Partial Effects for Binary Outcome Models with Unobserved Heterogeneity

Lucas Núñez *

California Institute of Technology

September 18, 2018

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 small samples. To demonstrate that accounting for unobserved heterogeneity has important consequences for empirical analysis, I use PF-CRE in three studies of voting behavior: tactical voting during the 2015 British Election, support for the Brexit referendum, and vote choice in the 2012 U.S. Presidential election. In all three cases, I find that ignoring the unobserved heterogeneity leads to an overestimation of the effects of interest, and that PF-CRE is a valid approach for the analyses.

*Ph.D. Candidate, Division of the Humanities and Social Sciences, California Institute of Technology. Email: `lnunez@caltech.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. In studies of political behavior, this heterogeneity sometimes takes the form of voters' core beliefs, which are hard to define, let alone to measure. It can also take more mundane forms. For example, researchers rarely get to observe how political parties choose which voters to contact during electoral campaigns. 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. Returning to the example, if a party contacts those voters who are already likely to support it (in a way that researchers do not observe), then the effect of party contact on the probability of voting for that party will be overestimated if researchers do not account for the unobserved heterogeneity in some way.

There are three main 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 semiparametric techniques (e.g., Abrevaya, 2000); or use random or correlated random effects (Mundlak, 1978; Chamberlain, 1980). 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.[1]

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. The PF-CRE estimator, explicitly accounts for the correlation between the

[1]Making restrictive assumptions about the individual heterogeneity also leads to biased estimates. I distinguish the bias and inconsistency that arise from unrealistic assumptions from the one that arises from the estimation procedure itself.

observed and unobserved components of the model, using a large flexible specification (more details below). Moreover, it includes a penalization step for variable selection to induce efficiency. This estimator addresses the three problems described above: it provides consistent estimates for the model parameters, allows for the estimation of partial effects, and makes mild assumptions about the unobserved heterogeneity.

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. This flexible specification is composed of functions of the observed covariates (such as individual time-means and other exchangeable functions[2]), 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. When the number of covariates is small, this does not pose a major hurdle. However, the number of parameters grows exponentially with the number of covariates in the model. For example, with 3 observed covariates, a relatively simple specification that models the unobserved heterogeneity on the time-means of the covariates with up to three-way interactions requires the estimation of 25 parameters, which is manageable; with 5 covariates, 63 parameters; with 10 covariates, 298 parameters.[3] Moreover, if the specification also includes additional time-invariant characteristics, the number of coefficients in the model can become unmanageable very fast.

To address the dimensionality issue, the estimation is performed via *penalized* Maximum Likelihood using the Smoothly Clipped Absolute Deviation (SCAD) penalty. Importantly, the penalization is only applied to the terms that model the unobserved

[2]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.

[3]With three covariates there are 3 coefficients associated with the covariates, a constant term, three asssociated with the time-means, 6 for two-way interactions, 10 for three-way interactions, and the variance of the random effect.

heterogeneity, but not to the covariates of interest. Like other penalized estimation methods, SCAD introduces a cost in the likelihood function for the size of each parameter to be estimated. Therefore, 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. 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 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 for researchers to decide when it is correct to use it.

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, and that it performs better than alternative estimators. In addition, the penalization step is the key for reducing uncertainty around the estimates. For the logistic case, the simulations show that the rejection rate of the specification test is close to theoretical levels.

To illustrate the performance of PF-CRE in a real-data environment, I provide an application to tactical voting during the 2015 United Kingdom General Election. The outcome of interest is whether a voter intends to cast a tactical vote, that is, vote for a party that is not her most preferred one. I use three waves of the British Election Study Online Panel. The effects of interest are the extent to which parties can influence the probability of a tactical vote through campaign contacts to voters. The unobserved heterogeneity in this application represents all the information that parties know about voters' that outsiders (the researcher) do not know. In particular, parties may know which voters may consider casting a tactical vote and be more likely to contact them. The specification test shows that PF-CRE's assumptions hold in this case. The results show that ignoring the unobserved heterogeneity leads to an overestimation of the effects of party contacts during the campaign on the probability that a voter casts a tactical vote.

