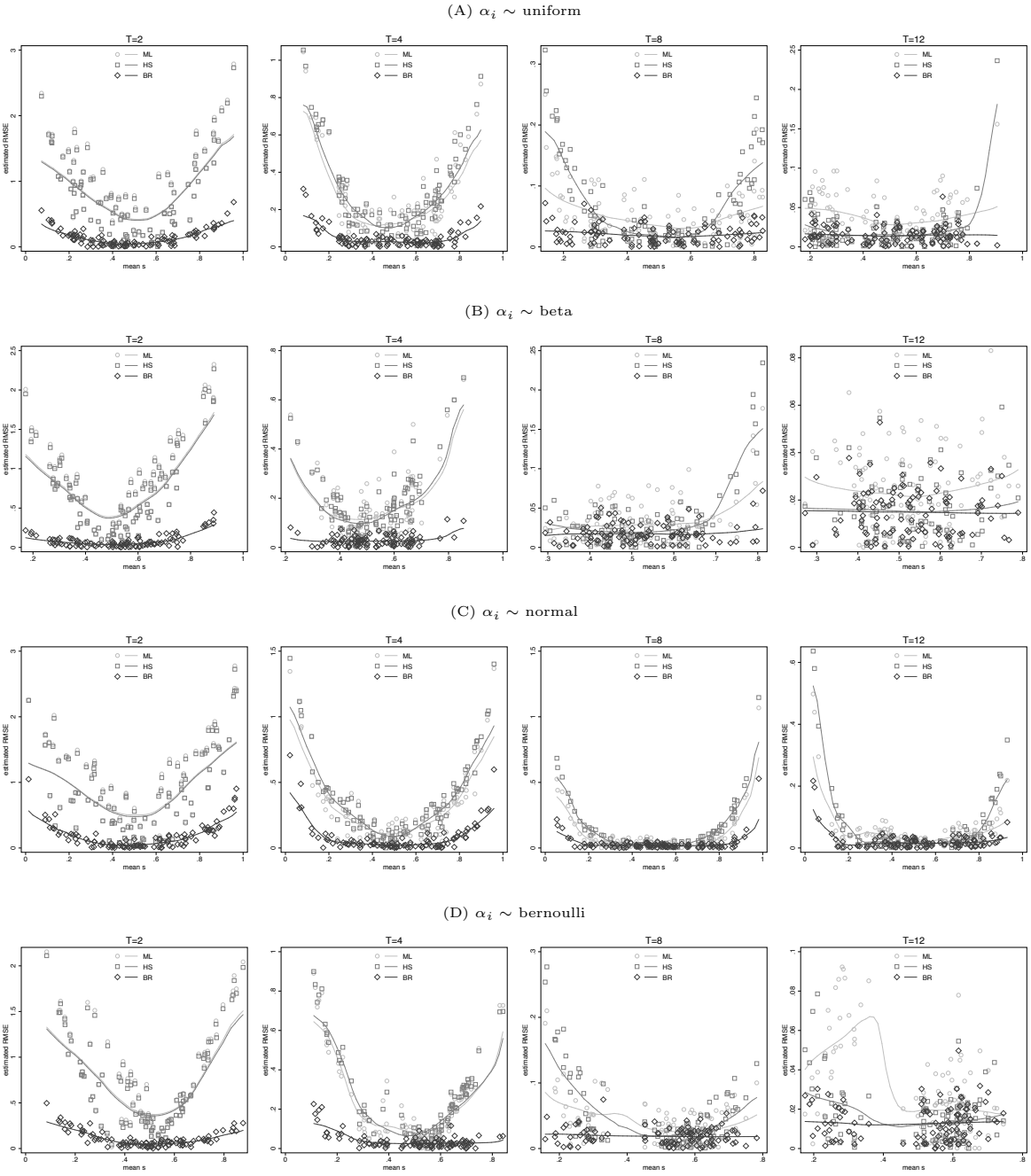


Figure 4: Root Mean Square Error of α_i and Incidence of Perfect Prediction

Notes: Graphs show RMSE estimated for each $\hat{\alpha}_i$ and for each of the three estimators Maximum Likelihood (ML), HS Penalised Likelihood (HS), and BR Likelihood (BR); against the average share of $y_{it} = 1$ within i, over 500 replications. The share of $y_{it} = 1$ within i is $s_i = \sum_{i=1}^{T} y_{it}/T$. The lines represent kernel-weighted local polynomial regressions of each estimator's RMSE on the average s_i. Scales of y-axes adjust for the reductions in the RMSE.

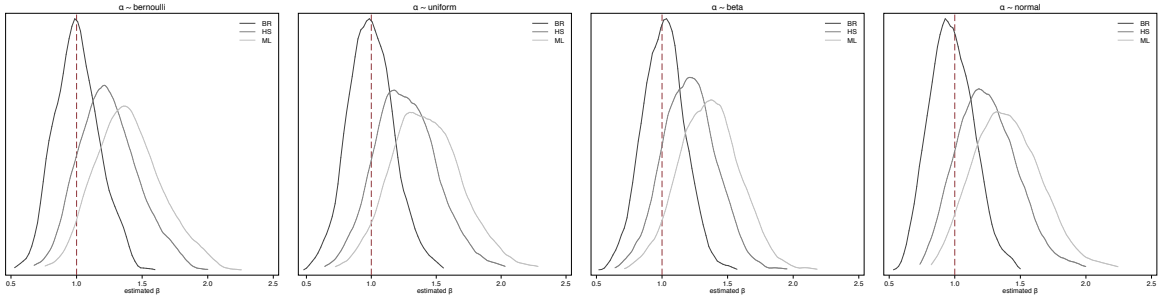


Figure 5: KERNEL DENSITY ESTIMATES OF $\hat{\beta}$ ($\beta = 1$) FOR FOUR DGP WITH $N = 100$ AND $T = 4$

Notes: Graphs show estimated kernel density of $\hat{\beta}$ based on 500 replications for the three estimators, BR Likelihood (BR), HS Penalised Likelihood (HS) and Maximum likelihood (ML). The vertical maroon dashed line represents the true value of β.

