Partial Effects for Binary Outcome Models with Unobserved Heterogeneity

Lucas Núñez *

February 19, 2021

Abstract

Unobserved heterogeneity is ubiquitous in empirical research. In this paper, I propose a method for estimating binary outcome models with panel data in the presence of unobserved heterogeneity, called the *Penalized Flexible Correlated Random Effects* (PF-CRE) estimator. I show that this estimator produces consistent and efficient estimates of the model parameters. PF-CRE also provides consistent estimates of partial effects, which cannot be calculated with existing consistent estimators. Using Monte Carlo simulations, I show that PF-CRE performs well in finite samples. I also illustrate the performance of PF-CRE in two real-data applications: a small T study on party contacts and tactical voting during the 2015 U.K. General Election; and a large T panel dealing with the effect of economic sanctions on government stability. In both cases I find that PF-CRE is a valid approach that reduces bias and/or generates efficiency gains in the estimation.

Keywords: Unobserved heterogeneity; binary outcome; logit; fixed effects

*Assistant Professor, Schar School of Policy and Government, George Mason University. Email: lnunez6@gmu.edu

1 Introduction

The presence of unobserved heterogeneity is ubiquitous in observational studies in political science, and the social sciences in general. It is generally defined as differences across units of analysis that are not measured, influence the outcome, and may correlate with observed characteristics of interest. Regardless of its origins and form, unobserved heterogeneity poses the same problem: ignoring it when it is correlated with the covariates of interest leads to biased and inconsistent estimates of the quantities of interest.

There are three main types of estimation approaches for binary outcome models with panel data in the presence of unobserved heterogeneity: treat the heterogeneity as parameters to be estimated; use conditional maximum likelihood estimation (Rasch, 1961; Chamberlain, 1980) and related semi-parametric techniques (e.g., Abrevaya, 2000); or use random or correlated random effects (Mundlak, 1978; Chamberlain, 1980).[1] Each of these approaches suffers from one of three problems. They produce inconsistent and biased estimates, cannot produce estimates of the probability of the outcome nor partial effects of the covariates of interest, or they require making restrictive assumptions about how the unobserved heterogeneity relates to the observed covariates in the model.[2] A fourth problem applies to the first two approaches and is particularly pervasive in rare-events data: accounting for unobserved heterogeneity can absorb most of the cross-sectional variation in the data leading to very large uncertainty about the quantities of interest and consequently inconclusive results (see, for example Beck and Katz, 2001).

In this paper I develop an estimator that deals with unobserved heterogeneity in binary outcome models, the *Penalized Flexible Correlated Random Effects* (PF-CRE) estimator.[3] In the PF-CRE estimator, I explicitly account for the correlation between the observed and unobserved components of the model, using a large flexible specification (more details below). Moreover, I include a penalization step for variable selection to induce efficiency. This estimator addresses the four problems described above: it provides consistent estimates for the model parameters, allows for the estimation of partial

[1] Each of these approaches has received attention in the political science literature, in the past as well as recently. See for example Beck and Katz (2001); Coupe (2005); Greene (2004); Beck (2018, 2020); Cook et al. (2020) on FE; Beck (2018, 2020) on CMLE; and Bell and Jones (2015); Clark and Linzer (2015) for the linear CRE and Crisman Cox (2019) for the binary case. Further discussion of this literature is included in Section 3

[2] Making restrictive assumptions about the individual heterogeneity also leads to biased estimates if those assumptions do not hold. I distinguish the bias and inconsistency that arise from unrealistic assumptions from the one that arises from the estimation procedure itself.

[3] I have also developed an `R` package called `PFCRE` that implements this estimator.

effects, makes mild assumptions about the unobserved heterogeneity, and does not discard all cross-sectional variation in the data.

The PF-CRE estimator builds upon the correlated random effects (CRE) approach by using a rich and *flexible* specification of the correlation between the unobserved heterogeneity and the observed covariates in the model derived from a higher level assumption called exchangeability. This flexible specification is composed of functions of the observed covariates (such as individual time-means and other exchangeable functions[4]), additional observed time-invariant characteristics, and higher order interactions between these terms. The flexible specification in PF-CRE requires making weaker assumptions about the unobserved heterogeneity than in the traditional CRE approach. Weaker assumptions mean that PF-CRE is more likely to capture the underlying heterogeneity correctly and lead to correct inferences.

The key challenge of the specification in PF-CRE is that it requires the estimation of additional parameters, which can increase uncertainty. To address this dimensionality issue, I estimate the model via *penalized* Maximum Likelihood using the Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan and Li, 2001). When the penalized likelihood is maximized, the polynomial coefficients with little or no predictive power are shrunk to zero, a form of variable selection. In the case of PF-CRE, the penalization selects the polynomial terms that are necessary to control for the unobserved heterogeneity and discards the rest. Since the main covariates of interest are not penalized in PF-CRE, no shrinkage is introduced to those parameters directly. The reduction of dimensionality is especially useful in small samples, as it can significantly reduce the variance of the estimates, leading to more accurate inferences. The estimator is implemented with an approximate algorithm that shows good performance in simulation studies.

The assumptions underlying the PF-CRE estimator may not always be sufficient to capture the unobserved heterogeneity in the data. The underlying heterogeneity may be correlated with the observed covariates in a highly convoluted way that PF-CRE may fail to successfully approximate. Thus, for the logistic case, I present a model specification test to determine whether the PF-CRE approach is appropriate for the data at hand. This provides an indirect test of the assumptions in PF-CRE and a tool to help researchers decide when it is correct to use it.

[4]Exchangeable functions are those for which the order of their arguments does not change their value. For example, moments are exchangeable functions: an average does not change if the order in which the terms enter the sum is altered.

I study the small sample performance of the PF-CRE estimator using Monte Carlo simulations. The simulations show that the asymptotic properties of PF-CRE hold in small samples despite the approximate nature of the algorithm, and that it performs better than alternative estimators. For the logistic case, the simulations show that the rejection rate of the specification test is close to theoretical levels.

Additionally, I illustrate the performance of PF-CRE in two applications: a small T panel on tactical voting in the United Kingdom; and a large T panel that replicates Marinov (2005)'s study of economic sanctions. In both cases, the specification test suggests PF-CRE's assumptions hold and both highlight the advantages of the method.

2 Penalized Flexible Correlated Random Effects

A binary outcome model with unobserved heterogeneity consists of a binary reponse, y_{it}, and a k-dimensional vector of time-varying characteristics, x_{it}, such that the response for individual i at time t is generated by:

$$y_{it} = \mathbb{I}[\alpha + x_{it}\beta + c_i - \varepsilon_{it} > 0], \quad i = 1, ..., n, \ t = 1, ..., T, \tag{1}$$

where $\mathbb{I}[A]$ is an indicator function that takes the value of one if A holds and zero otherwise; α is a constant; β is a k-dimensional parameter vector; c_i is the unobserved heterogeneity that is constant over time; and ε_{it} is an individual- and time-specific error.[5]

When the error terms are independently and identically distributed according to a known cumulative distribution $G(\cdot)$, equation 1 can be alternatively written as:

$$Prob(y_{it} = 1|x_{it}, c_i) = G(\alpha + x_{it}\beta + c_i). \tag{2}$$

Typical choices of $G(\cdot)$ are the normal distribution, which gives the probit model, or the logistic distribution, which gives the logit model.

While model parameters β may be of interest in themselves, researchers are usually interested in estimating partial effects and probabilities. In the presence of unobserved heterogeneity these partial effects are calculated by taking expectations over c.[6] The partial effects for the model in equation 2 are defined by:

[5] The focus on a balanced panel is for simplicity; however, T can differ across individuals.

[6] Alternatively, one can calculate effects for particular values of c. However, I prefer not to take this approach, as it presumes knowledge about which values of c are interesting, even though it is an unobserved quantity.

$$PE_j(x) = E\left[\frac{\partial}{\partial x_j} G(\alpha + x\beta + c)\middle| x\right], \quad j = 1, ..., k \tag{3}$$

where x_j denotes that jth element of x. Additionally, researchers may be interested in the average partial effect, defined as:

$$APE_j = E\left[\frac{\partial}{\partial x_j} G(\alpha + x\beta + c)\right], \quad j = 1, ..., k \tag{4}$$

where the last expectation is taken with respect to both x and c.[7]

2.1 Assumptions for Identification and Estimation

The PF-CRE method focuses on a particular challenge in panel data: the presence of time-invariant unobserved heterogeneity. This setting is represented in Figure 1 where the identification challenge lies in c_i being unobserved, influencing the outcome, *and* being correlated with the time-varying covariates x_{it}.[8] Thus, this setting does not encompass other potentially interesting relationships in panel data like dynamic effects, which have also received attention recently (e.g., Blackwell and Glynn, 2018). The assumption underpinning all models like the one represented in Figure 1 is that the observed covariates are strictly exogenous conditional on the unobserved heterogeneity:

$$h(y_{it}|x_{i1}, ..., x_{iT}, c_i) = h(y_{it}|x_{it}, c_i), \quad \forall\, t = 1, ..., T,$$

where h is a density function. This exogeneity assumption, common to all common approaches to unobserved heterogeneity in binary outcome models, rules out models with lagged dependent variables as well as other models with dynamic effects (see, for example Wooldridge, 2010, p610–611).

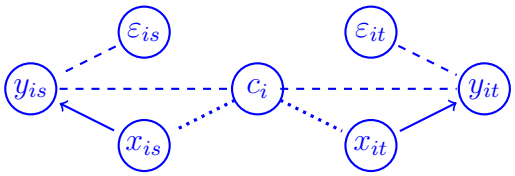
While the model parameters (and probabilities) can be non-parametrically identified (with some mild assumptions), consistent estimation is typically not possible without further restrictions, particularly when T is small and the incidental parameters problem is at its greatest.

A common approach is to restrict the distribution of c_i conditional on $(x_{i1}, ..., x_{iT})$. Here, rather than directly imposing ad hoc assumptions about this distribution, I begin with a higher-level assumption about the relationship between the unobserved heterogeneity and the observed covariates.

[7] Note that some authors refer to equation 3 as the *average partial effect*, as it is averaging over the distribution of the unobserved heterogeneity. However, researchers also use the term average partial effect for equation 4. I reserve the term average partial effect for equation 4.

[8] When c is independent of x, it is sometimes known as a random effect. This case does not pose significant challenges to traditional estimators. The PF-CRE approach is also valid for this case. However, if it is known that the unobserved heterogeneity is fact independent of x, a random effects estimator will be preferable.

Figure 1: *Relationships between the variables*



Solid lines represent the effects of interest. The subindex i refers to a unit of analysis, whereas s and t refer to distinct time periods. Other periods are excluded to simplify exposition

Assumption 1 (Exchangeability)

$$f(c_i|x_{i1},...,x_{iT}) = f(c_i|x_{is_1},...,x_{is_T}),\ where\ s_j \in \{1,...,T\},\ s_j \neq s_{j'}.$$

Assumption 1 requires that the distribution of the unobserved heterogeneity conditional on the observed covariates, $f(c_i|x_{i1},...,x_{iT})$, does not depend on the order in which x_{it} enters the density $f(c_i|\cdot)$.

Under Assumption 1, without loss of generality, $f(c_i|x_{i1},...,x_{iT})$ can be written as a polynomial on $z_i^1,...,z_i^T$, where $z_i^t = \sum_{s=1}^{T}(x_{is})^t$ (Altonji and Matzkin, 2005, and references therein for further details).[9] Note that when divided by T, $(z_i^1,...,z_i^T)$ are in fact the first T non-central moments of $(x_{i1},...,x_{iT})$ for each i.[10]

In most circumstances, researchers also observe time-invariant information, w_i, about each individual i, such as birth gender, race, and year of birth. These time-invariant characteristics can be added to the conditional distribution of c_i to improve fit. Moreover, the inclusion of these auxiliary variables can help the exchangeability assumption hold.

The exogeneity assumption together with exchangeability are sufficient to non-parametrically identify the model parameters and probabilities in theory (Altonji and Matzkin, 2005). However, assumption 1 alone is not sufficient in practice. The reason is that the first T non-central moments characterize the T observations per individual i, thus exhausting the degrees of freedom.[11] Therefore, additional restrictions are necessary for identification and practical estimation:

[9]The Weierstrass approximation theorem establishes that a function with bounded support can be uniformly approximated by a polynomial function. Because of exchangeability, this is a symmetric polynomial. By the fundamental theorem of symmetric polynomials, it may be written as a polynomial in the power functions (i.e., the moments). See Altonji and Matzkin (2005, p. 1062). Other polynomial bases can be used. I use the power functions because they have a more intuitive interpretation.

[10]Polynomial bases other than the moments (power functions) could be used to the same effect as they ultimately create the same conditioning set (see Altonji and Matzkin, 2005). Using moments, however, creates a direct comparison to CRE.

[11]The exact same problem arises if other polynomial bases are used.

Assumption 2 (Linear Index) *The conditional density function* $f(c_i|z_i^1, ..., z_i^T, w_i)$ *depends on a linear index of* $(z_i^1, ..., z_i^T, w_i)$ *and interaction terms, for some* $\tau < T$*. That is:*

$$f(c_i|z_i^1, ..., z_i^T, w_i) = f(c_i|z_i\gamma),$$

where z_i *is the vector of the first* τ *moments, the observed time-invariant characteristics,* w_i*, and interaction terms.*

Under Assumption 2, I restrict attention to a linear index of the first τ moments of $(x_{i1}, ..., x_{iT})$, observed time-invariant characteristics, and interaction terms. Notice that this is simply a truncation of the polynomial derived from exchangeability. This implies a stronger condition than exchangeability alone, but it maintains sufficient flexibility to capture (or approximate) the conditional distribution of the unobserved heterogeneity.

With Assumptions 1 and 2 (together with exogeneity) it is possible to consistently estimate model parameters semi-parametrically up-to-scale by relying, for example, on the Maximum Score estimator (Manski, 1987). But estimation of probabilities is not directly possible.[12] Full parametric estimation, then, requires specifying a distributional form.

Assumption 3 (Normality) $f(c_i|\cdot)$ *is a normal density function with variance* σ^2*.*

Here I use a normal distribution, as its use is widespread in correlated random effects models; in fact, it is the only distribution used, to the best of my knowledge.[13] However, other distributions are possible, as long as they have finite moments.[14] Combining assumptions 1, 2, and 3, the unobserved heterogeneity and its density function can be written as:

$$c_i = z_i\gamma + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma^2), \quad f(c_i|x_{i1}, ..., x_{iT}) = \mathcal{N}(z_i\gamma, \sigma^2). \tag{5}$$

2.2 Estimation

Imposing Assumptions 1, 2, and 3 to the model in equation 2 results in the following specification:

$$Prob(y_{it} = 1|x_{it}, c_i) = G(\alpha + x_{it}\beta + z_i\gamma + \eta_i), \text{ with } \eta_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \tag{6}$$

[12] One alternative is to combine parameter estimates from the Maximum Score estimator with a kernel method to estimate probabilities. While appealing because it does not introduce distributional assumptions, the combination of a semiparametric with a non-parametric method makes convergence slow and uncertainty very large, possibly rendering estimates useless unless a very large dataset is available.

[13] In fact, Beck and Katz (2007) show that random effects models perform well, even when the normality assumption is violated.

[14] Finite moments are required because expectations are not well defined otherwise.

where z_i is a vector of moments of $(x_{i1}, ..., x_{iT})$, observed time-invariant characteristics, and interaction terms among these (the polynomial); and η_i is a normally distributed random effect with variance σ^2 that is independent of the covariates of the model.[15]

In principle, the parameters β in equation 6 can be estimated via Maximum Likelihood. The log-Likelihood function for this model is:

$$\log L(\beta, \alpha, \gamma, \sigma) = \sum_{t=1}^{T} \sum_{i=1}^{n} [y_{it} \log(p_{it}) + (1 - y_{it}) \log(1 - p_{it})] \tag{7}$$

$$p_{it} \equiv Prob(y_{it} = 1|x_{it}) = \int_{-\infty}^{\infty} G(\alpha + x_{it}\beta + z_i\gamma + \eta_i) \frac{1}{\sigma} \phi(\eta_i/\sigma) d\eta_i, \tag{8}$$

where $\phi(\cdot)$ is the standard normal density function.