I also provide two additional applications that show that the assumptions of PF-CRE hold in other political science applications. In particular, I show that PF-CRE provides consistent and efficient estimates of (1) the effect of preferences for immigration and economic fears on voting for the 2016 Brexit Referendum in the U.K.; and (2) the effect of ideological preferences and candidate characteristics on vote choice during the 2012 U.S. Presidential Election. In both these cases, ignoring the unobserved heterogeneity leads to significant differences in the estimated partial effects of the covariates of interest and to our understanding of voter behavior.

2 Penalized Flexible Correlated Random Effects

In this section I first provide a short introduction to binary outcome models with unobserved heterogeneity and define the quantities of interest. Second, I present the identification strategy, estimation, and asymptotic properties of PF-CRE.

2.1 Binary Outcome Models with Unobserved Heterogeneity

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.

The focus in this paper is on balanced panels. However, unbalanced panels are not uncommon. If the type of selection that generates the unbalancedness is completely at random, then the estimator proposed here can be applied without modification. If the selection is instead correlated with the covariates and the unobserved heterogeneity, then the same modification proposed by Wooldridge (2010a) for Correlated Random Effects models applies for PF-CRE, as long as selection is ignorable.[4]

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.[5]

In some applications researchers may only be interested in the sign and relative sizes of the β coefficients. In many others, however, the interest lies in the partial effects that reflect how the probability of the outcome changes with respect to a change in the covariates x. 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

[4]Selection is ignorable if it is independent of the error term, conditional on the covariates and the unobserved heterogeneity.

[5]The independence assumption is not a necessary condition for PF-CRE, as along with other CRE-type models, it is robust to violations of conditional independence (see, Wooldridge, 2010a, for CRE's robustness). However, this assumption is maintained throughout the paper as it is necessary for CMLE, which is the basis of comparison in the specification test presented in Section 4.

[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.

2 are defined as:

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

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

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

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

2.2 Assumptions for Identification and Estimation

The identification challenge in the model of equation 2 lies in c_i being unobserved *and* correlated with x_{it}.[8] The identification strategy is to specify a distribution for c_i conditional on $(x_{i1}, ..., x_{iT})$ without imposing excessively strong restrictions on the unobserved heterogeneity. I begin with the following assumption:

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 $f(c_i|\cdot)$. Returning to the example of party contacts, this assumption requires that what matters for the conditional distribution of the unobserved heterogeneity is, for example, how many times a voter was contacted, but not exactly when she was contacted.

Under Assumption 1, without loss of generality, $f(c_i|x_{i1}, ..., x_{iT})$ can be written as a

[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 known as a random effect. This case does not pose significant challenges to traditional estimators. However, the PF-CRE approach is also valid.

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.

In most circumstances, researchers also observe time-invariant information, w_i, about each individual i, such as 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.

Assumption 1 alone is not sufficient for identification. The reason is that the first T non-central moments characterize the T observations per individual i, exhausting the degrees of freedom. Therefore, additional restrictions are necessary for identification:

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.

Assumption 2 restricts 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 actually represents a polynomial). This implies a stronger condition than exchangeability alone, but it maintains sufficient flexibility to capture (or approximate) the conditional distribution of the unobserved heterogeneity.

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

In order to obtain parametric identification, it is necessary to specify a distribution

[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.

for the unobserved heterogeneity, hence Assumption 3. However, other distributions are possible, as long as they have finite moments.[10]

Combining assumptions 1, 2, and 3, the unobserved heterogeneity and its density function can be written as:

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

2.3 Estimation

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

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

where z_i is a vector of moments of $(x_{i1}, ..., x_{iT})$, observed time-invariant characteristics, and interaction terms among these; and η_i is a normally distributed random effect with variance σ^2 that is independent of the covariates of the model.[11]

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

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

with

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

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

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

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

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. However, the dimensionality of γ increases exponentially with the number of covariates in the model. With three covariates, a simple specification of z_i that includes the time-means of the covariates and up to three-way interactions requires the estimation of 20 additional parameters.[12] The same type of specification with 5 covariates requires the estimation of 56 additional parameters with 10 covariates, 286 parameters. Moreover, the inclusion of time-invariant characteristics exacerbates this problem. However, the assumptions establish that the polynomial $z_i\gamma$ is sufficient to capture the unobserved heterogeneity, but do not establish 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 and simplify the 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. Then, β is estimated 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 Clipped Absolute Deviation (SCAD) penalty (Fan

[12] 3 time-means, 6 two-way interactions, 10 three-way interactions, and the variance of the random effect.