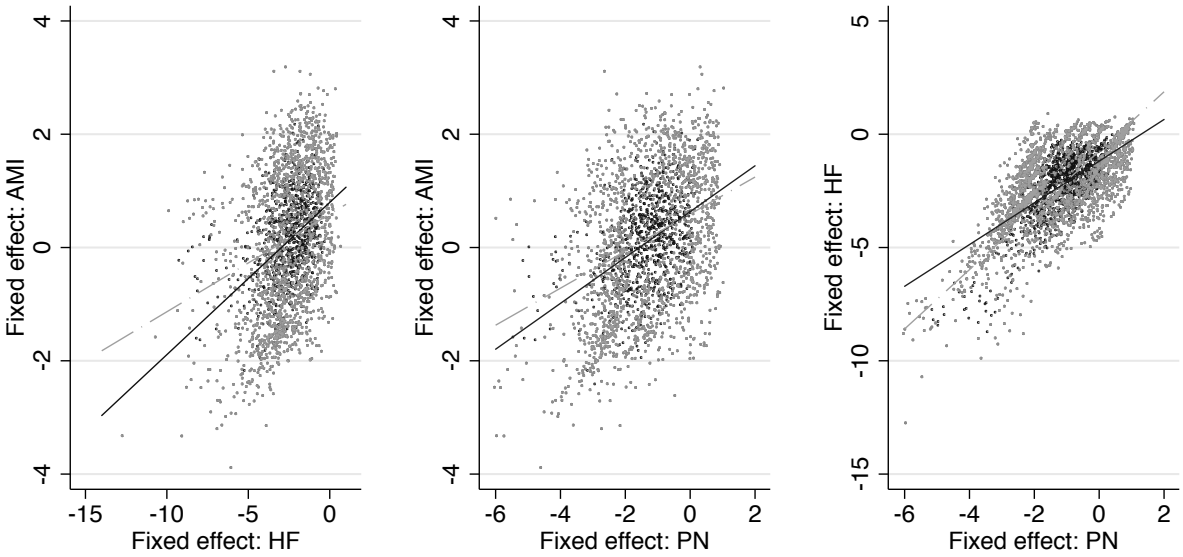


Figure 6: PAIR-WISE SCATTER PLOTS OF HOSPITALS UNOBSERVED HETEROGENEITY IN READMISSION PENALTIES

Notes: The panels plot estimates of α_i^c based on (14) and Table 4, pairwise for all combination of $c = \{AMI, HF, PN\}$, which denote acute myocardial infarction (AMI), heart failure (HF), or pneumonia (PN). The solid black line shows the correlation using all points; dashed grey line shows the correlation among (bias reduced) discordant pairs. Although overlying greatly, black dots depict all fixed effects and grey dots only those that are concordant in at least one condition (hence not estimable by a naïve probit).
Source: Hospital Compare Dataset and Final Rule Impact files 20012-2016, ACS, Dartmouth Atlas of Health Care, own calculations.

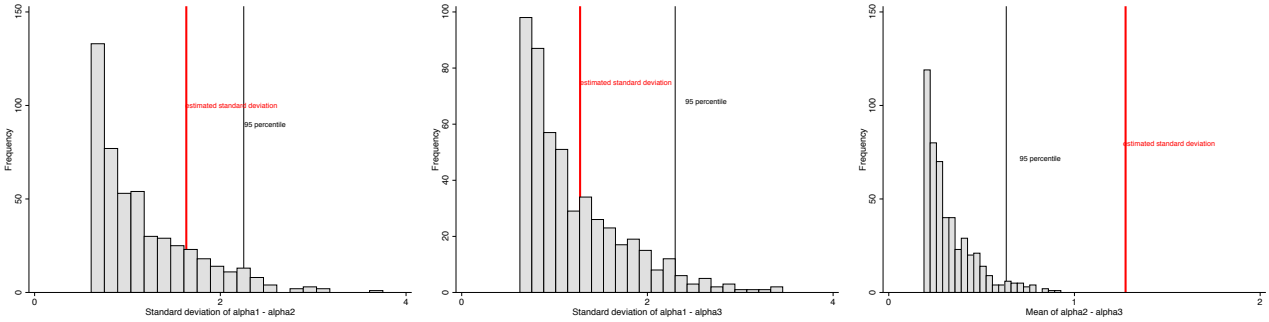


Figure 7: Monte Carlo test for condition-specific heterogeneity

Notes: For pairs $c, c' = \{AMI, HF, PN\}$, the vertical black line represents the test statistic $\hat{\sigma}(\alpha_i^c - \alpha_i^{c'}) = ((N-1)^{-1}\sum_i(\hat{\alpha}_i^c - \hat{\alpha}_i^{c'})^2)^{1/2}$. The null hypothesis is $\sigma(\alpha_i^c - \alpha_i^{c'}) = 0$. The histogram is an estimate of the distribution of the test statistic under the null hypothesis, calculated based on 500 random permutations of the dependent variable. The vertical red line is the 90th percentile of this distribution.
Source: Hospital Compare Dataset and Final Rule Impact files 20012-2016, ACS, Dartmouth Atlas of Health Care, own calculations.

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: MC Simulation: Predicted probabilities, averaged over distribution of α_i (500 replications)

	$T=2$				$T=4$			
	Mean	D_1	D_5	D_9	Mean	D_1	D_5	D_9
$\alpha_i \sim Bernoulli$								
True	*0.494*	*0.227*	*0.495*	*0.760*	*0.494*	*0.227*	*0.495*	*0.760*
ML	1.040	0.474	1.061	1.192	1.007	0.791	1.008	1.074
HS	1.040	0.498	1.060	1.186	1.006	0.864	1.006	1.052
BR	1.001	1.218	1.000	0.940	1.001	1.134	1.000	0.964
$\alpha_i \sim Uniform$								
True	*0.487*	*0.235*	*0.485*	*0.741*	*0.487*	*0.235*	*0.485*	*0.741*
ML	1.059	0.434	1.082	1.229	1.002	0.734	1.000	1.089
HS	1.058	0.454	1.081	1.223	1.003	0.804	1.001	1.067
BR	1.000	1.179	1.002	0.944	1.004	1.115	1.005	0.969
$\alpha_i \sim Beta$								
True	*0.512*	*0.231*	*0.512*	*0.788*	*0.511*	*0.231*	*0.512*	*0.788*
ML	0.934	0.418	0.917	1.109	0.971	0.790	0.966	1.029
HS	0.935	0.439	0.918	1.103	0.970	0.864	0.966	1.006
BR	0.985	1.224	0.982	0.918	0.994	1.132	0.992	0.955
$\alpha_i \sim Normal$								
True	*0.516*	*0.273*	*0.518*	*0.754*	*0.516*	*0.273*	*0.518*	*0.754*
ML	0.931	0.339	0.918	1.170	0.970	0.680	0.967	1.080
HS	0.930	0.357	0.919	1.163	0.969	0.738	0.965	1.058
BR	0.978	1.111	0.975	0.934	0.992	1.069	0.990	0.967

Notes: Entries in rows "True" are mean predicted probabilities. Entries in other rows are mean predicted probabilities divided by the value in the corresponding "True" row. Entries in Columns "Mean" are mean predicted probabilities marginal of x. Entries in Columns "D_1", "D_5" and 'D_9" are mean predicted probabilities evaluated at the first, fifth (median) and ninth decile of x.