The model in equation 6 represents a *flexible* specification of a Correlated Random Effects (CRE) model. It is a CRE-type model because it assumes a specific correlation form between the unobserved heterogeneity and the covariates in the model (represented by $z_i\gamma$). It is flexible because, under Assumptions 1 and 2, it can accommodate a wide range of correlation forms.

The flexible specification derived from Assumptions 1 and 2 requires the estimation of additional coefficients (γ). When the number of covariates is small, γ is relatively low dimensional. But the dimensionality of γ increases exponentially with the number of covariates in the model. However, the assumptions establish that the polynomial $z_i\gamma$ is sufficient to capture the unobserved heterogeneity, but do not require that all its terms are necessary for this. That is, the underlying unobserved heterogeneity may have a simpler form that relies only on some of the terms of the polynomial. For this reason, detecting unnecessary terms in the polynomial and removing them can produce more efficient estimates of the parameters of interest by simplifying the final specification.

To address the dimensionality issue introduced by the flexible specification, I use a *penalized* Maximum Likelihood estimation technique. This technique performs variable selection in an efficient way that avoids computing an infeasible number of models to choose the one with the better fit. I estimate β using *Penalized Flexible Correlated Random Effects* (PF-CRE), which is defined by:

$$(\widehat{\beta}, \widehat{\alpha}, \widehat{\gamma}, \widehat{\sigma}) = \arg \max_{(\beta,\alpha,\gamma,\sigma)} \log L(\beta, \alpha, \gamma, \sigma) - \Pi_\lambda(\gamma), \tag{9}$$

where $\Pi_\lambda(\cdot)$ is a penalty function that penalizes only the terms used to model the unobserved heterogeneity (γ), but not the parameters associated with the observed covariates (β). I use the Smoothly

[15] Independence follows from Assumptions 1 and 2, and normality from Assumption 3.

Clipped Absolute Deviation (SCAD) penalty (Fan and Li, 2001), defined as:

$$\Pi_\lambda(\gamma) = \begin{cases} \lambda|\gamma| & \text{if } |\gamma| \leq \lambda, \\ -\frac{|\gamma|^2 - 2a\lambda|\gamma| + \lambda^2}{2(a-1)} & \text{if } \lambda < |\gamma| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\gamma| > a\lambda, \end{cases} \tag{10}$$

where a and λ are constants that govern the penalization. The SCAD penalty shrinks small values of γ towards zero, while leaving larger values of γ mostly unpenalized. This way, SCAD selects those terms in z_i that are most predictive of the outcome and discards those that are not. Importantly, the shrinkage introduced by the SCAD penalty in PF-CRE does not affect the coefficients of interest, β, directly since they are left unpenalized.[16] I use the SCAD penalty because it has the Oracle property for this problem. The Oracle property establishes that the penalized estimation selects the correct set of non-zero polynomial terms and that, in pointwise convergence, the asymptotic distribution of the estimates is the same as the one obtained by estimation with the non-penalized likelihood using only the correct (but unknown) set of terms. That is, it establishes that, asymptotically, there is no efficiency cost to variable selection.[17]

2.3 Asymptotic Properties

Here I discuss the asymptotics of the PF-CRE estimator. I leave issues arising from the computational implementation to the next subsection. The PF-CRE estimator with the SCAD penalty produces consistent, efficient, and asymptotically normal estimates of the model parameters, β. I state this result in the following Theorem 1 for easy reference:

Theorem 1 *Under Assumptions 1, 2, and 3,*

$$\sqrt{nT}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, V(\beta)), \tag{11}$$

where $V(\beta)$ is the $k \times k$ submatrix for β from the inverse of the Fisher Information matrix of the full model (that depends on all parameters, $\alpha, \beta, \gamma, \sigma$).

Theorem 1 follows from standard properties of Maximum Likelihood estimation and the Oracle property of the SCAD penalty. First, the consistency of the estimator stems from the combination of standard properties of Maximum Likelihood estimators together with the *flexible* specification derived

[16] The parameter a in the SCAD penalty is usually set to $a = 2.3$ (Fan and Li, 2001). The parameter λ can be chosen via cross-validation.

[17] Alternative penalties that have the Oracle property can be used and the asymptotic properties of PF-CRE will be the same. However, I use SCAD because it shows good finite sample performance relative to, for example, adaptive LASSO for this problem.

from the exchangeability and related assumptions. Second, the Oracle property of SCAD establishes that the penalized estimator has the same asymptotic distribution as the underlying (and unknown) data generating process (Ibrahim et al., 2011; Hui et al., 2017). Consequently, it has the same pointwise asymptotic properties of the Maximum Likelihood estimator of the data generating process. Efficiency and normality of the PF-CRE estimator thus follow from the properties of Maximum Likelihood estimators.[18]

It is important to note here that the Oracle property is a pointwise asymptotic convergence result, rather than a uniform convergence result, which is stronger. The finite sample behavior of estimators that converge pointwise is not necessarily well behaved as that of estimators that converge uniformly (Leeb and Pötscher, 2005, 2008). However, poor finite sample behavior is less likely (although possible) in selection based on regular estimators (Leeb and Pötscher, 2005) like PF-CRE. Moreover, the PF-CRE estimator does not penalize the main coefficients of interest themselves. As such, it greatly isolates them from potentially poor finite sample behavior.[19] While this poor behavior can occur in parameters involved in the polynomial that composes the flexible specification, the specific values of these coefficients are not important per-se individually but as a whole that captures the unobserved heterogeneity for a particular unit.[20] Therefore, while the type of convergence of the estimator could produce problematic finite sample estimates, poor behavior is less likely to occur in the specific case of PF-CRE. Additionally, at least for the logistic case, a sanity check is always possible for parameter estimates by comparing them to the estimates from CMLE (see below). A related concern is that estimators that include penalization might shrink the variance of main parameter estimates (and derived quantities) artificially.[21] However, since the penalization applies to the polynomial and not the main coefficients, this is a lesser concern in the practical estimation of PF-CRE in finite samples. Finally, the simulations presented in Section 5 show no manifestations of poor behavior in finite samples, although it is possible it may arise.

The next result establishes that the PF-CRE estimates of partial effects are also consistent:

Corollary 1 *Under Assumptions 1, 2, and 3, the partial effects are identified, and for all x:*

[18]The asymptotic properties of Maximum Likelihood estimation hold under a number of regularity conditions, which the PF-CRE model satisfies.

[19]In fact, Leeb and Pötscher (2008) note that when the parameter space is partitioned issues of maximal risk underpinning poor finite sample behavior do not apply to the non-sparse portion, which would correspond to β in PF-CRE.

[20]This observation is similar to findings in Belloni et al. (2016) who show, for different models, that inference for a set of parameters of interest in the presence of selection for a different set of variables can achieve uniform convergence under weak conditions (see also Belloni et al., 2012, 2014)

[21]See, for example, Knight (2008)

$$\widehat{PE}_j(x) \equiv \int_{-\infty}^{\infty} g(\hat{\alpha} + x\widehat{\beta} + z\widehat{\gamma} + \eta)\frac{1}{\widehat{\sigma}}\phi(\eta/\widehat{\sigma})\widehat{\beta}_j d\eta \xrightarrow{p} PE_j(x), \quad j = 1, ..., k,$$

where $g(\cdot)$ is the probability density function of $G(\cdot)$.

Moreover, it is asymptotically normal and efficient:

$$\sqrt{nT}(\widehat{PE}_j(x) - PE_j(x)) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

The Oracle properties of SCAD guarantee that $z\widehat{\gamma}$ is a consistent estimator of $z\gamma$. Corollary 1 follows from this and Theorem 1 by direct application of the continuous mapping theorem.[22] Standard errors for the partial effects can be obtained via the Delta method or bootstrap.

To estimate the partial effects, it is necessary to specify a value of z. In principle, any value of z is valid for estimating the partial effects. However, a significant proportion (or all) of the terms in z are functions of x. For this reason, it is advisable to ensure that the values of x and z used to calculate the partial effects are consistent with one another to avoid issues similar to those of extreme counterfactuals (King and Zeng, 2006).

The PF-CRE estimator relies on the SCAD penalty to reduce the dimensionality of the polynomial used to capture the unobserved heterogeneity. The asymptotic results rely on knowledge of the optimal penalty parameters λ and a, which in reality are unknown. Fan and Li (2001) note that selection criteria are not very sensitive to values of a and suggest that a choice of $a = 3.7$ results in good practical performance of the penalization. For selection of the λ parameter in SCAD, I rely on a grid search using the Akaike Information Criterion (AIC) to select the optimal value.[23]

2.4 Computational Implementation

The PF-CRE estimator is a form of generalized linear mixed model, with fixed coefficients β and γ and only one random coefficient, the random effect (or random intercept). The computational difficulty in implementing an algorithm for these type of estimators with penalization is the combination of non-convexity in the penalized likelihood together with the integration required for the random coefficient(s). Few alternatives exist that implement penalization in generalized linear mixed models

[22]The continuous mapping theorem states that continuous functions are limit-preserving. Therefore, a continuous function, $G(\cdot)$, of a random vector, $(\widehat{\beta}, \widehat{\alpha}, \widehat{\gamma}, \widehat{\sigma})$, converges in distribution to the function of the random vector.

[23]An alternative criterion can also be used. Additionally, it is possible to rely on cross-validation for the selection of the optimal penalty parameter. However, cross-validation can impose very significant computational costs given the large number of model estimations required. The simulations presented in Section 5 suggest that AIC has a good performance; alternative procedures are unlikely to produce any meaningful gains.

that result in computationally fast and stable algorithms. Ibrahim et al. (2011), for example, produce an algorithm that simultaneously penalizes and estimates fixed and random coefficients. However, the algorithm is computationally slow and can produce unstable numerical results.

Given the complexities of computationally implementing penalization in the presence of random coefficients in non-linear models, I rely on an approximate algorithm with a re-fitting (or hybrid) step that produces faster results that are numerically more stable and remain accurate. As such, the estimation proceeds on an approximation that first estimates the unpenalized fixed parameters (β) and the penalized fixed parameters (γ) using a local quadratic approximation, LQA (Ulbricht, 2012), that abstracts from the random effect (the only random coefficient in PF-CRE). In a second step, only the fixed parameters selected by using LQA are used in a maximum likihood re-fitting of the model that includes the random effect. This approach is somewhat related to the ones used by Schelldorfer et al. (2011) and Groll and Tutz (2014), both emphasizing the gains in numerical stability as well as more accurate estimation of the variance of the random coefficients. Schelldorfer et al. (2011) in particular, notes that the hybrid algorithm performs better in practice, producing results closer to the Oracle estimator. This is mainly due to the numerical stability of the algorithm more than compensating for the approximation.[24]

The particular computational implementation of the model generates some ambiguity about the extent to which the theoretical asymptotic results are achieved, and establishing the statistical properties of the algorithm itself is not straightforward. However, the simulations that compare PF-CRE to the Oracle estimator show that this algorithm is capable of producing estimates that are quite close to those of the Oracle, strongly suggesting that the approximate algorithm is a very good approximation to otherwise computationally challenging calculations.

3 Relation to Existing Estimators

As previously mentioned, there are three main strategies for the estimation of binary outcome models with panel data in the presence of unobserved heterogeneity.[25] I briefly discuss each of them and how

[24]Other researchers have followed related approaches for penalized methods with random effects in different settings. For example, Lai et al. (2012) in a nonparametric setting for additive mixed models; Lin et al. (2013) use a two-step algorithm that also penalizes random effects.

[25]A fourth alternative is the use of the Linear Probability Model, which in some circumstances can produce good estimates of average partial effects but can produce nonsensical estimates of partial effects for many values of the independent variables.

they relate to the PF-CRE estimator.[26]

The first approach is estimation via Fixed Effects (FE), where the c_is are treated as parameters to be estimated. This is operationalized through dummy variables for each individual in the sample. When the panel is short (small T), this requires estimating each dummy with a handful of observations, a problem known as the incidental parameters problem (first noted by Neyman and Scott, 1948). The incidental parameters problem implies that estimates from the FE approach are inconsistent for small T. This bias can be substantial. For example, simulations in Greene (2004) show that with $T = 5$ this bias can be 40% of the true parameter value. The asymptotic bias is of order $O_p(T^{-1})$, meaning that it disappears as T tends to infinity. Monte Carlo evidence in Heckman (1981) suggest that this bias is negligible for a panel of size $T = 8$, although more recent studies in Coupe (2005) suggest that a larger size of $T = 16$ is preferable.[27]

In light of the inconsistency of the FE estimator, bias correction procedures have been proposed.[28] These corrections reduce the bias; however, they do not eliminate it.[29] A related strand of literature seeks to ameliorate the incidental parameters problem (as well as the computational burden of estimating $n + k$ parameters) by assuming that the individual heterogeneity is in fact group heterogeneity. However, these group fixed-effects estimators also suffer from the incidental parameters problem (although to a lesser extent) and may not be appropriate for short panels.[30]

A different, although related, problem with the FE estimator relates to the presence of all-zero units(see King and Zeng, 2001, for example).[31] The FE estimator will effectively remove all zero units. This can lead to a loss of information resulting in higher uncertainty over the estimates. Moreover, the effective loss of these observations and the reliance on a subset of the data may result in a form of sample selection bias, although its consequences cannot really be settled empirically (Beck, 2018, 2020). This issue becomes increasingly relevant with rare-events data, which leads to a large number of all-zero units. Cook et al. (2020) propose an alternative to estimate fixed-effects models in this

[26]See Greene (2015) for a review of the literature on parametric estimation of discrete choice models.

[27]In the case of $T = 2$ Abrevaya (1997), shows that the maximum likelihood estimates of β using the FE approach converge to 2β. Thus, dividing the FE estimate by 2 results in a consistent estimate of β. However, the incidental parameters problem persists in the estimation of partial effects.

[28]See, for example, Fernandez-Val (2009); Fernandez-Val and Vella (2011); Hahn and Newey (2004); Dhaene and Jochmans (2015).

[29]In fact, Dhaene and Jochmans (2015) show that the elimination of the leading term of the bias leads to larger magnitudes of the higher order terms of the bias in the bias-corrected estimator.

[30]See, for example, Bonhomme and Manresa (2015); Ando and Bai (2016); Su et al. (2016). Bonhomme et al. (2017) do not assume group heterogeneity, but assume that the heterogeneity can be coarsened into groups without significant loss.

[31]All-one units create an exact mirror problem.

type of data by relying on a maximum likelihood approach that penalizes the unit dummy variables using Jeffreys prior. While the imposition of Jeffreys prior on these dummy variables is somewhat ad hoc, this method can help resolve these issues. Its performance is also assessed by Crisman Cox (2019), who finds that it is typically outperformed by the traditional CRE method (see below).[32]

A separate issue in FE, that is also shared by CMLE, occurs when independent variables of interest rarely change. The variation in such variables becomes hard to distinguish from unobserved heterogeneity and, therefore, methods that allow for essentially unrestricted time-invariant unobserved heterogeneity will remove most of the variation in the rarely changing covariate thus making it extremely hard to detect any effect coming from it (see, for example Beck and Katz, 2001; Green et al., 2001; King, 2001).[33]

The second approach is estimation via Conditional Maximum Likelihood (CMLE), which results in consistent estimates of β (Rasch, 1961; Andersen, 1970; Chamberlain, 1984). This approach relies on conditioning the estimation only on those individuals with variation in the outcome across time. By restricting the estimation to these individuals, the conditional likelihood only depends on β and not the unobserved heterogeneity c_i, avoiding the incidental parameters problem. However, this property only holds for the logistic distribution.[34]

The CMLE approach has two main shortcomings. First, it does not provide estimates of the partial effects.[35] This is because location parameters, c_i and α, are not estimated. The second shortcoming is inefficiency. The CMLE approach allows the heterogeneity to be completely unrestricted, which implicitly assumes that individuals with no variation in the outcome provide no information about β. However, if the heterogeneity has a less general form, conditioning on these individuals results in a loss of information, and consequently larger standard errors in the estimates.[36] The same issues relating to rare-events as well as rarely changing covariates that apply to FE estimation also apply to

[32]While Cook et al. (2020) shares the idea of penalization with PF-CRE, they penalize different things. Additionally, the penalization in Cook et al. (2020) is imposed directly, rather than being derived from higher level assumptions.

[33]This is an issue addressed by Plümper and Troeger (2007, 2011) and Greene (2011) for the linear regression case, for example.

[34]Chamberlain (2010) shows that if the support of the observed predictor variables is bounded, then identification is only possible in the logistic case. Moreover, if the support is unbounded, the information bound is zero unless the distribution is logistic. This means that consistent estimation at the standard asymptotic rates is only possible in the logistic case. For alternative semi-parametric estimators that require unbounded support and have slower convergence rates, see Manski (1987); Abrevaya (2000).

[35]This is also a problem with semi-parametric alternatives to CMLE.

[36]Note that the FE approach results in the same kind of information loss without 'technically' discarding observations outright. The behavior of individuals with no variation in the outcome is fully explained by the dummy variables corresponding to these individuals. Thus, these individuals do not contribute to the estimation of the model parameters β (see, for example, Beck and Katz, 2001).

CMLE.

The third approach is estimation via Correlated Random Effects (CRE). This approach requires making explicit assumptions about the unobserved heterogeneity. The strongest restriction is assuming that the heterogeneity is independent of the covariates in the model, leading to the Random Effects (RE) model. Mundlak (1978) proposes to model the unobserved heterogeneity as a linear combination of the time-means of the covariates and a random effect, which allows for correlation between the model covariates and the unobserved heterogeneity.[37]

The main advantage of CRE is that, by providing an explicit model of the unobserved heterogeneity, it allows for the estimation of partial effects. However, it does so at the cost of restricting the unobserved heterogeneity with ad-hoc specifications, which depending on the unknown data generating process they can be severe. When this restriction is not satisfied by the data generating process (which is unobserved), CRE models are misspecified and provide incorrect estimates of the model parameters and partial effects.

Implementations of RE and CRE for the linear regression case have received attention form political scientists relatively recently. For example Clark and Linzer (2015) compare RE models with FE models and provide guidance in choosing one over the other, emphasizing that while RE models have more restrictive assumptions they are have lower variances.[38] Bell and Jones (2015) discuss CRE estimation in linear models strongly arguing in their favor relative to FE models, finding them always preferable.[39] While linear models do not suffer from the inconsistency derived from the incidental parameters problem of non-linear models, the efficiency issues discussed in these papers are closely related to the efficiency issues that PF-CRE addresses.

Crisman Cox (2019) discusses the CRE estimator in the binary outcome case, with particular attention to rare-events data. His findings show that CRE models are particularly useful in rare-events models with unobserved heterogeneity. Moreover, he finds that CRE models can perform relatively well even when the unobserved heterogeneity is somewhat misspecified. The gains of CRE

[37]Chamberlain (1980) proposes a more general version of Mundlak's model, modeling the unobserved heterogeneity by projecting the time dimension of the model into one dimension. This is akin to a weighted mean of the covariates across time.

[38]In fact, the authors note that in some cases, a biased estimator (like RE) can be preferable to an unbiased estimator (like FE in linear models) provided it results in a sufficiently large reduction in variance.

[39]It should be noted, however, that part of their strong recommendation may stem for simulations in which the data generating process was particularly beneficial to CRE models. However, the main point remains: CRE models can capture a good amount of unobserved heterogeneity that combined with a smaller variance make them attractive alternatives to fixed-effects models in linear environments.

are most notable relative to the fixed effects estimator, but he also finds that it typically outperformes other alternatives, like Beck's two-step estimator (Beck, 2015) and the penalized maximum likelihood estimator from Cook et al. (2020).

The PF-CRE estimator introduced in this paper represents a compromise between the unrestricted unobserved heterogeneity that FE and CMLE allow for and the restrictive and ad-hoc assumptions underlying CRE models. I achieve this compromise through the exchangeability assumption proposed in Altonji and Matzkin (2005), which allows me to derive a flexible specification of the unobserved heterogeneity. This flexible specification can capture a wide range of correlation forms between the unobserved heterogeneity and the observed covariates in the model.

If the exchangeability assumption holds, the PF-CRE estimator has several advantages relative to the FE and CMLE approaches. Unlike the FE approach, it does not suffer from the incidental parameters problem. It also allows for the estimation of probabilities and partial effects, which cannot be done with CMLE. PF-CRE also provides more efficient estimates of the model parameters than FE and CMLE. This is because FE and CMLE account for every possible form of correlation between the covariates and the unobserved heterogeneity, even when it is not necessary. Additionally, PF-CRE addresses some of the issues stemming from rare-events data with unobserved heterogeneity by avoiding the pitfalls of 'losing' all-zero units in a similar way as CRE models do in both linear (Clark and Linzer, 2015; Bell and Jones, 2015) and non-linear (Crisman Cox, 2019) environments.

PF-CRE also has benefits relative to traditional CRE models. First, by using a flexible (and thus more general) specification for the unobserved heterogeneity it can better capture it; in fact, traditional CRE specifications are nested within PF-CRE. Additionally, the penalization step in PF-CRE selects the minimal specification that captures the correlation between the observed covariates and the unobserved heterogeneity, which can lead to efficiency gains as well. In other words, FE and CMLE assume there is no information in pure cross-sectional variation (an assumption that exacerbates issues in rare-events data). PF-CRE allows cross-sectional variation to be informative of the parameter vector β when the estimated specification is sufficiently sparse (i.e., when few γ parameters are non-zero). This penalization step can also lead to more efficient estimates relative to CRE when the unobserved heterogeneity is non-pervasive.

4 Specification Test

The method outlined in Section 2 requires that the unobserved heterogeneity in the data can be appropriately captured through the flexible correlation specification represented by the polynomial terms. This does not necessarily hold in every application and this assumption is not directly testable. However, it is possible to indirectly test PF-CRE assumptions in the logistic case.

If the correlation between the observed and unobserved components of the model can be correctly captured by the polynomial terms (that is, when the assumptions hold), then the PF-CRE estimator developed in this paper is both consistent and efficient. For the logistic case, the CMLE estimator provides a consistent estimator of the model parameters. Under the null hypothesis that the unobserved heterogeneity can be sufficiently captured by the PF-CRE specification, the PF-CRE estimator is both consistent and efficient, whereas the CMLE estimator is consistent but inefficient. Under the alternative hypothesis, the PF-CRE estimator is inconsistent, but the CMLE estimator remains consistent.[40] Following Hausman (1978), I construct a specification test based on the standardized squared difference between these two estimators. That is, the test statistic is defined as:

$$\delta = d' V(d)^{-1} d, \quad with \;\; d = \widehat{\beta}_{CMLE} - \widehat{\beta}_{PF-CRE}, \tag{12}$$

where $V(d)$ is the variance of d.

Under the null hypothesis, δ is asymptotically distributed χ^2 with k degrees of freedom. This is because both estimators are asymptotically normal with identical means under the null hypothesis, and therefore their difference, d, is asymptotically normal with mean zero. The $\chi^2_{(k)}$ distribution follows from δ being the sum of the squares of k normally distributed terms.

Under the null hypothesis, the variance $V(d)$ has a simple expression due to the efficiency of the PF-CRE estimator:[41]

$$V(d) = V(\widehat{\beta}_{CMLE}) - V(\widehat{\beta}_{PF-CRE}). \tag{13}$$

Hence, putting equations 12 and 13 together:

$$\delta \equiv \left(\widehat{\beta}_{CMLE} - \widehat{\beta}_{PF-CRE}\right)' \left(V(\widehat{\beta}_{CMLE}) - V(\widehat{\beta}_{PF-CRE})\right)^{-1} \left(\widehat{\beta}_{CMLE} - \widehat{\beta}_{PF-CRE}\right). \tag{14}$$

Thus, when the test statistic δ takes a small value, there is no evidence to reject the null hypothesis

[40] The reason the test is restricted to the logistic case is that CMLE is consistent only for the logistic case. Semi-parametric alternatives to CMLE provide consistent estimates of the model parameters for any distribution. However, the convergence rates of these estimators is slower than $\sqrt{n}$. For this reason, asymptotic comparisons with the PF-CRE estimator, which converges at rate $\sqrt{n}$, are not well defined.

[41] Hausman (1978) shows that the variance of the difference between two consistent estimators when one of them is efficient is the difference of the variances.

that the PF-CRE estimator of β is consistent and efficient. This, in turn, provides indirect evidence of the validity of the PF-CRE assumptions.

5 Simulations

I conduct a series of studies to analyze the performance of the PF-CRE estimator in finite samples and compare it to that of alternative methods. This is important generally, but particularly given the pointwise asymptotic convergence of PF-CRE, as well as the approximate nature of the computational algorithm.

In the first set of simulations I analyze the performance of PF-CRE in estimating model parameters comparing it Mundlak's Correlated Random Effects (CRE), the Fixed Effects estimator (FE), and the Conditional Maximum Likelihood estimator (CMLE).[42] [43]

In the second set, I compare the estimator's performance for Average Partial Effects. This set includes the Linear Probability Model and excludes CMLE (since it does not provide estimates of partial effects and probabilities). The third set of simulations deals with the estimation of probabilities. Finally, the fourth set of simulations study the specification test for PF-CRE in the logistic case. Appendices B, C, and D present additional simulations: for rare-events data, for a case in which the exchangeability assumption in PF-CRE is violated, for the coverage rate of confidence intervals, and for computing time.

The data generating process in all simulations is given by:

$$Prob(y_{it}=1|x_{it},c_i)=\Lambda(\alpha+x_{it}\beta+c_i),\ \ with\ \ x_{it}\in\mathbb{R}^4,\ \ \beta=(1.5,1,0.5,1),\ \ \alpha=0.5, \tag{15}$$

where $\Lambda(\cdot)$ is the logistic cumulative distribution and:

$$\tilde{\mathbf{u}}_\mathbf{i}\sim\mathcal{N}(0,\mathbf{V}),\ with\ \ \mathbf{V}_{jk}=0.25^{|j-k|},\ j,k=1,...,4, \tag{16}$$

$$u_{1i}=\tilde{u}_{1i}^2,\ \ u_{2i}=\tilde{u}_{2i},\ \ u_{3i}=\tilde{u}_{2i}\times\tilde{u}_{3i},$$

$$x_{kit}=\tilde{u}_{ki}+\mathcal{N}(0,0.5),\forall\ k=1,...,4,$$

$$c_i=u_{1i}+u_{2i}+u_{3i}$$

[42] Mundlak (1978)'s specification of CRE uses the time-means of the covariates to model the unobserved heterogeneity. It is the same CRE specification used in Crisman Cox (2019)

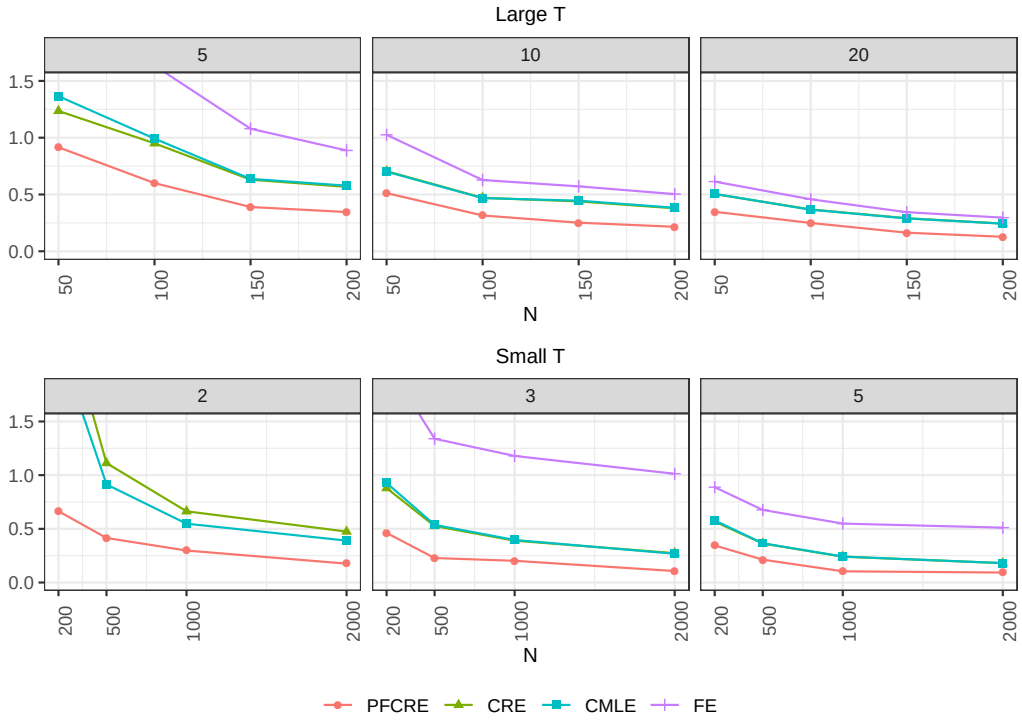
[43] Other estimators are available. However, I exclude the bias-corrected Fixed-Effects estimator from Fernandez-Val (2009) as its performance is quite poor (unreported simulations). I also exclude the methods from Beck (2015) and Cook et al. (2020) since Crisman Cox (2019) shows they are both typically outperformed by the traditional CRE, which is nested in PF-CRE.

This specification for the data generating process has several properties that make it interesting to study. First, the model is relatively simple and it can therefore illustrate the efficiency gains of PF-CRE relative to CMLE and FE. Second, because of the inclusion of an interaction and quadratic term, the traditional CRE approach is misspecified.

The simulations are separated into two groups: 'small T', with $T \in \{2, 3, 5\}$ and $N \in \{200, 500, 1000, 2000\}$; and the 'large T' group, with $T \in \{5, 10, 20\}$ and $N \in \{50, 100, 150, 200\}$. Simulation results are based on 250 draws from the data generating process in each case, except for the specification test simulations which rely on 1,000 simulations.

Figure 2 presents the Root Mean Squared Error (RMSE) of β_1 for the four estimators considered. PF-CRE has the lowest RMSE among the estimators for all N and T sizes. CMLE and Fixed-Effects account for more heterogeneity than there actually exists in the data generating process, leading to higher standard deviations, and consequently a higher RMSE. In the case of Fixed-Effects, this higher standard deviation is compounded by the bias stemming from the incidental parameters problem when T is small, further increasing the RMSE for this estimator. The final estimator, the traditional CRE approach, has a performance that is similar to that of CMLE. However, the similarity in RMSEs between these two estimators comes from a different composition of bias and variance: the CRE estimator is biased, especially for smaller T sizes, but it achieves a smaller variance because of the simplicity of its specification compared to FE and CMLE.

Figure 2: *RMSE for* β_1



The penalty parameter λ *is selected in each individual iteration for PF-CRE; thus, these simulations also incorporate uncertainty over the optimal penalty parameter.*

The better performance of PF-CRE in terms of RMSE is a reflection of two characteristics of the estimator. First, PF-CRE's flexible specification allows the estimator to better capture the underlying unobserved heterogeneity in the data generating process. In fact, a version of the PF-CRE estimator without the penalization step, denoted as F-CRE in Appendix Figures A1 A2 and A3, has a performance in terms of bias that is the same as the Oracle estimator.[44] Second, the penalization step selects the necessary terms to capture the unobserved heterogeneity and discards those that are unnecessary, thus reducing the standard error of the estimator by removing nuisance parameters. This is clearly evidenced in Figure A3 which depicts the standard errors from estimates of PF-CRE, the unpenalized F-CRE, and the Oracle estimator. Thus, the flexible specification of PF-CRE addresses the bias concerns, whereas the penalization of PF-CRE reduces the variance by focusing on the flexible specification terms that actually capture the unobserved heterogeneity and discards those that do not.