Table 4: Probit and bias-reduced fixed effects probit models for penalty status by condition

Dependent variable: readmission penalty indicator						
	Probit			BR-Probit		
	AMI	*HF*	*PN*	*AMI*	*HF*	*PN*
	(1)	(2)	(3)	(4)	(5)	(6)
Number of condition-specific discharges per 1'000	1.939	3.863	0.682	1.306	2.512	0.452
	(0.902)	(0.691)	(0.320)	(0.675)	(0.463)	(0.229)
Total number of discharges other conditions per 1'000	0.350	−0.433	0.856	0.232	−0.251	0.599
	(0.272)	(0.345)	(0.429)	(0.206)	(0.234)	(0.304)
Discharges ACSCs per 1'000 enrollees, in HRR	0.020	0.017	0.009	0.014	0.011	0.006
	(0.007)	(0.006)	(0.006)	(0.005)	(0.004)	(0.004)
Hospital entry in HRR (Yes/No)	−0.144	−0.025	0.023	−0.093	−0.016	0.017
	(0.071)	(0.062)	(0.060)	(0.053)	(0.045)	(0.044)
Hospital exit in HRR (Yes/No)	−0.086	−0.004	−0.022	−0.055	−0.001	−0.017
	(0.065)	(0.056)	(0.055)	(0.049)	(0.041)	(0.042)
Household median income in 10'000$, in county	−0.367	0.109	−0.023	−0.242	0.074	−0.014
	(0.172)	(0.123)	(0.118)	(0.130)	(0.097)	(0.096)
Percent unemployed, in county	−0.071	−0.027	−0.019	−0.049	−0.021	−0.013
	(0.044)	(0.036)	(0.034)	(0.033)	(0.027)	(0.026)
Percent of total population living in poverty, in county	−0.010	−0.004	−0.010	−0.007	−0.002	−0.007
	(0.023)	(0.017)	(0.017)	(0.018)	(0.013)	(0.013)
Total population in 100'000, in county	0.013	0.079	0.038	0.012	0.063	0.034
	(0.113)	(0.100)	(0.091)	(0.089)	(0.081)	(0.074)
Number of observations	6,071	7,989	8,369	10,880	14,889	15,106
Number of hospitals	1,256	1,633	1,713	2,301	3,075	3,112
Share of concordant observations, 0				22.6	23.8	24.2
Share of concordant observations, 1				21.6	22.5	20.4
$\overline{\phi(x'\beta)}$				0.34	0.09	0.23
Hospital fixed effects	✓	✓	✓	✓	✓	✓
Time fixed effects	✓	✓	✓	✓	✓	✓

Notes: Coefficient estimates from bias reduced probit regressions, clustered standard errors in parentheses. First three columns standard probit with dummy variables, separately by DRG-condition: acute myocardial infarction (AMI), heart failure (HF), or pneumonia (PN). Columns (4)-(6) bias-reduced probit estimation. Penalty status is only defined if there were more than 25 discharges in the specific condition across three years. All regressions include hospital and year fixed effects. Exit and entry in the hospital referral region are constructed by increase (decrease) in the number of hospitals in the region, which are observationally identical to mergers or separations. All regressions include indicators for missing values, whose main effects are set to zero to avoid sample selection issues. Descriptive Statistics for the variables and their definitions are presented in Appendix, Table F1.
Source: Hospital Compare Dataset and Final Rule Impact files 20012-2016, ACS, Dartmouth Atlas of Health Care, own calculations.

Table 5: Hospitals' unobserved time-constant heterogeneity and ownership structure, OLS regressions

Dependent variable: Unobserved hospital fixed effects by condition						
	α_i^{AMI}		α_i^{HF}		α_i^{PN}	
	(1)	(2)	(3)	(4)	(5)	(6)
For-profit hospital (Yes/No)	0.189	0.144	0.206	0.149	0.167	0.153
	(0.059)	(0.059)	(0.081)	(0.072)	(0.056)	(0.054)
Number of beds (100 - 400, Yes/No)		−0.254		−0.663		−0.181
		(0.070)		(0.062)		(0.050)
Number of beds (>400, Yes/No)		−0.889		−1.836		−0.768
		(0.096)		(0.113)		(0.087)
Minor teaching hospital (Yes/No)		−0.298		−0.439		−0.228
		(0.061)		(0.079)		(0.062)
Major teaching hospital (Yes/No		0.398		0.272		0.371
		(0.074)		(0.095)		(0.070)
Located in urban area (Yes/No)		0.213		−0.751		−0.299
		(0.066)		(0.059)		(0.051)
Number of hospitals	2,301	2,301	3,075	3,075	3,111	3,111

Notes: OLS regression coefficients and robust standard errors in parentheses. Conditions are acute myocardial infarction (AMI), heart failure (HF), or pneumonia (PN). For-profit is an indicator for Physican Ownership, Physician, Proprietary, and Tribal hospitals; non-profit, the reference category, is a mix of voluntary and government hospitals. The omitted categories are 'number of beds 0-100', 'no teaching hospital', and 'located in rural area'. Teaching intensity is measured by the resident-to-bed and resident-to-average-daily-census ratio; minor teaching if ratio was between 0-0.25, major teaching if at least one ratio >0.25. Descriptive statistics for the variables are presented in Appendix, Table F1.
Source: Hospital Compare Dataset and Final Rule Impact files 20012-2016, ACS, Dartmouth Atlas of Health Care, own calculations.

Appendix

A Details on derivation of the bias in the Probit model

In general the small sample bias can be written as Cox and Snell (1968, equation 16)

$$Bias = \frac{1}{\kappa_{\alpha\alpha}^2}\left(\kappa_{\alpha,\alpha\alpha} + \frac{1}{2}\kappa_{\alpha\alpha\alpha}\right)$$

where $\kappa_{\alpha\alpha}^2$ is the square of the expected Hessian (total information in the sample), $\kappa_{\alpha,\alpha\alpha}$ is the expected product of Hessian and score, and $\kappa_{\alpha\alpha\alpha}$ the expected derivative of the Hessian.