It is clear from Figure 2 that PF-CRE has a lower RMSE than alternative estimators and is therefore preferable from this point of view. However. PF-CRE has higher computational demands

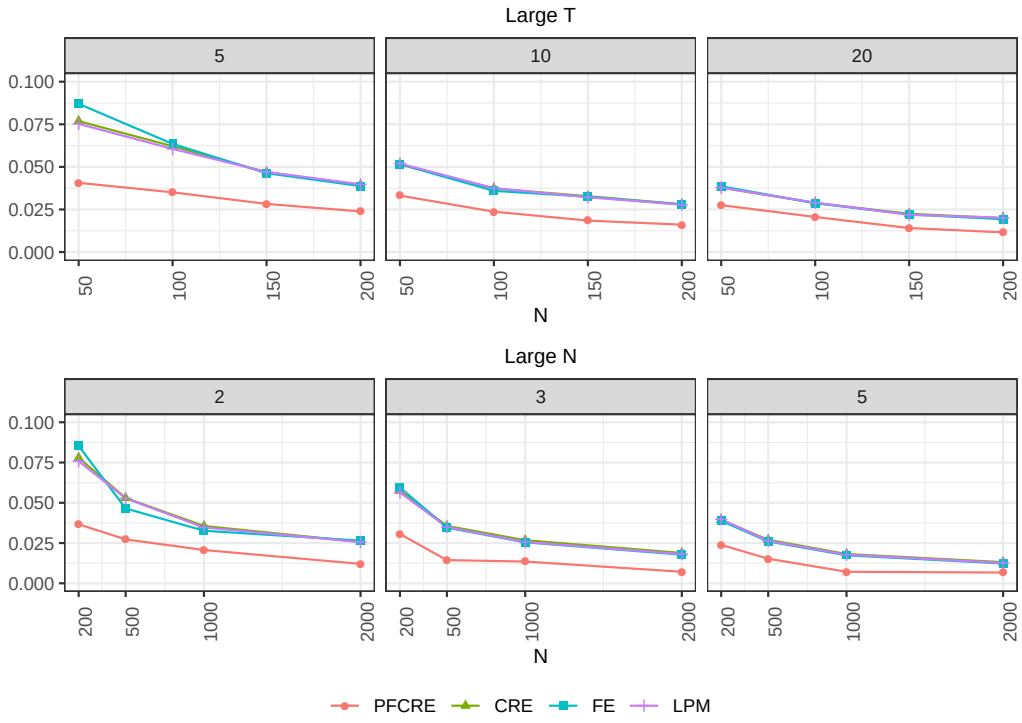
[44]The Oracle estimator is the maximum likelihood estimator that uses the exact specification for the unobserved heterogeneity of the data generating process (which is unfeasible in real applications).

that could hinder its applicability. Despite this, simulations in Appendix B.1 show that while PF-CRE has a higher computational costs, they are not prohibitive by any means.

Similarly to the parameter estimates, the RMSE of PF-CRE for estimates of the average partial effect (APE) is also the smallest among the estimators presented in Figure 3. This advantage is smaller for larger sample sizes, but it is not expected to disappear asymptotically given that PF-CRE is an efficient estimator and the alternative ones are not.

The performance of the Linear Probability Model, Fixed-effects, and CRE are similar to one another, despite being quite different estimators. The vast majority of this RMSE derives from higher standard errors, as the bias in estimating average partial effects tends to be smaller than that of the model coefficients themselves (partly a consequence of the 'average' in APEs). This is particularly interesting in the case of FE since while it is clearly inferior to CMLE in estimating the model parameters, it has a similar performance in terms of APEs.

Figure 3: *RMSE for Average Partial Effect of* x_1



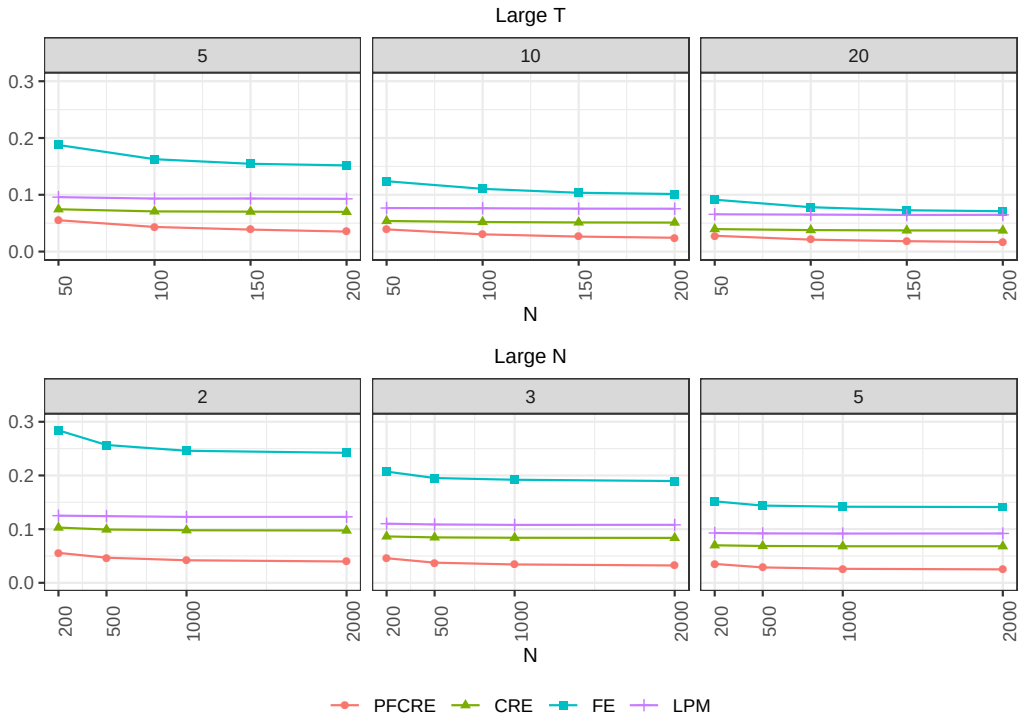
The penalty parameter λ *is selected in each individual iteration for PF-CRE; thus, these simulations also incorporate uncertainty over the optimal penalty parameter.*

While the average partial effect is usually the main quantity of interest to researchers, other quantities may be of interest too. Predicted probabilities are among those other quantities of interest.

Not only they may be of interest by themselves, but they are also a constituent part of partial effects at interesting values of the covariates.

PF-CRE has the least average absolute bias in estimating predicted probabilities for each individual in the sample (Figure 4). All other estimators have a higher absolute bias. CRE and LPM have a performance similar to each other. But their performance does not seem to improve noticeably with sample size (particularly not in N). Fixed-effects has the poorest performance in the estimation of predicted probabilities. This poorer performance will translate into poorer partial effect estimates.

Figure 4: *Average Absolute Bias for Individual Probabilities*



The penalty parameter λ is selected in each individual iteration for PF-CRE; thus, these simulations also incorporate uncertainty over the optimal penalty parameter.

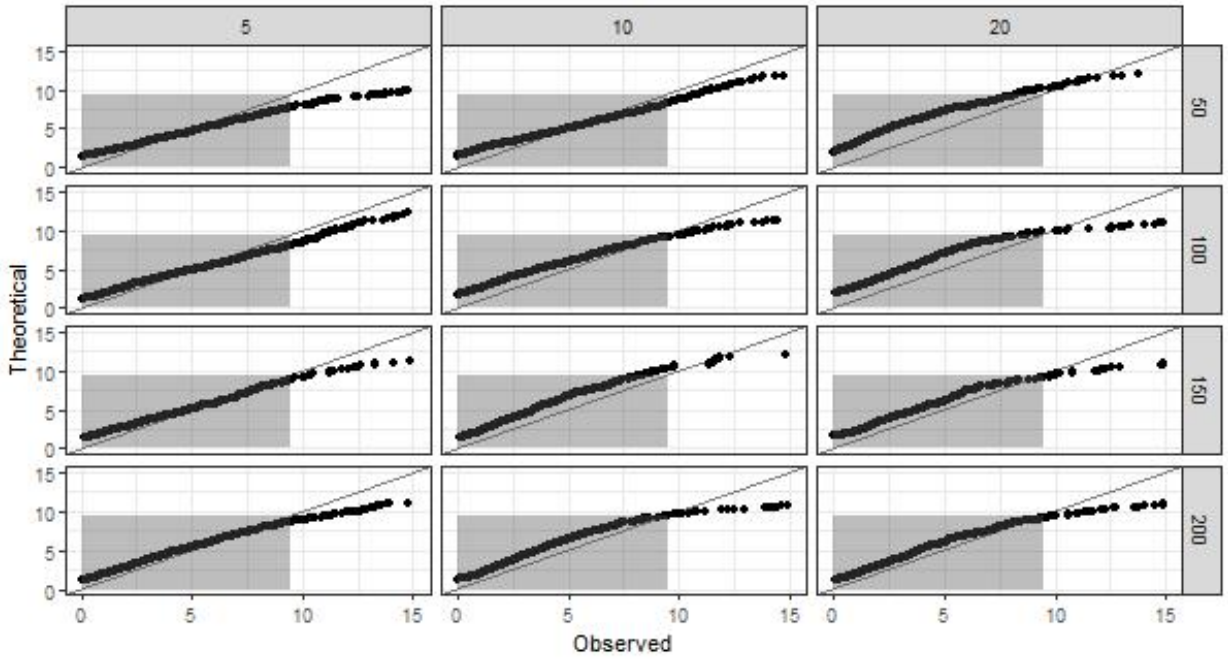
Figures 5 and 6 present the results of simulations for the specification test. These results show that, across 1,000 simulations, the empirical distribution of the specification test approximates theoretical expectations of a χ^2 distribution. However, for the larger T sizes, there is a slight tendency to over-reject, whereas for the smaller T sizes there is a slight tendency to under-reject the null hypothesis, especially when N is large too. This less-than-ideal performance of the specification test comes from a common issue in Hausman-type tests: the tendency to have low power, mainly because they make difficult comparisons between competing models. For the linear models, in fact, Clark and Linzer (2015) and Bell and Jones (2015) recommend interpreting the results of these types of test with care

because of this and related issues.

To help determine whether the less-than-ideal performance of the specification test is indeed due to these common problems rather than to the pointwise (as opposed to uniform) convergence of PF-CRE, I conduct additional simulations in Appendix E that compare estimates from the Oracle estimator and CMLE. The simulations show that the Oracle, which although unfeasible in real life applications converges uniformly, has a similar performance to PF-CRE.

While the specification test overall works as expected, the results from the test should be taken with some caution given its tendency to slightly over and under-reject, depending on the circumstances. Nonetheless, the test still follows the general expected theoretical behavior and is a useful indicator of whether the assumptions underlying PF-CRE are warranted in a particular application of the method, although it is always beneficial to additionally conduct a visual inspections of the estimates compared to those of CMLE.

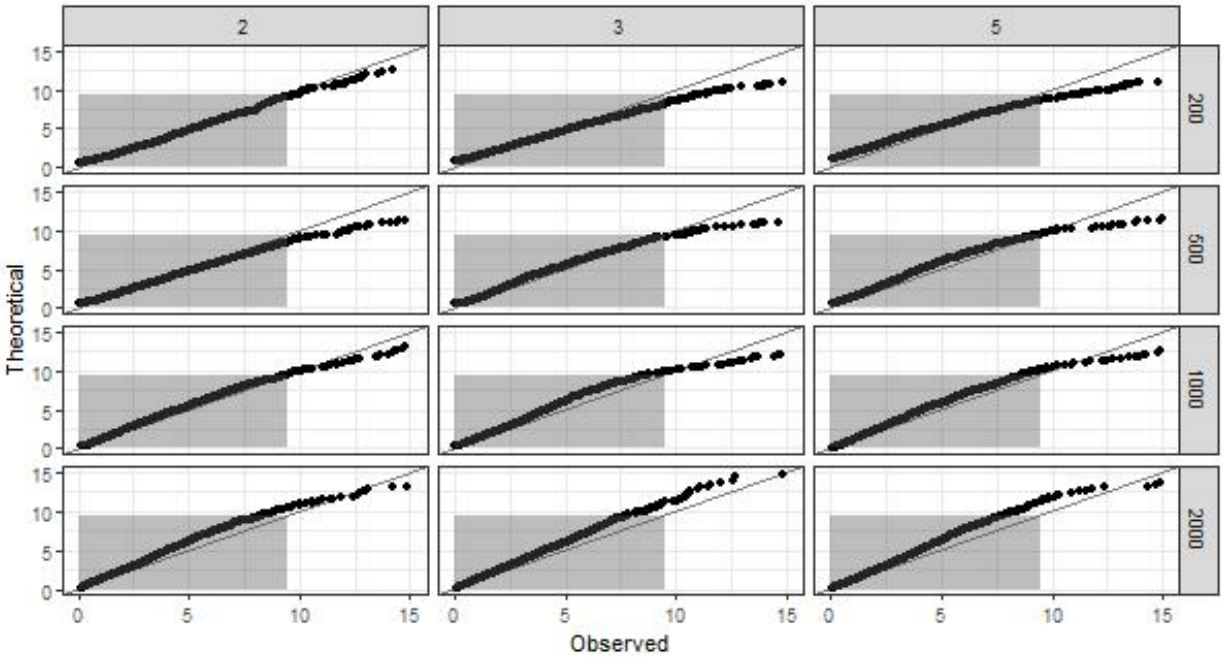
Figure 5: *Quantile-Quantile Plot for Specification Test: Large T*



Observed are the sample quantiles from the simulations. Theoretical are the theoretical quantiles from a $\chi^2_{(4)}$*. The shaded area represents the 95% theoretical quantile.*

Overall, the simulations presented here, together with those in the appendix, show that the asymptotic properties of PF-CRE derived in Section 2 travel well to finite samples, despite PF-CRE converging pointwise rather than uniformly and the approximate nature of the algorithm. The PF-CRE estimator produces estimates of the model parameters that are more efficient than those of the CMLE

Figure 6: *Quantile-Quantile Plot for Specification Test: Small T*



Observed are the sample quantiles from the simulations. Theoretical are the theoretical quantiles from a $\chi^2_{(4)}$*. The shaded area represents the 95% theoretical quantile.*

estimator when the data generating process for the unobserved heterogeneity satisfies the assumptions of PF-CRE. Simulations in Appendices C and D show that these advantages remain for rare-events data and are robust to some violations of the exchangeability assumption. In addition, the simulations also show the advantages of PF-CRE in estimating average partial effects and predicted probabilities, having better finite sample properties than the traditional correlated random effects estimator, the fixed effects, and the linear probability model. Finally, the simulations also show that the specification test has a distribution that is close to theoretical expectations, although it has a tendency to over-reject for small N, large T datasets and to under-reject in large N, small T datasets, and should therefore be interpreted with some care.

6 Short Panel Application: Tactical Voting in the U.K.

In elections with more than two candidates, voters often cast tactical votes. That is, when they believe their most preferred candidate is unlikely to win, they often vote for a less preferred candidate with chances of winning, if only to prevent their most disliked one from being elected (Duverger, 1954).[45]

[45] I use the term tactical voting instead of strategic voting, as it is the common denomination used for this behavior in Britain.