In the our example of the probit model using $\alpha_i = \tilde{\alpha}_i + \bar{x}_i'\gamma$, these expected moments are equal to (see, e.g., Alexander and Breunig, 2016):

$$\begin{aligned}
\kappa_{\alpha\alpha} &= \sum_{t=1}^{T} \mathrm{E}(v_{it}^{\alpha}), \\
\kappa_{\alpha,\alpha\alpha} &= \sum_{t=1}^{T} \mathrm{E}(v_{it}v_{it}^{\alpha}), \\
\kappa_{\alpha\alpha\alpha} &= \sum_{t=1}^{T} \mathrm{E}(-v_{it} - \alpha_i v_{it}^{\alpha} - 2v_{it}v_{it}^{\alpha}) = -\alpha_i \sum_{t=1}^{T} \mathrm{E}(v_{it}^{\alpha}) - 2\sum_{t=1}^{T} \mathrm{E}(v_{it}v_{it}^{\alpha}),
\end{aligned}$$

where the last equality uses the fact that the expected score is equal to zero.

For T observations, this gives

$$\begin{aligned}
Bias &= \frac{1}{T^2\mathrm{E}(v_{it}^{\alpha})^2}\left(T\mathrm{E}(v_{it}v_{it}^{\alpha}) + \frac{1}{2}T\left(-\alpha_i\mathrm{E}(v_{it}^{\alpha}) - 2\mathrm{E}(v_{it}v_{it}^{\alpha})\right)\right) \\
&= \frac{1}{T^2\mathrm{E}(v_{it}^{\alpha})^2}\left(\frac{1}{2}T\left(-\alpha_i\mathrm{E}(v_{it}^{\alpha})\right)\right) \\
&= -\frac{1}{T}\alpha_i\mathrm{E}(v_{it}^{\alpha})^{-1}.
\end{aligned}$$

Since $\mathrm{E}(v_{it}^{\alpha}) = -\phi(\eta_{it})^2/(\Phi(\eta_{it})(1-\Phi(\eta_{it}))$ this is identical to the bias shown in equation (2) in the main text.

B Details on BR estimator for other binary response panel models

B.1 Modified score for α_i

For a general binary response fixed effects panel model with

$$P(y_{it} = 1|\boldsymbol{x}_{it}, \alpha_i) = F(\eta_{it}) = F(\alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta}) \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

where $F(\cdot)$ is a known distribution function, the modified score of the bias-reduced estimator for the parameter α_i is

$$s^{BR}(\alpha_i) = s^{ML}(\alpha_i) + \frac{1}{2}\sum_{t=1}^{T} h_{it}\frac{f'_{it}}{f_{it}},$$

where $f_{it} = f(\eta_{it})$ and $f'_{it} = f'(\eta_{it})$ are the first and second derivative of $F_{it} = F(\eta_{it})$ with respect to α_i, and h_{it} is the it-th diagonal elements of the $NT \times NT$ projection matrix

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

with $\boldsymbol{X}$ the $NT \times K$ matrix of the K regressors, and $\boldsymbol{W}$ is the $NT \times NT$ diagonal matrix with typical element

$$w_{it} = \frac{f_{it}^2}{F_{it}(1-F_{it})}.$$

The expressions for probit, $F_{it} = \Phi(\eta_{it})$, are given in Section 2.2. For logit, $F_{it} = \Lambda_{it} = \Lambda(\eta_{it}) = \exp(\eta_{it})/(1+\exp(\eta_{it}))$, and so

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

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

B.2 IWLS estimation

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

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

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

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

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

B.3 Existence in perfect prediction cases

In the perfect prediction cases, $\sum_t y_{it} = 0$ and $\sum_t y_{it} = T$, the ML estimator for α_i does not exist. We consider the case $\sum_t y_{it} = T$ for the BR estimator:

$$s^{BR}(\alpha_i) = \sum_{t=1}^{T}\frac{f_{it}}{F_{it}} + \frac{1}{2}h_{it}\frac{f'_{it}}{f_{it}}.$$

For all the usual choices of F_{it} (normal, logistic, cloglog, etc.), the log-likelihood $\ln f_{it}$ is globally concave, and the corresponding score, f'_{it}/f_{it}, has a unique root, is 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, $\sum_t \frac{1}{2} h_{it} f'_{it}/f_{it}$, 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 $h_{it} \in (0, 1]$. For such F_{it}, the first term,$\sum_t f_{it}/F_{it}$, 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 $s^{BR}(\alpha_i)$ is continuous, there must exist a $\hat{\alpha}_i$ such that $s^{BR}(\hat{\alpha}_i) = 0$.

A detailed example was given in Section 2.2 for $F_{it} = \Phi_{it}$. For logit, $F_{it} = \Lambda_{it}$, and

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

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, $\lim_{\alpha \to -\infty} s^{BR}(\alpha_i) > 0$. 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, $\lim_{\alpha \to -\infty} s^{BR}(\alpha_i) < 0$. Therefore, $s^{BR}(\hat{\alpha}_i) = 0$ exists.

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

C 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 $(\eta_{i1}, \eta_{i2}) \in \boldsymbol{R}^2$, 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 2.3. 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, $1 < x'_{i1}\boldsymbol{\beta} - x'_{i2}\boldsymbol{\beta} \leq 6$. We only consider the limit case, as the results for differences smaller than 6 follow immediately using the same arguments.[10] 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 $\alpha_i \in \boldsymbol{R}$ and thus that for $x'_{i1}\boldsymbol{\beta} - x'_{i2}\boldsymbol{\beta} = 6$ the estimator $\hat{\alpha}_i$ does not exist.

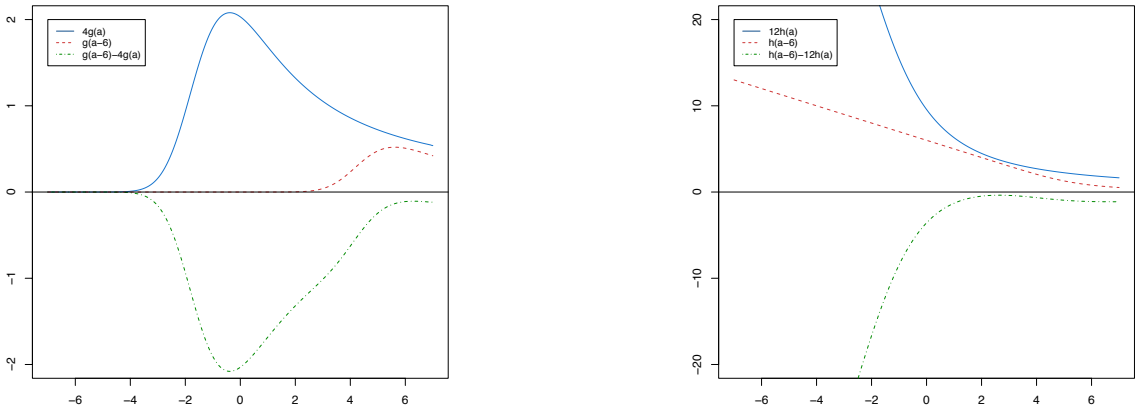
Figure C1 plots the two terms in brackets. Each one is negative over the entire plotted range (green dash-dotted lines). This holds in general as well. Consider the first term in brackets, $g(\eta_2) - g(\eta_1)$. It is straightforward to show that the function $g(\eta)$ is positive for all η and has 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

[10] This example is meant as an illustration. The interval $1 < x'_{i1}\boldsymbol{\beta} - x'_{i2}\boldsymbol{\beta} \leq 6$ is not a tight bound for the interval in which $\hat{\alpha}_i$ does not exist. However, differences such as $x'_{i1}\boldsymbol{\beta} - x'_{i2}\boldsymbol{\beta} = 6$ 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.

first term in brackets is strictly negative for all α_i. (In the left panel of Figure C1, 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 C1, this can be seen as the solid blue line, which depicts $12h(\alpha_i)$, always lies higher than the dashed red line, which depicts $h(\alpha_i)$.)

Figure C1: 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.

D Equivalence of Firth Penalised Likelihood Logit estimator and Bester-Hansen Penalised Likelihood "IE" Logit estimator

For generalised linear models in canonical parametrisation, Firth (1993) showed that first-order-bias-corrected scores lead to a penalised likelihood estimator. In panel data notation, this estimator is

$$Q^F(\boldsymbol{\theta}) = \sum_i \sum_t \log L_{it}(\boldsymbol{\theta}) + \pi^F(\boldsymbol{\theta}),$$

where $\log L_{it}(\boldsymbol{\theta})$ is the log-likelihood contribution of observation it, and $\pi^F(\boldsymbol{\theta})$ is the penalty term,

$$\pi^F(\boldsymbol{\theta}) = \frac{1}{2}\ln\left(\det \boldsymbol{I}(\boldsymbol{\theta})\right),$$

where $\boldsymbol{I}(\boldsymbol{\theta}) = \sum_i \sum_t (\partial \log L_{it}(\boldsymbol{\theta})/\partial\boldsymbol{\theta})(\partial \log L_{it}(\boldsymbol{\theta})/\partial\boldsymbol{\theta})'$ is the Fisher information matrix. Thus, for the logit panel model, we obtain the Firth penalised likelihood estimator as

$$Q^F(\boldsymbol{\theta}) = \sum_i \sum_t y_{it}\log\Lambda(\eta_{it}) + (1-y_{it})\log(1-\Lambda(\eta_{it})) + \frac{1}{2}\ln\left[\det(\Lambda(\eta_{it})(1-\Lambda(\eta_{it}))(\partial\eta_{it}/\partial\boldsymbol{\theta})(\partial\eta_{it}/\partial\boldsymbol{\theta})')\right], \tag{16}$$

where $\Lambda(\eta_{ij}) = \exp(\eta_{ij})/(1+\exp(\eta_{ij}))$ is the logistic cdf.

For the HS estimator,

$$Q^{HS}(\boldsymbol{\theta}) = \sum_i \sum_t \log L_{it}(\boldsymbol{\theta}) + \pi^{HS}(\boldsymbol{\alpha}),$$

the penalty term is given by

$$\pi^{HS}(\boldsymbol{\alpha}) = \sum_i -\frac{1}{2}\frac{\sum_t v_{it}^2}{\sum_t -v_{it}^{\alpha}} + \frac{1}{2}.$$

The IE version of this estimator is obtained by replacing v_{it} and v_{it}^{α} by $E(v_{it})$ and $E(v_{it}^{\alpha})$. For the logit model, analytical expressions for $E(v_{it})$ and $E(v_{it}^{\alpha})$ are available, and one obtains

$$Q^{IE} = \sum_i \sum_t y_{it}\log\Lambda(\eta_{it}) + (1-y_{it})\log(1-\Lambda(\eta_{it})) + \sum_i \frac{1}{2}\ln\left[\det\left(\sum_t \Lambda(\eta_{it})(1-\Lambda(\eta_{it}))\right)\right] + \frac{N}{2},$$

see also Bester and Hansen (2009, p.138).

For $\boldsymbol{\beta} = 0$, it is immediately evident that $Q^{IE} = Q^F$. For the general case with covariates, the two estimators would be equivalent for a penalty function applied $\pi^F(\boldsymbol{\theta})$ only to $\boldsymbol{\alpha}$, $\pi^F(\boldsymbol{\alpha})$; i.e., by replacing $(\partial\eta_{it}/\partial\boldsymbol{\theta})(\partial\eta_{it}/\partial\boldsymbol{\theta})'$ in (16) by $(\partial\eta_{it}/\partial\boldsymbol{\alpha})(\partial\eta_{it}/\partial\boldsymbol{\alpha})'$. Since the common parameter vector $\boldsymbol{\beta}$ uses information from the whole sample NT, the penalisation should be mild when the size of the cross-sectional dimension N is large, so that $Q^{IE} \approx Q^F$.