The literature on tactical voting has generally focused on measuring its extent, but less on why some voters behave tactically while others do not. As such, work has focused on voter demographics or electoral circumstances associate with tactical behavior: weak partisan or ideological attachments (e.g., Blais, 2002); political knowledge and education (e.g., Alvarez et al., 2006); as well as familiarity with the electoral system (e.g., Spenkuch, 2017; Duch and Palmer, 2002); the closeness of the election (e.g., Kiewiet, 2013; Elff, 2014; Núñez, 2016); or availability of a close substitute (Karp et al., 2002). While these correlates are important, they are, however, non-actionable: they are not variables an electoral participant can modify. For this reason, I study the one of the most common actionable variables in political campaigns: out-reach to voters.

The empirical challenge in determining the impact of local campaigns on voters' propensity to cast a tactical vote lies in separating the effect of party contacts themselves, from the fact that parties will try to contact those already likely to be swayed their way.

Another way of thinking about this empirical challenge is that from a researcher's point of view, the process or information about voters that parties use to choose which types of voters to contact constitutes unobserved heterogeneity in voters' behavior that is correlated with being contacted by a party. To the extent that which voters are in contention throughout the campaign *and* parties' local campaign strategies remain relatively stable, this unobserved heterogeneity is constant in time. This does not mean that parties' must be contacting the same voters, but instead, the same types of voters.

Thus, to address this challenge, I use a panel data survey collected prior to the 2015 United Kingdom General Election, covering the three months prior to election day. Controlling for unobserved heterogeneity using PF-CRE allows me to reduce or eliminate the concerns stemming from parties' choosing which voters to reach out to, that would otherwise generate upward bias in the estimates. I restrict the sample to respondents that reported vote intention and party preferences in at least two waves of the panel. This leaves 3,824 respondents for a total of 10,378 observations. I impute missing values for other variables using the package `mice` in `R` (Buuren and Groothuis-Oudshoorn, 2011).

The analysis focuses on those voters whose most preferred party is not viable, a common approach in the study of tactical voting (Alvarez et al., 2006). I define a party as viable if it finished among the top-two in a given district. I define voters' most preferred party in the following way: (1) the party with the highest thermometer score; (2) if there are ties, these are broken by the thermometer scores for the leaders of the corresponding parties; (3) if ties remain, then all tied parties are considered the

voters' most preferred party.[46] I define voters' *most preferred viable* party as the most preferred party from among the viable ones.

The covariates of interest are indicators for whether a voter's most preferred party or most preferred viable party contacted the voter during the four weeks prior to each wave. I also include as dependent variables the thermometer score for the most preferred and most preferred viable parties as reported by each respondent, measured on a scale from 1 to 10.[47]

Given that I use the logistic distribution in this application, I compare the coefficient estimates from the PF-CRE estimator with those of CMLE. While both PF-CRE and CMLE account for unobserved heterogeneity, only PF-CRE allows for the estimation of partial effects. Additionally, I also include the FE estimator for comparison, although since the T dimension of the data is small ($2 \leq T \leq 3$, depending on the individual) it is known to be biased. Finally, I also include the traditional CRE specification for comparison.

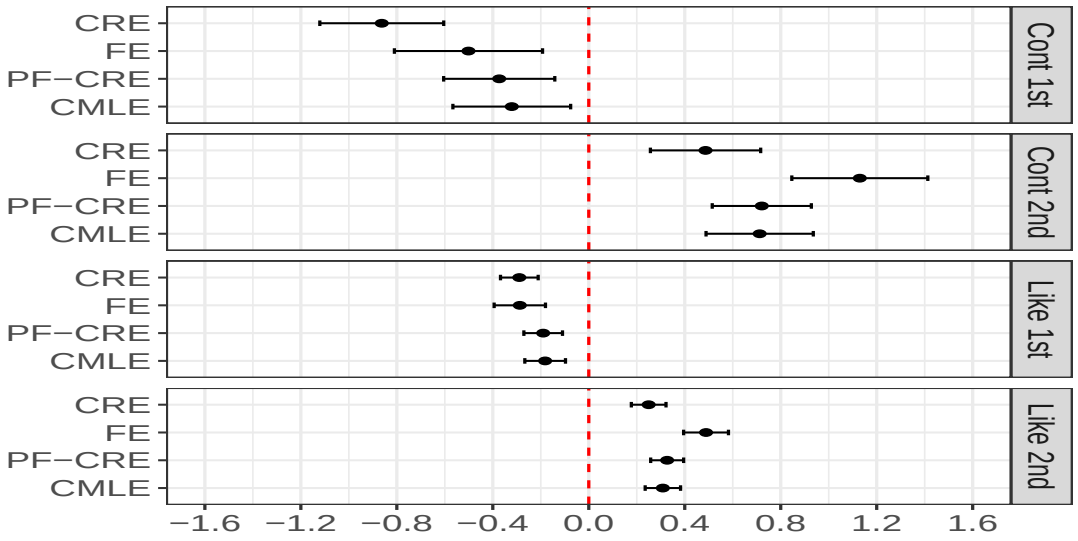
Figure 7 shows the coefficient estimates from PF-CRE, CMLE, CRE, and FE.[48] PF-CRE estimates look remarkably similar to the CMLE ones. Indeed, the specification test does not reject the null hypothesis that PF-CRE is consistent and more efficient than CMLE, with a p-vlaue of 0.48. The traditional CRE approach, however, shows estimates that differ from CMLE, with the specification test rejecting its validity. This application illustrates that the traditional CRE method is not always robust to misspecification of the unobserved heterogeneity. The FE estimates are biased, which is expected, and also have a higher standard error than the other estimates.

[46] In these cases, a tactical vote for these voters only occurs when none of their most preferred parties are viable and they cast a vote for the most liked viable party.

[47] I also produces estimates (not reported) that include a number of time-invariant characteristics: employment status, retirement status, education, gender, age, and home ownership. When including these time-invariant characteristics, the estimates are qualitatively and quantitatively similar.

[48] See Table F1 in the appendix for details with the estimates from the three models.

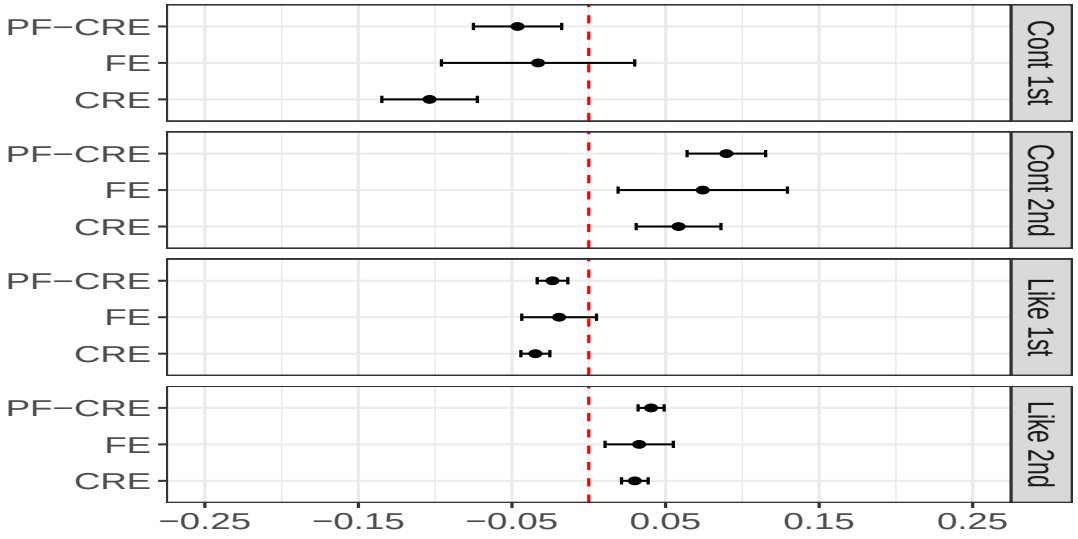
Figure 7: *Coefficient Estimates, Tactical Voting 2015 U.K. Election*



The λ parameter in the SCAD penalty was chosen using the Akaike Information Criterion. All confidence intervals are 95%

Figure 8 presents the partial effects for the PF-CRE, CRE, and FE estimators. While the CMLE and PF-CRE coefficient estimates are indistinguishable from one another, only the PF-CRE estimator provides estimates of probabilities, partial effects, and average partial effects.

Figure 8: *Average Partial Effects, Tactical Voting 2015 U.K. Election*



The λ parameter in the SCAD penalty was chosen using the Akaike Information Criterion. All confidence intervals are 95%

The PF-CRE estimates of the average partial effect show that when respondents are contacted

by their most preferred party, they are 4.6% less likely to cast a tactical vote for a less preferred party, suggesting that party contact enforces party loyalty or sincerity in voters. The CRE estimator, instead, estimates this effect at 7.5%. Interestingly, being contacted by the most preferred viable party has a countervailing effect that is stronger than being contacted by the most preferred party, increasing the probability of casting a vote for a less preferred party by 6.4% according to PF-CRE estimates. However, the CRE model underestimates this effect at 3%, about half the size of the PF-CRE's estimate. The FE estimates lie somewhat in between the estimates from CRE and PF-CRE, thus having lower bias than CRE in this case. However, they have a substantially larger variance, a reflection of the efficiency costs of allowing for unrestricted heterogeneity when it is not in fact necessary.

The results presented here illustrate that the CRE method cannot always fully account for the unobserved heterogeneity. In particular, coefficient estimates from CRE are statistically different from those of CMLE and PF-CRE. Additionally, average partial effects from CRE are noticeably different from those of PF-CRE. CRE estimates might be a tempting approach given PF-CRE's higher computational demands. However, the estimation of PF-CRE took 5.15 minutes, whereas that of CRE took 1.56 minutes. While this is a three-fold increase in computing time, it is still a relatively short wait. More importantly, while CRE does account for some unobserved heterogeneity (compared to unreported logit estimates), CRE estimates retain a noticeable amount of bias in this application. Therefore, the bias reduction from PF-CRE more than warrants its higher computation time.

7 Long Panel Application: Economic Sanctions

Here I consider a large T panel that analyzes the destabilization effects of economic sanctions in the countries targeted. Based on the equilibrium of a formal model, Marinov (2005) hypothesizes that economic sanctions should, on average, destabilize the governments of the countries they target.

The challenge in estimating these effects is the potential for omitted variables. In particular, Marinov (2005) notes that country-specific factors like unique political cultures or historical experiences are likely to influence government stability possibly generating bias in the estimates. Therefore, the use of models that account for time-invariant unobserved heterogeneity is necessary.

The data come from Marinov (2005). The dependent variable is an indicator for leadership change

in a given country and year. The data includes information about 160 countries from 1947 to 1999, with up to 50 years of data for each, for a total of 5,766 observations.[49] The main independent variable of interest is economic sanctions collected from an updated version of Hufbauer et al. (1990), transformed into an indicator for whether the country was subject to sanctions in a given year. The data also includes several time-varying control variables: indicator variables for involvement in a militarized-interstate dispute with use of force, economic growth rate, GDP per capita, indicators for democratic and for mixed regimes (also interacted with time), the incumbent leaders' age and time in office, and a cubic spline.

The main analysis in Marinov (2005) corresponds to the first column of Table 2, which is replicated here.[50] The original paper used CMLE to estimate the coefficients of the main table, but it also used the fixed-effects estimator to produce substantive effects. While the original paper estimated the average increase in *risk* of losing power from economic sanctions, here I estimate the average partial effect, or the increase in probability of losing power.[51]

Since the original results are also derived from a logistic model, I use the logistic distribution for PF-CRE. I compare both parameter estimates and estimates of the average partial effect for the effect of economic sanctions on leader survival using four estimators: CMLE, Fixed-Effects, a traditional CRE, and PF-CRE.

Figure 9 presents the estimated coefficients for the first five independent variables reported in the main results in Marinov (2005) (see all covariates in Table F2). As the figure shows, all four estimators produce very similar point estimates. Since CMLE imposes no restrictions on the unobserved heterogeneity, it can be used as a benchmark for the truth. The Fixed-Effects (FE) estimator produces point estimates that are similar to those of CMLE and with a similar degree of uncertainty. This is to be expected given that T is large and, therefore, FE should have a negligible bias.

The point estimates from the traditional CRE model and the PF-CRE model are also similar to those of CMLE, suggesting that both are valid approaches in this case and can capture the unobserved heterogeneity.[52] However, PF-CRE has smaller standard errors, which is reflected in the shorter

[49]There are a total of 6,782 observations before accounting for missing values in one of more covariates.

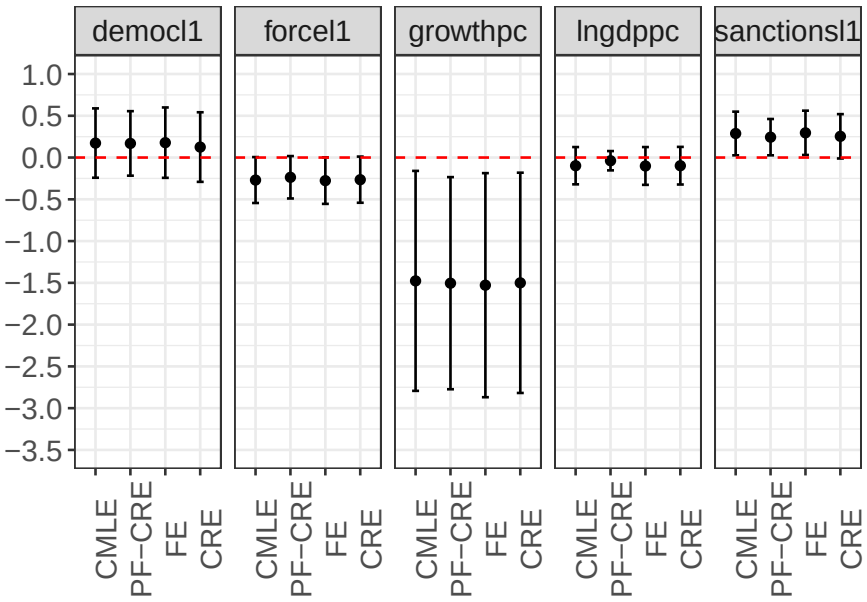
[50]Please note that there are minor differences between the estimates in the original paper and the ones presented here due to different rounding methods.

[51]These two quantities are derived from the same calculations. Change in risk refers to the relative change in the probability of the outcome occurring, whereas the average partial effect measures the absolute change in the probability of the outcome occurring.

[52]The specification tests supports the validity of PF-CRE in this application, with a p-value of 0.95. This is not the case with the CRE model, despite the extreme similarity of the point estimates. This is due to CRE not having a

confidence intervals. This illustrates the efficiency advantage of the PF-CRE approach relative to alternative estimators, even if they all produce very similar point estimates.

Figure 9: *Coefficient Estimates for Marinov (2005)*



The λ parameter in the SCAD penalty was chosen using the Akaike Information Criterion. All confidence intervals at 95%. Full results are presented in Table F2 in the Appendix.

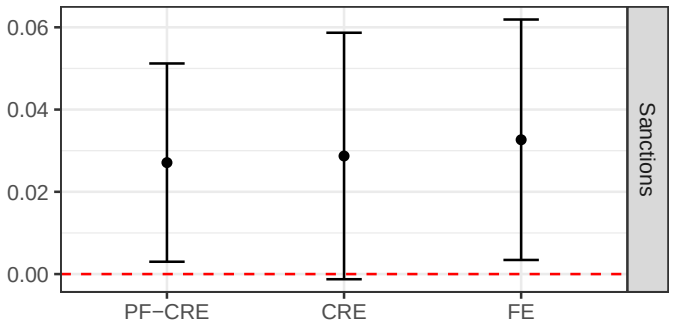
Figure 10 presents the average partial effect of economic sanctions on leader survival, the main variable of interest, obtained from the three estimators that allow for the estimation of these effects: the traditional CRE, FE, and PF-CRE. The point estimates are very similar for the three estimators. FE estimates that economic sanctions increase the probability of leadership change by 3.3%, while PF-CRE estimates this quantity at 2.7%, and CRE at 2.9%.

However, despite the point estimates being very similar across the three estimators, PF-CRE has a smaller standard error. In fact, the standard error of PF-CRE is 82% of the FE standard error and 80% of the CRE standard error. This instance clearly illustrates the efficiency advantage of PF-CRE relative to both FE and CRE. FE spends many degrees of freedom in estimating the individual fixed-effects, which translates into increased uncertainty for estimated probabilities and average partial effects. CRE, estimates fewer parameters, but does not necessarily capture the unobserved heterogeneity very precisely. The PF-CRE estimator, on the other hand, obtains a more efficient estimate thanks to the combination of a flexible specification that captures the unobserved heterogeneity more precisely and a penalization step that avoids the estimation of too many parameters. The precision

smaller variance than CMLE.

gains from PF-CRE relative to CRE do not come at a high computational cost: the estimation of the CRE model took 0.5 minutes, while that of PF-CRE took 2.1 minutes. While it is a 4-fold increase in time, the additional 1.5 minutes in estimation time are worth the efficiency gains.

Figure 10: *Average Partial Effects for Marinov (2005)*



The λ parameter in the SCAD penalty was chosen using the Akaike Information Criterion. All confidence intervals at 95%.

8 Conclusion

Unobserved heterogeneity is pervasive in observational studies in political science, and the social sciences in general. Whatever its origins and form, all unobserved heterogeneity poses the same problem: if ignored, and correlated with the covariates of interest, it leads to biased and inconsistent estimates. One of the best ways to deal with unobserved heterogeneity is to use panel data. However, a standing problem in the case of binary outcomes (and discrete outcomes generally) is that consistent estimators of the model parameters do not allow for the estimation of partial effects, which are usually the quantity of interest to researchers.

In this paper, I develop the *Penalized Flexible Correlated Random Effects* (PF-CRE) estimator for binary outcome models with panel data. PF-CRE provides consistent and efficient estimates of the model parameters and partial effects. It relies on adopting a flexible specification for the unobserved heterogeneity that is complemented with a penalization step for variable selection. The flexibility is derived from mild assumptions on the unobserved heterogeneity that contribute to the estimator's consistency, and the penalization step induces a parsimonious model that results in efficiency gains. Moreover, I provide a model specification test for the logistic case as an indirect test of PF-CRE's underlying assumptions. While PF-CRE is more computationally demanding that alternative estimators, the additional time required is not excessive and is more than warranted by the gains in consistency and efficiency.

The PF-CRE estimator has a number of advantages relative to alternative estimators. Unlike Fixed Effects, it does not suffer from the incidental parameters problem that leads to inconsistent estimates. PF-CRE allows for the estimation of partial effects that the Conditional Maximum Likelihood estimator does not provide. Finally, its assumptions are significantly less restrictive than those of traditional Correlated Random Effects models, meaning that PF-CRE's assumptions are more likely to hold in real world applications.

PF-CRE can be applied in other areas of social science and applications well beyond those included in this paper. The expected benefits from applying PF-CRE will likely vary by application. In small T environments, the most important gain from PF-CRE is the ability to produce consistent estimates of partial effects and probabilities under much milder assumptions than alternative estimators. In large T environments, the PF-CRE's main advantage are its efficiency gains relative to alternative consistent estimators. This efficiency gain might be substantial in some applications with implications to substantive results; for example in areas like comparative political institutions and international relations, where most of the variation in the data is usually across units and within unit variation is typically much smaller. Methods like CMLE and Fixed Effects tend to discard most of the information in the data, often leading to statistically non significant results in these environments. The alternative is to ignore unobserved heterogeneity, which is also not desirable as it introduces bias. The appeal of PF-CRE in these cases is that while it accounts for unobserved heterogeneity, it does not discard all cross-sectional variation in the data, as my large T panel replication demonstrates. This is what the penalization step accomplishes: if it selects a relatively sparse specification for the unobserved heterogeneity, a significant portion of cross-sectional variation will still be used to estimate the parameters of interest, partial effects, and probabilities.

Acknowledgements

I would like to thank R. Michael Alvarez, Jonathan Katz, and Robert Sherman for helpful advice and invaluable guidance in this project. I would also like to thank Marcelo Fernández, Alejandro Robinson-Cortes, Jun Chen, and Laural Doval for their assistance and encouragement, as well as other members of the Caltech community. Finally, I also thank the editor and three anonymous referees for their detailed and conscientious feedback that helped improve this work significantly.

References

Abrevaya, J. A. (1997). The Equivalence of Two Estimators of the Fixed-Effects Logit Model. *Economic Letters*, 55:41–43.

Abrevaya, J. A. (2000). Rank Estimation of a Generalized Fixed Effects Regression Model. *Journal of Econometrics*, 95:1–23.

Altonji, J. G. and Matzkin, R. L. (2005). Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors. *Econometrica*, 73(4):1053–1102.

Alvarez, R. M., Boehmke, F. J., and Nagler, J. (2006). Strategic Voting in British Elections. *Electoral Studies*, 25:1–19.

Andersen, E. B. (1970). Asymptotic Properties of Conditional Maximum-Likelihood Estimators. *Journal of the Royal Statistical Society, Series B (Methodological)*, 32(2):283–301.

Ando, T. and Bai, J. (2016). Panel Data Models with Grouped Factor Structure under Unknown Group Membership. *Journal of Applied Econometrics*, 31(1).

Beck, N. (2015). Estimating Grouped Data Models with a Binary Dependent Variable and Fixed Effects: What are the Issues? *Unpublished manuscript*.

Beck, N. (2018). Estimating Grouped Data Models with a Binary Dependent Variable and Fixed Effects: What are the Issues? *http://arxiv.org/abs/1809.06505*.

Beck, N. (2020). Estimating Grouped Data Models with a Binary-Dependent Variable and Fixed Eects via a Logit versus a Linear Probability Model: The Impact of Dropped Units. *Political Analysis*, 28:139–145.

Beck, N. and Katz, J. (2007). Random Coefficient Models for Time-Series-Cross-Section Data: Monte Carlo experiments. *Political Analysis*, 15(2):182–195.

Beck, N. and Katz, J. N. (2001). Throwing out the Baby with the Bath Water: A Comment on Green, Kim, and Yoon. *International Organization*, 55:487–495.

Bell, A. and Jones, K. (2015). Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods*, 3(1):133–153.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. *Econometrica*, 80:2369–2429.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2014). arXiv:1311.2645. *Program Evaluation with High-Dimensional Data*.

Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2016). Inference in High-Dimensional Panel Models with an Application to Gun Control. *Journal of Business & Economic Statistics*, 34:590–605.

Blackwell, M. and Glynn, A. N. (2018). How to Make Causal Inferences with Time-Series Cross-

Sectional Data under Selection on Observables. *American Political Science Review*, 112:1067–1082.

Blais, A. (2002). Why is there So Little Strategic Voting in Canadian Plurality Rule Elections? *Political Studies*, 50:445–454.

Bonhomme, S., Lamadon, T., and Manresa, E. (2017). Discretizing Unobserved Heterogeneity. Wokring paper.

Bonhomme, S. and Manresa, E. (2015). Grouped Patterns of Heterogeneity in Panel Data. *Econometrica*, 83(3):1147–1184.

Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):33–48.

Chamberlain, G. (1980). Analysis of Covariance with Qualitative Data. *Review of Economic Studies*, 47:225–238.

Chamberlain, G. (1984). Panel Data. In Griliches, Z. and Intriligator, M. D., editors, *Handbook of Econometrics*, volume II, chapter 22. North-Holland, Amsterdam.

Chamberlain, G. (2010). Binary Response Models for Panel Data: Identification and Information. *Econometrica*, 78(1):159–169.

Clark, T. S. and Linzer, D. A. (2015). Should I Use Fixed or Random Effects? *Political Science Research and Methods*, 3(2):399–408.

Cook, S. J., Hays, J. C., and Franzese, R. J. (2020). Fixed effects in rare events data: A penalized maximum likelihood solution. *Political Science Research and Methods*, 8:92–105.

Coupe, T. (2005). Bias in Conditional and Unconditional Fixed Effects Logit Estimation: A Correction. *Political Analysis*, 13:292–295.

Crisman Cox, C. (2019). Estimating Substantive Effects in Binary Outcome Panel Models: A Comparison. *Journal of Politics*, Accepted.

Dhaene, G. and Jochmans, K. (2015). Split-Panel Jackknife Estimation of Fixed-Effect Models. *Review of Economic Studies*, 82:991–1030.

Duch, R. M. and Palmer, H. D. (2002). Strategic Voting in Post-Communist Democracy? *British Journal of Political Science*, 32:63–91.

Duverger, M. (1954). *Political Parties: Their Organization and Activity in the Modern State*. Wiley, New York.

Elff, M. (2014). Separating Tactical from Sincere Voting: A Finite Mixture Discrete Choice Modelling Approach to Disentangling Voting Calculi. Paper presented at the 2014 Annual Meeting of the Midwest Political Science Association, Chicago, April 3-6.

Fan, J. and Li, R. (2001). Variable Selection Via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96:1348–1360.

Fernandez-Val, I. (2009). Fixed Effects Estimation of Structural Parameters and Marginal Effects in

Panel Probit Models. *Journal of Econometrics*, 150:71–85.

Fernandez-Val, I. and Vella, F. (2011). Bias Corrections for Two-Step Fixed Effects Panel Data Estimators. *Journal of Econometrics*, 163:144–162.

Green, D., Kim, S. Y., and Yoon, D. H. (2001). Dirty Pool. *International Organization*, 55(2):441–468.

Greene, W. H. (2004). The Behavior of the Fixed Effects Estimator in Nonlinear Models. *Econometrics Journal*, 7:98–119.

Greene, W. H. (2011). Fixed Effects Vector Decomposition: A Maginal Solution to the Problem of Time-Invariant Variables in Fixed Effects Models? *Political Analysis*, 19:135–146.

Greene, W. H. (2015). Panel Data Models for Discrete Choice. In Baltagi, B. H., editor, *The Oxford Handbook of Panel Data*. Oxford University Press.

Groll, A. and Tutz, G. (2014). Statistics and Computing. *Variable Selection in Generalized Linear Mixed Models by L_1-Penalized Estimation*, 24:137–154.

Hahn, J. and Newey, W. (2004). Jackknife and Analytical Bias Reduction for Nonlinear Panel Models. *Econometrica*, 72:1295–1319.

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, 46(6):1251–1271.

Heckman, J. J. (1981). The Incidental Parameters Problem and the Problem of Initial Conditions in Discrete Time-Discrete Data Stochastic Process. In Manski, C. and McFadden, D., editors, *Structural Analysis of Discrete Data with Econometric Applications* . MIT Press, Cambridge.

Hufbauer, G., Shott, J., and Elliott, A. (1990). *Economic Sanctions Reconsidered*. Institute for International Economics, Washington.

Hui, F. K., Muller, S., and Welsh, A. H. (2017). Hierarchical Selection of Fixed and Random Effects in Generalized Linear Mixed Models. *Statistica Sinica*, 27:501–518.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and Random Effects Selection in Mixed Effects Models. *Biometrics*, 67(2):495–503.

Karp, J. A., Vowles, J., Banducci, S. A., and Donovan, T. (2002). Strategic Voting, Party Activity, and Candidate Effects: Testing Explanations for Split Voting in New Zealand's New Mixed System. *Electoral Studies*, 21:1–22.

Kiewiet, D. R. (2013). The Ecology of Tactical Voting in Britain. *Journal of Elections, Public Opinion and Parties*, 23(1):86–110.

King, G. (2001). Proper Nouns and Methodological Propriety: Pooling Dayds in International Relations Data. *International Organization*, 55(2):497–507.

King, G. and Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, 9(2):137–163.

King, G. and Zeng, L. (2006). The Dangers of Extreme Counterfactuals. *Political Analysis*, 14:131–159.

Knight, K. (2008). Econometric Theory. *Shrinkage Estimation for Nearly Singular Designs*, 24.

Lai, R. C., Huang, H.-C., and Lee, T. C. (2012). Fixed and Random Effects Selection in Nonparametric Additive Mixed Models . *Electronic Journal of Statistics*, 6:810–842.

Leeb, H. and Pötscher, B. (2005). Model Selection and Inference: Facts and Fiction. *Econometric Theory*, 21:21–59.

Leeb, H. and Pötscher, B. (2008). Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator. *Journal of Econometrics*, 142:201–2011.

Lin, B., Pang, Z., and Jiang, J. (2013). Fixed and Random Effects Selection by REML and Pathwise Coordinate Optimization. *Journal of Computational and Graphical Statistics*, 22:341–355.

Manski, C. (1987). Semiparametric Analysis of Random Effects Linear Models From Binary Panel Data. *Econometrica*, 55:357–362.

Marinov, N. (2005). Do Economic Sanctions Destabilize Country Leaders. *American Journal of Political Science*, 49:564–576.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica*, 46:69–85.

Neyman, J. and Scott, E. L. (1948). Consistent Estimates Based on Partially Constistent Observations. *Econometrica*, 16:1–32.

Núñez, L. (2016). Expressive and Strategic Behavior in Legislative Elections in Argentina. *Political Behavior*, 38(4):899–920.

Plümper, T. and Troeger, V. E. (2007). Efficient Estimation of Time-Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects. *Political Analysis*, 15:124–139.

Plümper, T. and Troeger, V. E. (2011). Fixed Effects Vector Decomposition: Properties, Reliability, and Instruments. *Political Analysis*, 19:147–164.

Rasch, G. (1961). On General Laws and the Meaning of Measurement in Psychology. *The Danish Institute of Educational Research, Copenhagen.*

Schelldorfer, J., Meier, L., and Buhlmann, P. (2011). GLMMLasso: An Algorithm for High-Dimensional Generalized Linear Mixed Models using L1-Penalization. *Journal of Computational and Graphical Statistics*, 23(2):460–477.

Spenkuch, J. L. (2017). Expressive vs. Pivotal Voters: An Empirical Assessment. Wokring paper.

Stammann, A. (2017). Fast and feasible estimation of generalized linear models with high-dimensional k-way fixed effects.

Su, L., Shi, Z., and Phillips, P. C. B. (2016). Identifying Latent Structures in Panel Data. *Econometrica*, 84(6):2215–2264.

Ulbricht, J. (2012). *lqa: Penalized Likelihood Inference for GLMs.* R package version 1.0-3.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data.* MIT Press, Cambridge, MA, second edition.