E Additional simulation results

Table E1: MC Simulation: Predicted probabilities, averaged over distribution of α_i (500 replications)

	$T = 8$				$T = 12$			
	Mean	D_1	D_5	D_9	Mean	D_1	D_5	D_9
$\alpha_i \sim Bernoulli$								
True	*0.494*	*0.227*	*0.495*	*0.760*	*0.494*	*0.227*	*0.495*	*0.760*
ML	1.014	0.985	1.015	1.021	1.002	0.997	1.003	1.004
HS	1.013	1.041	1.013	1.005	1.003	1.041	1.002	0.992
BR	1.000	1.078	0.999	0.978	0.998	1.054	0.998	0.983
$\alpha_i \sim Uniform$								
True	*0.486*	*0.235*	*0.485*	*0.741*	*0.487*	*0.235*	*0.485*	*0.741*
ML	1.003	0.940	1.003	1.024	0.998	0.979	0.997	1.004
HS	1.004	0.988	1.004	1.009	0.999	1.015	0.999	0.994
BR	1.000	1.061	1.001	0.981	0.998	1.042	0.999	0.985
$\alpha_i \sim Beta$								
True	*0.511*	*0.231*	*0.512*	*0.788*	*0.511*	*0.231*	*0.512*	*0.788*
ML	0.991	0.962	0.990	1.001	0.997	0.994	0.997	0.999
HS	0.989	1.019	0.988	0.983	0.996	1.039	0.995	0.984
BR	0.997	1.082	0.996	0.973	0.998	1.060	0.997	0.980
$\alpha_i \sim Normal$								
True	*0.516*	*0.273*	*0.518*	*0.754*	*0.516*	*0.273*	*0.518*	*0.754*
ML	0.987	0.881	0.986	1.027	1.001	0.955	1.001	1.018
HS	0.985	0.918	0.983	1.012	0.999	0.981	0.998	1.007
BR	0.995	1.037	0.993	0.981	0.996	1.025	0.994	0.986

Notes: Entries in rows "True" are mean predicted probabilities. Entries in other rows are mean predicted probabilities divided by the value in the corresponding "True" row. Entries in Columns "Mean" are mean predicted probabilities marginal of x. Entries in Columns "D_1", "D_5" and 'D_9" are mean predicted probabilities evaluated at the first, fifth (median) and ninth decile of x.

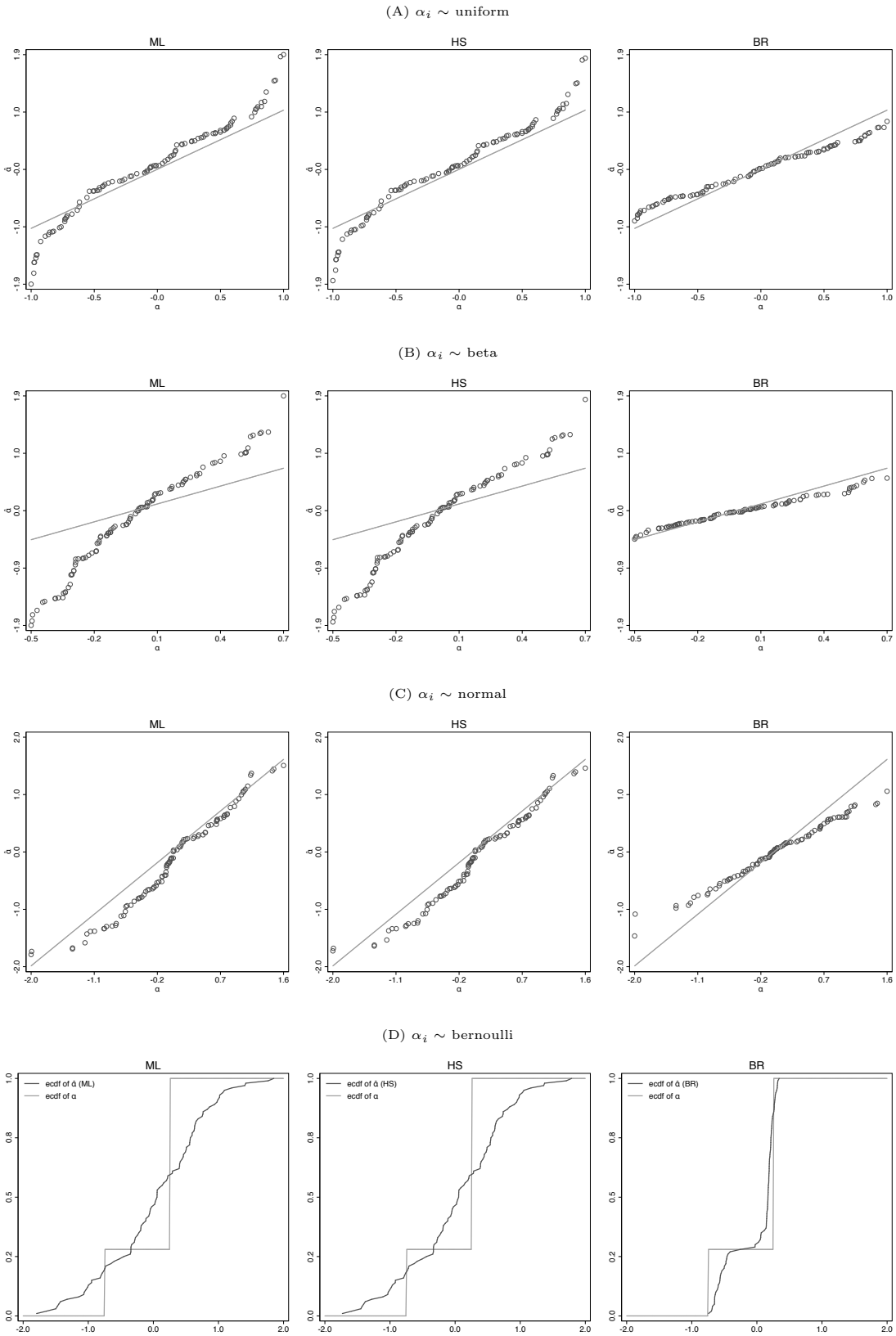


Figure E2: Appendix: Estimated versus true distributions of α_i, N=100, T=2

Notes: Graphs in panels (A), (B), (C) show average estimates of $\alpha_1, \ldots, \alpha_{100}$ over 500 replications against their true values. Graphs in panel (D) show the empirical cdf of the hundred true α_i against the empirical cdf of the hundred average $\hat{\alpha}_i$ estimated over 500 replications.