Online Appendix to

Partial Effects for Binary Outcome Models with Unobserved Heterogeneity

Lucas Núñez

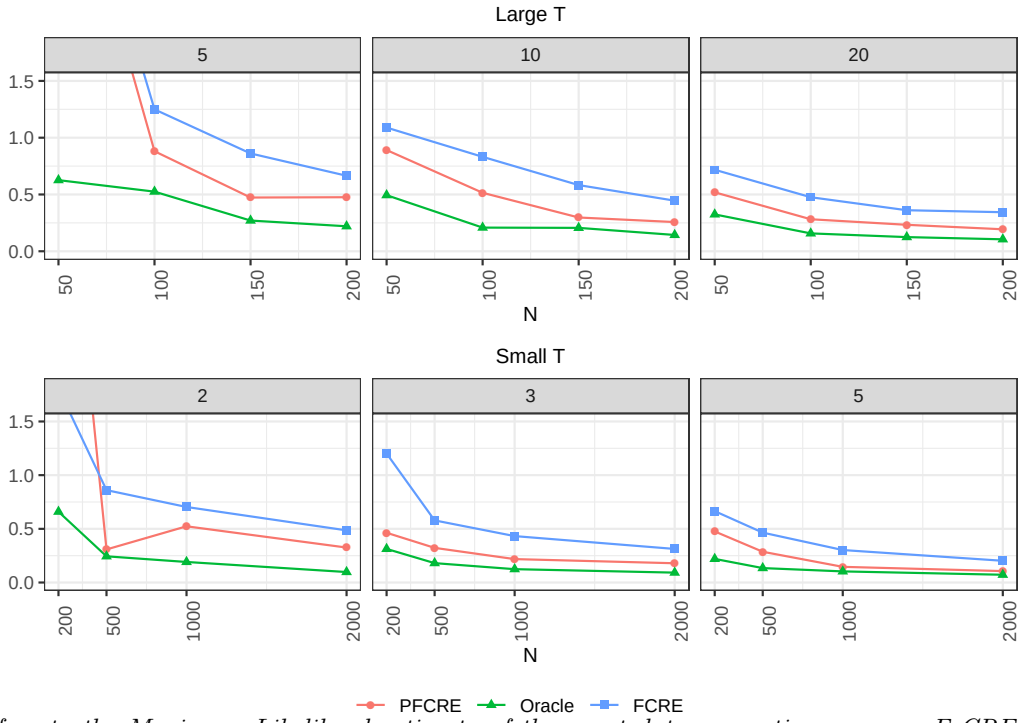
A PF-CRE, F-CRE, and the Oracle Estimator

In this section I provide additional simulations that compare the PF-CRE estimator with the Oracle as well as an unpenalized version of PF-CRE, denoted F-CRE. The goal is to illustrate how to two components of PF-CRE, flexible specification and penalization, work together to accomplish first consistency and then efficiency in the estimates. Each set includes 100 simulations. The data generating process in these simulations is similar to that in the main body of the paper with only one modification. The variance of one of the independent variables was increased from 1 to 4. This makes the selection of polynomial terms more difficult and was purposely included so as not to overstate the ability of the estimator to recover the Oracle estimates.

The simulations show the estimators perform as expected. Figure A1 shows that the Oracle has the smaller RMSE, followed by PF-CRE, and then F-CRE, which is expected, since F-CRE does not include the penalization step and therefore includes many nuisance terms in the polynomial that increase uncertainty over the estimates.

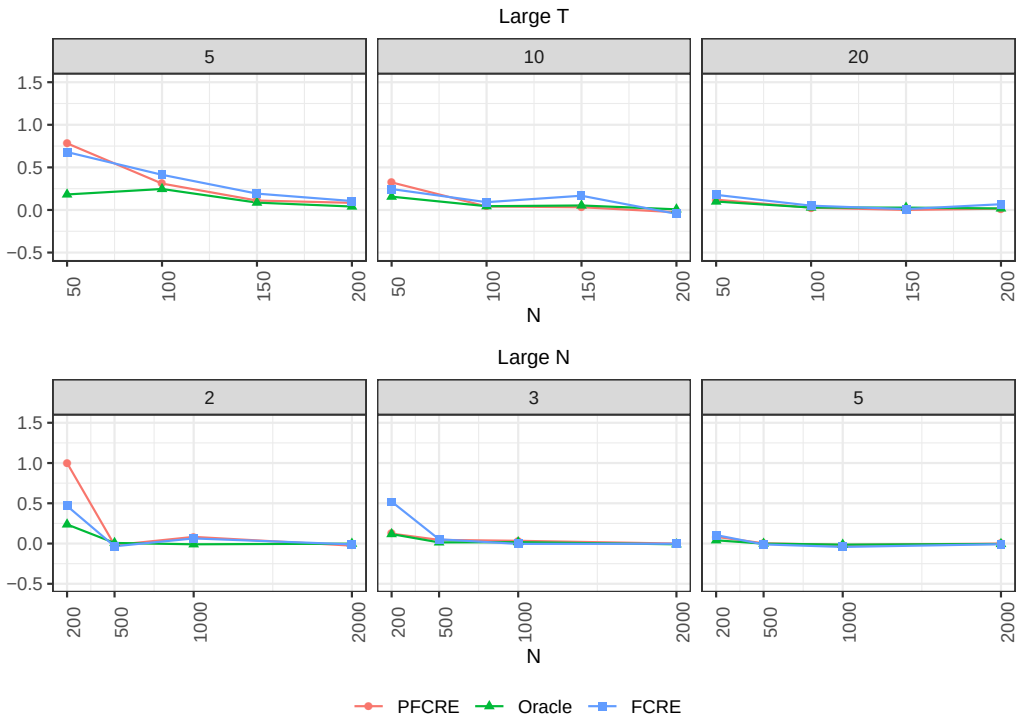
Figures A2 and A3 present the bias and standard error of the three estimators. It is clear from these figures that the consistency result in PF-CRE is in fact achieved by its flexible specification, as F-CRE has an almost identical performance to PF-CRE in terms of bias. Both estimators behave almost identically to the Oracle, having no bias, with the exception of simulations in which both N and T are small. Finally, a comparison of the standard errors of F-CRE and PF-CRE shows that the RMSE gains of the latter accrue from the penalization step.

Figure A1: *RMSE for β from PF-CRE, F-CRE, and the Oracle*



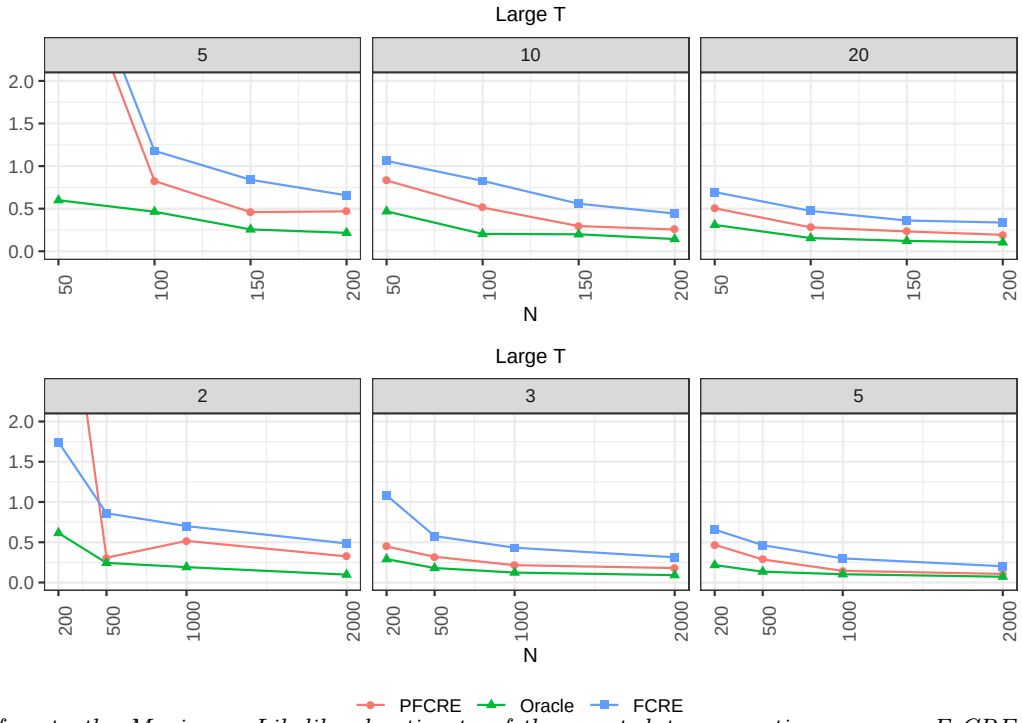
The Oracle refers to the Maximum Likelihood estimate of the exact data generating process. F-CRE is an estimator that uses the same flexible specification of PF-CRE but without the penalization. The penalty parameters λ in PF-CRE is selected in each individual iteration using the Akaike Information Criterion (AIC).

Figure A2: *Bias for β from PF-CRE, F-CRE, and the Oracle*



The Oracle refers to the Maximum Likelihood estimate of the exact data generating process. F-CRE is an estimator that uses the same flexible specification of PF-CRE but without the penalization. The penalty parameters λ in PF-CRE is selected in each individual iteration using the Akaike Information Criterion (AIC).

Figure A3: *Standard Error for β from PF-CRE, F-CRE, and the Oracle*



The Oracle refers to the Maximum Likelihood estimate of the exact data generating process. F-CRE is an estimator that uses the same flexible specification of PF-CRE but without the penalization. The penalty parameters λ in PF-CRE is selected in each individual iteration using the Akaike Information Criterion (AIC).

B Additional Simulation Results

B.1 Computing Times

It is clear from Figure 2 in the text that PF-CRE has a lower Root Mean Squared Error than alternative estimators and is therefore preferable from this point of view. However, PF-CRE has higher computational demands than the other estimators which could hinder its applicability.[53] However, Table B1 shows that while PF-CRE's computing times are indeed larger than those for the traditional CRE, they are not prohibitive by any means. Overall, PF-CRE estimation takes up to around three times as much as that of CRE, and this ratio tends to be lower when the T-dimension of the data is larger.[54] While a threefold increase in computing time could become a concern in very large datasets (models with a large number of variables and observations), it is not so for the more typical datasets used in political science. As described in the text, the estimation time for PF-CRE in the two applications presented in this paper are 5.15 minutes (versus 1.56 minutes for CRE) and 2.1 minutes (compared to 0.5 minutes for CRE), all acceptable wait times.[55]

Table B1: *Computing Times, PF-CRE v. CRE*

Large T				Small T			
PF-CRE	CRE	N	T	PF-CRE	CRE	N	T
9.5	3.4	50	5	17.0	7.1	200	2
11.7	5.6	50	10	15.9	8.5	200	3
18.1	10.8	50	20	20.2	10.9	200	5
12.5	6.1	100	5	40.1	18.3	500	2
17.2	9.9	100	10	46.3	19.7	500	3
29.4	18.9	100	20	61.2	26.5	500	5
15.1	8.7	150	5	92.3	36.2	1000	2
24.8	14.7	150	10	111.6	42.1	1000	3
43.9	27.2	150	20	154.4	63.4	1000	5
20.2	10.9	200	5	243.9	75.6	2000	2
32.6	18.7	200	10	308.0	92.9	2000	3
60.6	38.4	200	20	398.0	124.9	2000	5

Results are in seconds. All times are averages from 100 simulations for each N and T combination.

[53] Estimation times for FE depend on which algorithm is applied. Directly using dummy variables for the fixed-effects can lead to enormous computation times. However, estimation via alternative algorithms like that proposed by Stammann (2017) are very fast.

[54] This relatively low ratio of speeds might seem counter-intuitive at first, since in principle PF-CRE requires the estimation of multiple CRE-type models in the grid-search for the penalization parameter λ. However, the computational implementation takes advantage of the continuous shrinking induced by the SCAD penalty (Fan and Li, 2001) and initializes the algorithm for each λ value on the parameter estimates from the previous one, thus speeding convergence substantially.

[55] These computing times were obtained using non-parallel computing on a Laptop PC with Intel Core i7 2.00GHz Quad Core processor with 8GB of RAM.

B.2 Coverage Probabilities

While the PF-CRE estimator tends to produce smaller standard errors, it is important to determine whether they are not "too" small, in the sense that they underestimate the underlying uncertainty. This can be done by analyzing the coverage probability of 95% confidence intervals derived from the simulations. Coverage should be close to the theoretical 95%, although in finite samples it could be somewhat below the nominal values.

Table B2 presents the coverage probabilities for selected sample sizes for the following estimators: PF-CRE, the Oracle, CRE, CMLE, and Fixed Effects. Coverage probabilities for the CMLE estimator are generally very close to the nominal 95%, although for N = 1000, T = 2 they are slightly high. The Fixed-effects estimator has close to nominal coverage probabilities when T is small, despite it being biased. However, for large T, coverage probabilities are far from nominal levels. The CRE estimator has close to nominal coverage for small the T, cases (although slightly high when $T = 2$), but it has extremely low coverage for the $N = 200,\ T = 10$ case. The PF-CRE estimator, on the other hand, has coverage probabilities that are somewhat below the nominal values in all cases. However, they are never substantially below the nominal values and, importantly, it is the only feasible estimator that can calculate partial effects that does not have a very poor coverage performance in any of the sample sizes considered in these simulations.

Table B2: *Coverage Probabilities, selected sample sizes*

	PF-CRE	Oracle	CRE	CMLE	FE
N = 1000, T = 2	90.00	94.40	98.00	98.00	96.00
N = 1000, T = 5	90.30	94.40	94.80	94.80	93.50
N = 100, T = 10	86.30	89.60	58.20	96.00	60.20
N = 100, T = 20	91.00	93.10	91.40	94.30	70.60

C Rare Events Simulations

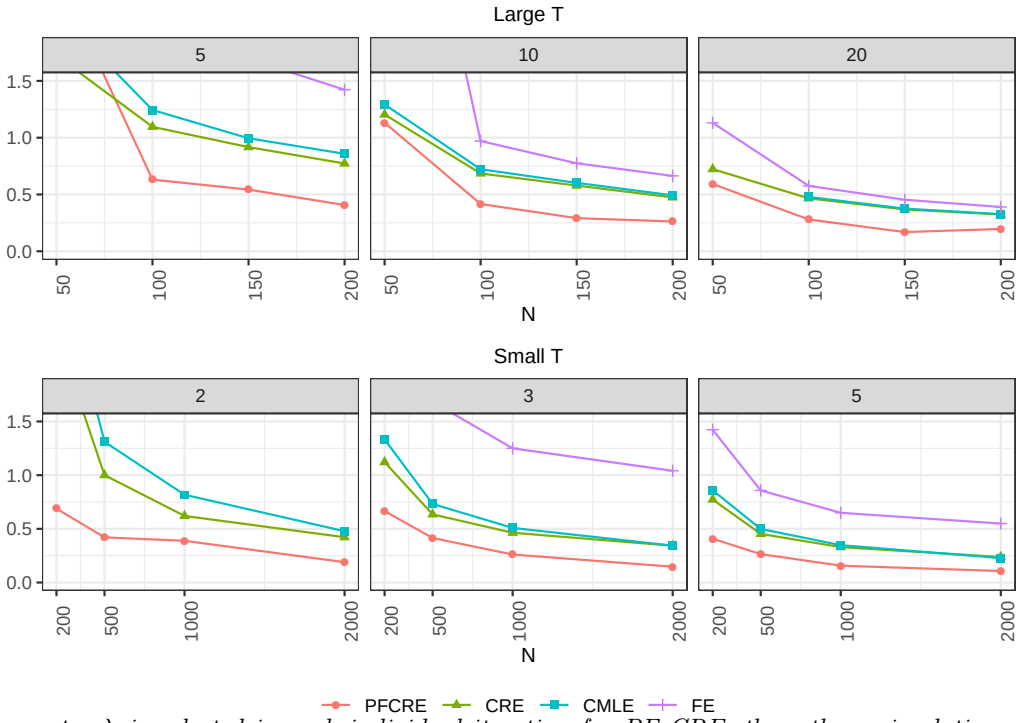
The estimation of models that account for unobserved heterogeneity with rare events data presents multiple challenges. The presence of many all-zero units in this context, means that estimators that do not restrict the unobserved heterogeneity in any way, like CMLE and FE, will effectively (or by design) remove all these units from the estimation of model parameters. This has two implications: first, it can lead to severe efficiency losses which results in typically useless estimates (see, e.g., Beck and Katz, 2001); second, it can introduce sample selection bias in the final estimates (see, e.g., Beck, 2020).

Because of these issues, some researchers choose to not control for unobserved heterogeneity, but this has can leave estimates exposed to bias. Researchers have instead advocated for the use of random effects or correlated random effects in the linear model context (Clark and Linzer, 2015; Bell and Jones, 2015). Crisman Cox (2019) advocates for CRE models in the context of binary outcome models. Cook et al. (2020) instead propose using a version of the Fixed-Effects estimator but imposing Jeffreys prior on the unit dummies, which effectively restricts the unobserved heterogeneity avoiding the effective loss of observations and issues derived from it.

Given that the presence of many all-zero units can potentially stress a computational algorithm (as it can create near singularities and other problems), I present simulations for rare events data. These simulations follow the same data generating process as those described by equations 15 and 16, with the only exception that the intercept α is set to -4. This more negative intercept together with the unobserved heterogeneity itself induces a very high number of all zero units. As comparison, I use the same estimators used in the body of the paper: CMLE , FE, and CRE. I do not include Cook et al. (2020)'s penalized fixed-effects model as Crisman Cox (2019) shows that CRE typically outperforms it and CRE is nested within PF-CRE.

These simulations, presented in Figure C1 show results that are largely the same as those presented in the main body of the paper. PF-CRE has the lowest Root Mean Squared Error, followed by CRE and, slightly above by CMLE. The FE estimator has the poorest performance of all. The only difference with the results from the main body of the paper is that the RMSE tends to be somewhat higher for all estimators (but particularly FE), which is expected since less information can be extracted from data with less time-variation in the outcome overall.

Figure C1: *RMSE for* β_1



The penalty parameter λ *is selected in each individual iteration for PF-CRE; thus, these simulations also incorporate uncertainty over the optimal penalty parameter.*

D Violation of Exchangeability

While a method should ideally not be used when its assumptions are violated, it is nonetheless useful to assess its performance when that is the case. A reasonable performance under violations of the underlying assumptions provides some robustness to the estimator and can increase confidence in its use. This is especially the case when the underlying assumptions are not always easily verifiable, or only partially verifiable with imperfect tools.

The main assumption in the PF-CRE estimator is assumption 1, which requires that the unobserved heterogeneity be exchangeable with respect to the time periods for the observed covariates. To see this, consider a simple case with only one covariate, x, and two time periods, $t = 1, 2$. Exchangeability requires that the distribution of the unobserved heterogeneity conditional on x_{i1} and x_{i2} be the same, even if we switch x_{i1} with x_{i2}; that is:

$$f(c_i|x_{i1}, x_{i2}) = f(c_i|x_{i2}, x_{i1}) \qquad \text{Exchangeable} \tag{D1}$$

$$f(c_i|x_{i1}, x_{i2}) \neq f(c_i|x_{i2}, x_{i1}) \qquad \text{Not Exchangeable} \tag{D2}$$

Here I present a set of simulations in which the exchangeability assumption is violated. I use the same data generating process presented in equations 15 and 16, with the following modification:

$$c_i = \begin{cases} z_1 & \text{if } x_{i11} > x_{i12} \\ -z_2 & \text{if } x_{i11} \leq x_{i12} \end{cases} \tag{D3}$$

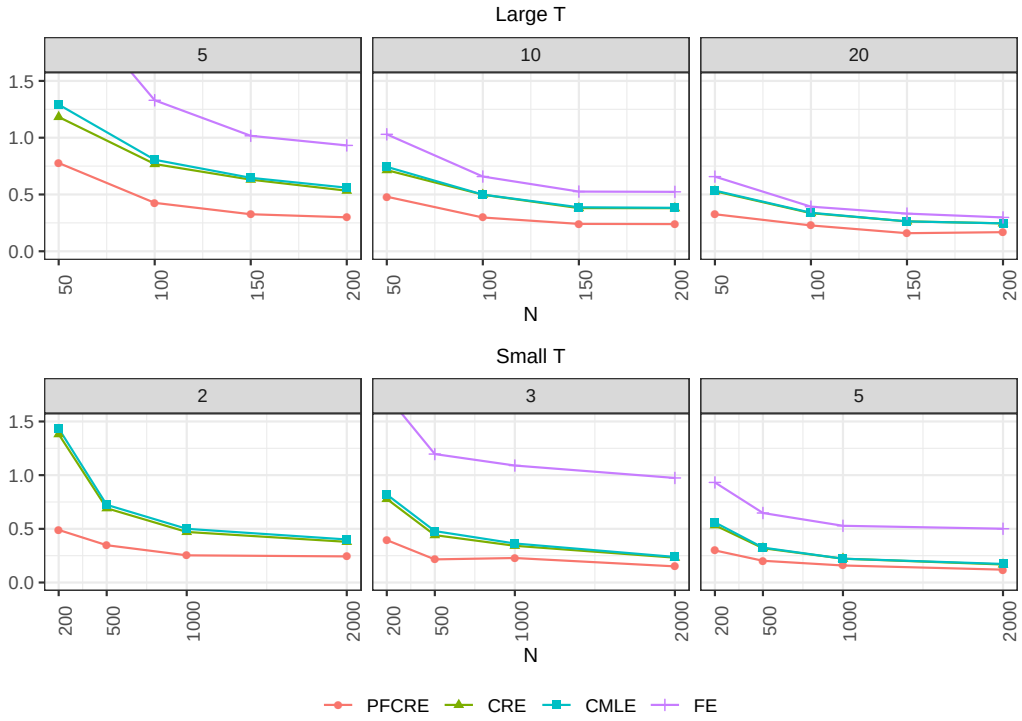
The heterogeneity described in D3 violates exchangeability because its distribution depends on whether the value of variable x_1 in the first time period, x_{i11} is larger than the value of variable x_1 in the second period, x_{i12}; thus, $f(c_i|\cdots)$ is not exchangeable with respect to variable x_1.

Figure D1 presents the results simulations for both large T and small T panels, with 250 simulations for each case.

The results from these simulations with a violation of the exchangeability assumption show that the PF-CRE estimator still outperforms the other estimators in terms of RMSE in both large and small T environments, although it does have a small bias relative to CMLE.

It is important to mention, however, that this is one possible violation of the exchangeability

Figure D1: *RMSE for β_1 with violation of exchangeability*

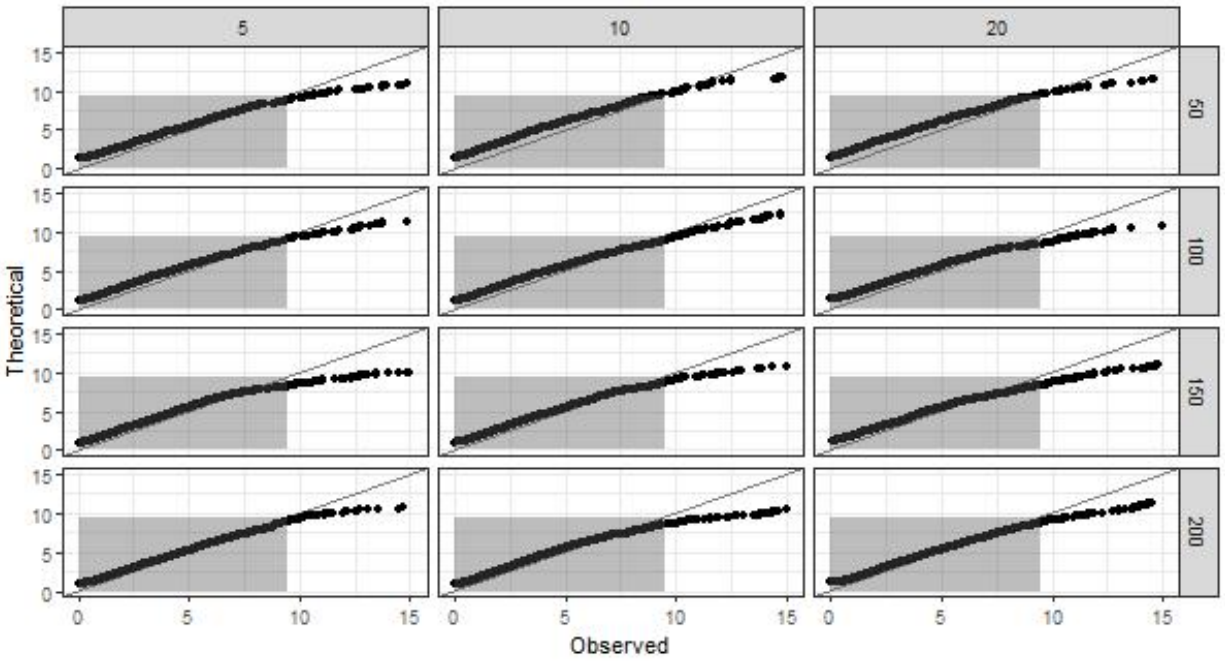


The penalty parameter λ is selected in each individual iteration for PF-CRE; thus, these simulations also incorporate uncertainty over the optimal penalty parameter.

assumption. It is possible that other violations could show different results, with other estimators having a lower RMSE than PF-CRE. In particular, it is possible that CRE could outperform PF-CRE in some potential violation of the exchangeability assumption. However, it is important to remember that if PF-CRE's assumptions are violated, so are CRE's assumptions, as CRE is nested in PF-CRE. In that sense, it is more likely than not, that when PF-CRE's performance is compromised because of the violation of it's assumptions, so will be CRE's performance.

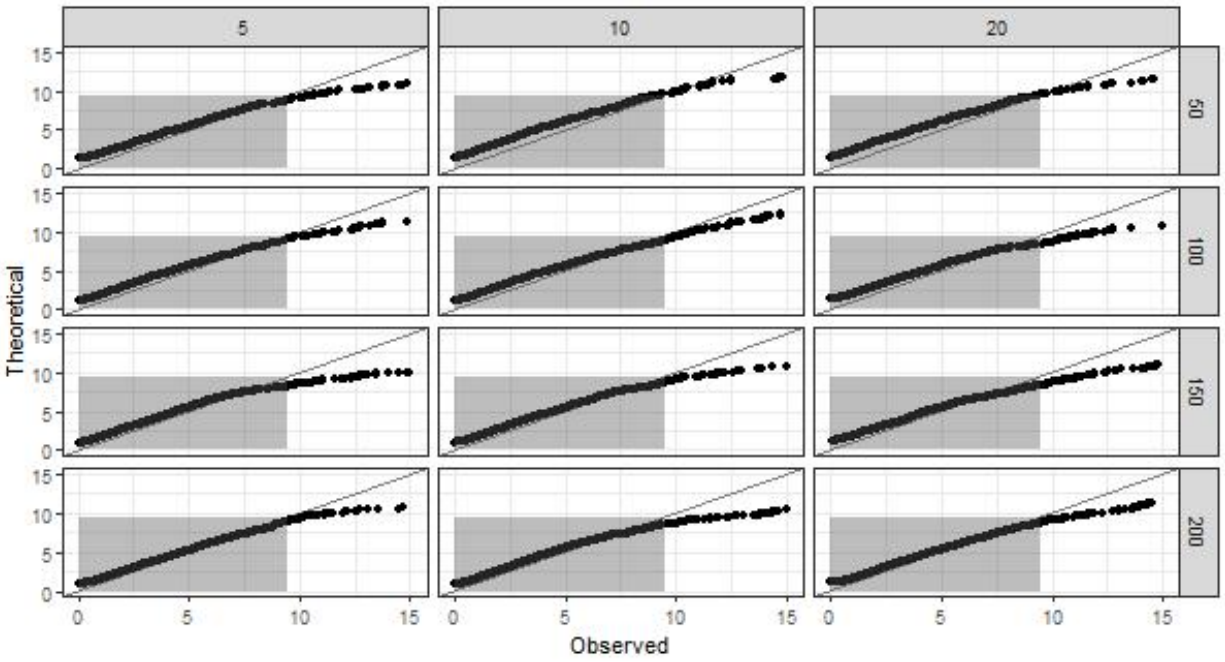
E Hausman Test for the Oracle Estimator

Figure E1: *Quantile-Quantile Plot for Specification Test: Large T*



Observed are the sample quantiles from the simulations. Theoretical are the theoretical quantiles from a $\chi^2_{(4)}$*. The shaded area represents the 95% theoretical quantile.*

Figure E2: *Quantile-Quantile Plot for Specification Test: Large T*



Observed are the sample quantiles from the simulations. Theoretical are the theoretical quantiles from a $\chi^2_{(4)}$*. The shaded area represents the 95% theoretical quantile.*

F Additional Tables from Applications

Table F1: *Coefficient Estimates, Tactical Voting 2015 U.K. Election*

	PF-CRE			CMLE			CRE			FE		
	β	Low	High	β	Low	High	β	Low	High	β	Low	High
Contact Preferred	-0.37	-0.60	-0.14	-0.32	-0.57	-0.08	-0.86	-1.12	-0.60	-0.50	-0.81	-0.19
Contact Viable	0.72	0.51	0.93	0.71	0.49	0.94	0.49	0.26	0.72	1.13	0.85	1.41
Therm. Preferred	-0.19	-0.27	-0.11	-0.18	-0.27	-0.10	-0.29	-0.37	-0.21	-0.29	-0.39	-0.18
Therm. Viable	0.33	0.26	0.40	0.31	0.24	0.38	0.25	0.18	0.32	0.49	0.40	0.58
n	3,824			3,824			3,824			3,824		
Effective n	3,824			1,164			3,824			1,164		
Observations	10,378			10,378			10,378			10,378		
Effective Obs.	10,378			3,263			10,378			3,263		
$\chi^2_{(4)}$	3.44			n/a			-129.42			n/a		
p-value	0.487			n/a			n/a			n/a		

All confidence intervals are at the 95% level. Low and High represent the upper and lower bounds of the confidence intervals. Logit standard errors are clustered at the individual level. The effective n and effective number of observations refers to the number of actual observations used in CMLE. There is no χ^2 test reported for CMLE since this estimator is the basis for that test, nor for FE. The p-value of CRE is not reported since it's test-statistic is negative.

Table F2: *Coefficient Estimates, Marinov (2005) Replication*

	PF-CRE			CMLE			CRE			FE		
	β	Low	High	β	Low	High	β	Low	High	β	Low	High
Sanctions	0.24	0.03	0.46	0.29	0.03	0.55	0.25	-0.01	0.52	0.30	0.03	0.56
Force	-0.24	-0.49	0.02	-0.27	-0.54	0.01	-0.26	-0.54	0.01	-0.28	-0.55	0.00
Econ. Growth	-1.50	-2.77	-0.23	-1.48	-2.79	-0.16	-1.50	-2.82	-0.18	-1.53	-2.87	-0.19
Wealth	-0.04	-0.15	0.08	-0.10	-0.32	0.13	-0.10	-0.32	0.13	-0.10	-0.33	0.13
Democracy	0.17	-0.22	0.56	0.17	-0.24	0.59	0.13	-0.29	0.54	0.18	-0.24	0.60
Dem * ln(t)	0.78	0.55	1.01	0.76	0.53	0.99	0.86	0.63	1.09	0.78	0.54	1.02
Mixed Regime	0.42	0.06	0.78	0.39	0.03	0.76	0.38	0.01	0.75	0.41	0.03	0.78
Mixed * ln(t)	0.37	0.16	0.57	0.38	0.17	0.60	0.40	0.19	0.60	0.40	0.18	0.61
Age	0.02	0.01	0.03	0.02	0.01	0.03	0.02	0.01	0.03	0.02	0.01	0.03
Years in Office	-1.00	-1.24	-0.77	-0.97	-1.20	-0.73	-1.10	-1.33	-0.87	-1.00	-1.24	-0.76
Spline 1	-0.19	-0.23	-0.14	-0.18	-0.22	-0.13	-0.19	-0.23	-0.14	-0.18	-0.23	-0.14
Spline 2	0.07	0.05	0.09	0.07	0.05	0.09	0.07	0.05	0.09	0.07	0.05	0.09
Spline 3	-0.01	-0.02	-0.01	-0.01	-0.02	-0.01	-0.01	-0.02	-0.01	-0.01	-0.02	-0.01
n	160			160			160			160		
Effective n	160			136			160			136		
Observations	5,766			5,766			5,766			5,766		
Effective Obs.	5,766			5,295			5,766			5,295		
χ^2	1.11			n/a			-2.31			n/a		
p-value	0.95			n/a			n/a			n/a		

All confidence intervals are at the 95% level. Low and High represent the upper and lower bounds of the confidence intervals. Logit standard errors are clustered at the individual level. The effective n and effective number of observations refers to the number of actual observations used in CMLE. There is no χ^2 test reported for CMLE since this estimator is the basis for that test, nor for FE. The p-value of CRE is not reported since it's test-statistic is negative.