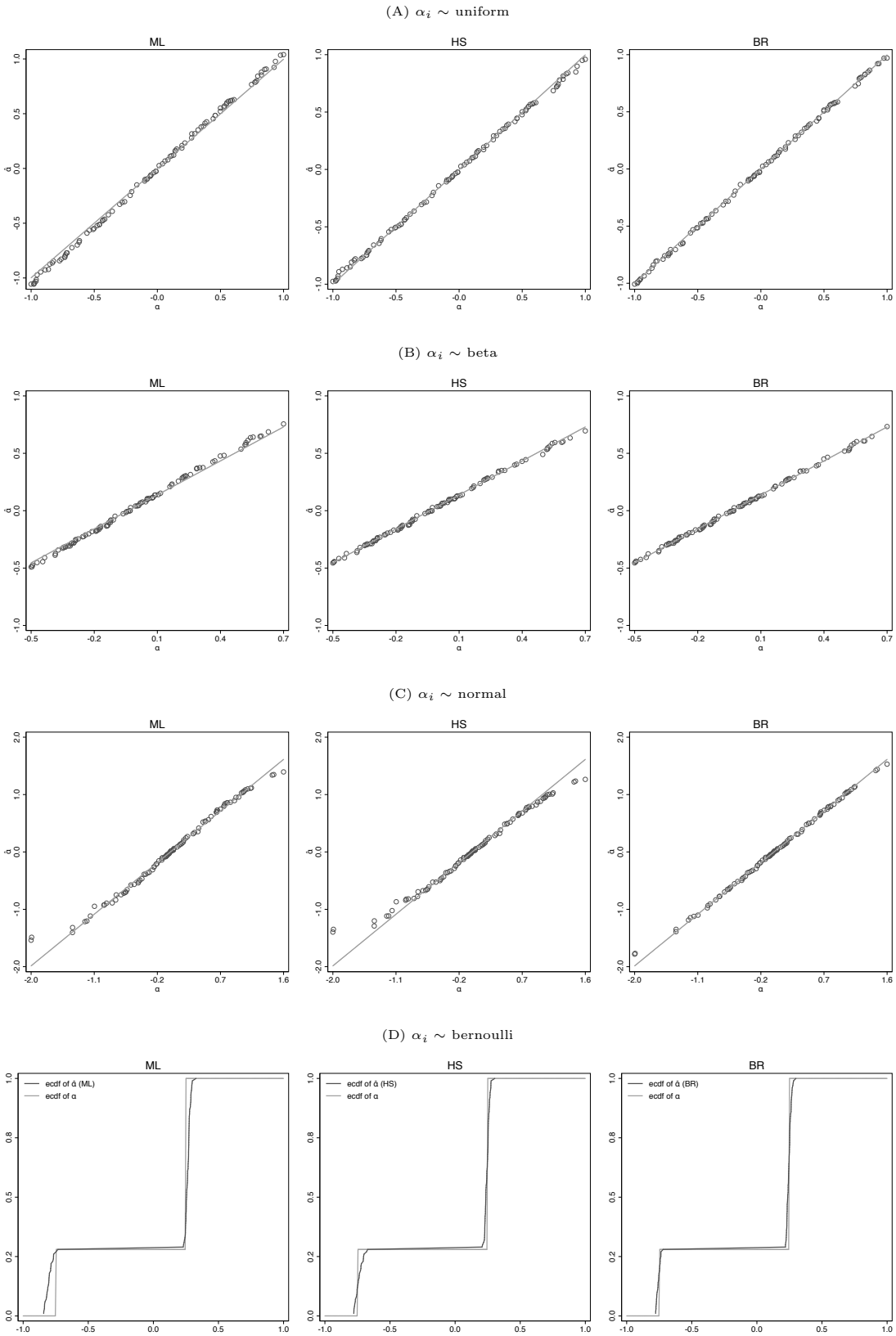


Figure E3: Appendix: Estimated versus true distributions of α_i, N=100, T=12

Notes: Graphs in panels (A), (B), (C) show average estimates of $\alpha_1, \ldots, \alpha_{100}$ over 500 replications against their true values. Graphs in panel (D) show the empirical cdf of the hundred true α_i against the empirical cdf of the hundred average $\hat{\alpha}_i$ estimated over 500 replications.

F Additional information about Section 4, Application to hospital readmissions

F.1 Data sources

FIPS crosswalk

We start by performing minor corrections on the file `CBSAtoCountycrosswalk_FY13.xls`[11] to make the crosswalk between county and FIPS State county linkable to the hospital compare data (i.e. SAINT CLAIR we set equal to ST. CLAIR). Note, that island states such as AMERICAN SAMOA are dropped because we could not merge them to county or HRR information.

County information

We compiled for the years 2011-2015:
`Rural_Atlas_Update14/Jobs.csv` and `Rural_Atlas_Update14/People.csv` files[12] from which we get the variables: yearly unemployment rate, and yearly total population/100,000.

Next, we use the file `SAIPESNC_05APR17_15_02_58_98.csv`,[13] which provides yearly measures of all ages in poverty (in percent) and the median household income (in dollars/10,000). We then merge them via the FIPS crosswalk, all hospitals which could not be merged are included in the regressions with a missing indicator for county.

Hospital Referral Region information

We use zip code crosswalks:[14] `ZipHsaHrr10.xls-ZipHsaHrr14.xls` from the *Dartmouth Atlas*, which allows us to connect the Zip codes to HRRs. We use the one year lagged values as hospital data is published with a lag. We calculate the number of hospitals for each year and define two indicators, one if there are more hospitals (in HRR) than in the previous year, and one if there where less. Note, that we can not distinguish, whether these are actually openings/closings of hospitals or a result of mergers or separations.

We use the number of Discharges for Ambulatory Care Sensitive Conditions from the selected medical discharge rates files:[15] `2010_med_discharges_hrr.xls-2014_med_discharges_hrr.xls` where we subtract the conditions that are equal to our outcome measures (BacterialPneumoniaDischargesp and CongestiveHeartFailureDischar) form the total discharges (DischargesforAmbulatoryCareS). We then merge them via the zip code crosswalk, all hospitals which could not be merged are included in the regressions with a missing indicator for HRR.

Hospital Compare data

Our main data set is provided by the Centers for Medicare & Medicaid Services.

Acute Inpatient PPS:[16]

[11] downloaded from http://www.nber.org/ssa-fips-state-county-crosswalk/ (accessed 26.03.17).
[12] downloaded from https://www.ers.usda.gov/data-products/atlas-of-rural-and-small-town-america/download-the-data/ (accessed 26.03.17).
[13] downloaded from https://www.census.gov/data-tools/demo/saipe/saipe.html (accessed 26.03.17).
[14] downloaded from http://www.dartmouthatlas.org/tools/downloads.aspx?tab=39 (accessed 26.03.17).
[15] downloaded from http://www.dartmouthatlas.org/tools/downloads.aspx?tab=41 (accessed 26.03.17).
[16] downloaded from https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/index.html (accessed 26.03.17).

- FY 2012 Final Rule- IPPS Impact File PUF-August 15, 2011_1.txt
- FY 2013 Final Rule CN - IPPS Impact File PUF-March 2013.txt
- FY 2014 Final Rule IPPS Impact PUF-CN1-IFC-Jan 2014.txt
- FY 2015 IPPS Final Rule Impact PUF-(CN data).txt
- FY 2016 Correction Notice Impact PUF - (CN data).txt

The construction of these variables is taken from Gu et al. (2014). First, we use the information on the number of hospital beds which we include as 2 indicators for 100-399 beds and for more than 400. Second, we resident to bed or daily ratio (rday) is larger than 0.25 as major teaching hospitals and lower than 0.25 but larger than 0 as minor teaching hospitals. Also urban if either urgeo or urspa indicate an urban area. These covariates are almost always constant within hospital, for the very few minor changes we set the to the maximum observed state, to make them time-consistent.

Next, we use Hospital Compare data archive:[17]

- HOSArchive_Revised_Flatfiles_20121001/Hospital_Data.csv and READMISSION REDUCTION.csv
- HOSArchive_Revised_Flatfiles_20131001/Hospital_Data.csv and READMISSION REDUCTION.csv
- HOSArchive_Revised_Flatfiles_20141218/Hospital General Information.csv and READMISSION REDUCTION.csv
- HOSArchive_Revised_FlatFiles_20151210/Hospital General Information.csv and READMISSION REDUCTION.csv
- Hospital_Revised_Flatfiles/Hospital General Information.csv and READMISSION REDUCTION.csv

from which we get for each health condition READM-30-AMI-HRRP, READM-30-HF-HRRP, READM-30-PN-HRRP the excess readmission ratio, which we define as a penalty if larger than 1, we drop the hospitals with missing information in this (our key) variable. We use for each condition its corresponding number of discharges (in 1'000). Note, that missing in this variable correspond to too few discharges, which we use as explanatory variable. Hence, we set missing values to 0 and included with a missing indicator to measure the impact of *too few discharges*. Further, across the three conditions we calculate the total number of discharges (in 1'000) leaving-out the current condition's discharges. Finally, the hospital's ownership is defined for-profit, if neither governmental nor non-profit (as above very minor changes, which we made time-consistent by taking the maximum observed value).

[17] downloaded from https://data.medicare.gov/data/archives/hospital-compare (accessed 26.03.17).

F.2 Additional results

Table F1: Descriptive Statistics, by penalty status and condition

	AMI-fine			*Heart Failure*-fine			*Pneumonia*-fine		
	Never	Some	Always	Never	Some	Always	Never	Some	Always
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Number of condition-specific discharges per 1'000	0.23	0.21	0.22	0.41	0.34	0.42	0.33	0.31	0.39
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)
Total number of discharges other conditions per 1'000	0.84	0.89	0.93	0.59	0.45	0.50	0.51	0.49	0.65
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
For-profit hospital (Yes/No)	0.15	0.19	0.21	0.14	0.20	0.24	0.14	0.21	0.22
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
Number of beds, between 100 and 400 (Yes/No)	0.66	0.69	0.67	0.54	0.54	0.59	0.47	0.56	0.60
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of beds, more than 400 (Yes/No)	0.16	0.18	0.23	0.17	0.12	0.17	0.12	0.12	0.21
	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
Minor teaching hospital (Yes/No)	0.30	0.24	0.24	0.30	0.18	0.16	0.23	0.18	0.20
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
Major teaching hospital (Yes/No)	0.10	0.17	0.29	0.10	0.13	0.23	0.06	0.14	0.24
	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)
Located in urban area (Yes/No)	0.80	0.84	0.87	0.78	0.71	0.77	0.68	0.74	0.79
	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)
Discharges ACSCs per 1'000 enrollees, in HRR	23.03	26.08	28.17	23.49	26.35	29.06	23.82	26.18	28.87
	(0.27)	(0.19)	(0.32)	(0.23)	(0.17)	(0.28)	(0.24)	(0.17)	(0.29)
Hospital entry in HRR (Yes/No)	0.17	0.15	0.11	0.16	0.15	0.12	0.18	0.14	0.11
	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
Hospital exit in HRR (Yes/No)	0.16	0.18	0.20	0.16	0.18	0.21	0.17	0.18	0.21
	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
Household median income in 10'000$, in county	5.27	5.29	5.35	5.19	5.14	5.04	5.13	5.15	5.00
	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.03)
Percent unemployed, in county	6.79	7.27	7.75	6.79	7.39	7.96	6.84	7.36	8.03
	(0.05)	(0.03)	(0.05)	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)
Percent of total population living in poverty, in county	15.33	16.02	16.71	15.60	16.51	17.80	15.70	16.56	17.93
	(0.10)	(0.07)	(0.12)	(0.08)	(0.06)	(0.11)	(0.09)	(0.06)	(0.11)
Total population in 100'000, in county	7.57	9.14	12.89	6.29	8.38	10.51	5.96	8.43	10.47
	(0.32)	(0.21)	(0.44)	(0.22)	(0.20)	(0.33)	(0.24)	(0.19)	(0.34)
Observations	2,459	6,071	2,350	3,549	7,993	3,347	3,651	8,365	3,086
Number of hospitals	531	1'256	514	741	1'633	701	756	1'713	643
Share in %	22.6	55.8	21.6	23.8	53.8	22.5	24.2	55.4	20.4

Notes: Means and standard deviations by condition and fine status. Total number of discharges, if there were to few too report we set the number to 0, in all regressions we include an indicator for this bottom-coding "too few condition-specific cases to report discharges". We follow the same procedure for the total number of discharges in other conditions (we do not present the indicators here). Teaching intensity via the indirect medical education adjustment which measures how many residents are employed at the hospital relative to either the number of beds or to average daily census, which measures the occupancy rate rather than beds. When either ratio was larger than 0 the hospital is classified as a minor teaching hospital and if either one being larger than 0.25 as a major teaching hospital. Dischargees in ambulatory care sensitive conditions, exclude pneumonia and heart failure related causes and is lagged by one year to avoid endogeneity. Hospital entry/exit was assessed by whether the number of hospitals in an HRR went up or down relative to the previous year.
Source: Hospital Compare Dataset and Final Rule Impact files 20012-2016, ACS, Dartmouth Atlas of Health Care, own calculations.