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 does not affect the coefficients of interest, β, directly since they are left unpenalized.[13] 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 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.[14]

2.4 Asymptotic Properties

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,*

$$\sqrt{nT}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, I(\beta)^{-1}), \tag{11}$$

where $I(\beta)$ is the Fisher information matrix.

[13] The parameter a in the SCAD penalty is usually set to $a = 3.7$ (Fan and Li, 2001). The parameter λ is chosen via cross-validation.

[14] 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 small sample performance.

Theorem 1 follows from standard properties of Maximum Likelihood estimation and the Oracle property of the SCAD penalty. The Oracle property of SCAD establishes that the penalized estimator has the same asymptotic distribution as the underlying (and unknown) Data Generating Process (DGP).[15] Consequently, it has the same asymptotic properties of the Maximum Likelihood estimator of the DGP. Consistency, efficiency, and normality of the PF-CRE estimator thus follow from the properties of Maximum Likelihood estimators.[16,17]

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:*

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

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.[18] 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

[15]See, Ibrahim et al. (2011); Hui et al. (2017)

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

[17]In principle, uncertainty over the penalty parameter λ, selected via cross-validation, generates more uncertainty over the parameters of interest β. This uncertainty is not included in equation 11. However, in Section 5.1, I show that uncertainty over λ does not, in practice, generate a significant amount of uncertainty over β.

[18]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.

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). For example, suppose x represents individuals' ideology, and z corresponds to the average ideology of each individual across panel waves. If we want to estimate the effect of changing x from liberal to very liberal, then the value of z should also correspond to a liberal (or very liberal) individual. Although using a value of z corresponding to a very conservative individual is technically correct, inferences in this case will rely heavily on extrapolation from the model.[19]

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. I briefly discuss each of them and how they relate to the PF-CRE estimator.[20]

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 asymptotic 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.[21,22]

[19]This is because individuals who report being a liberal in a wave, but have generally reported to be very conservative in other waves, are rare or non-existent.

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

[21]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.

[22]The asymptitic 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.

In light of the inconsistency of the FE estimator, bias correction procedures have been proposed.[23] These corrections reduce the bias; however, they do not eliminate it.[24] 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.[25] 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.

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.[26]

The CMLE approach has two main shortcomings. First, it does not provide estimates of the partial effects.[27] This is because location parameters are not estimated; in fact, β is estimated by eliminating the location parameters c_i and α from the likelihood function. 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

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

[24]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.

[25]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.

[26]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, see Manski (1987); Abrevaya (2000).

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

of information, and consequently larger standard errors in the estimates.[28,29]

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.[30]

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 severely restricting the unobserved heterogeneity with ad-hoc specifications. 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.

The PF-CRE estimator 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

[28]Note that CMLE's conditioning on those individuals with variation in the outcome can also introduce errors if this subpopulation behaves differently than the overall population, beyond the unobserved heterogeneity. However, an implicit assumption in this paper, for all estimators, is that despite the presence of unobserved heterogeneity, individuals' behavioral rules are the same. That is, they all have the same β.

[29]Note that the FE approach results in the same kind of information loss without 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 parameter β (see, for example, Beck and Katz, 2001).

[30]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.

suffer from the incidental parameters problem. It also allows for the estimation of probabilities and partial effects, which cannot be done with CMLE. Finally, 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. PF-CRE, on the other hand, selects the minimal specification for this correlation that is necessary to control for the unobserved heterogeneity, leading to efficiency gains. In other words, FE and CMLE assume there is no information in cross-sectional variation. 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).

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 $z\gamma$ terms. This does not necessarily hold in every application. Therefore, I present a model specification test for one of the most commonly used models in applied research: the logistic case.

If the correlation between the observed and unobserved components of the model can be correctly captured by the $z\gamma$ terms, then the PF-CRE estimator developed in this paper is both consistent and efficient. The Oracle property of the penalized estimator plays a crucial role here, as it ensures that the penalized model asymptotically attains the same information bound as the Oracle estimator, which is 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.[31] Following Hausman (1978), I construct an specification test based on the standardized squared difference between these two estimators. That is, the test statistic is defined as:

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

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:[32]

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

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 that the PF-CRE estimator of β is consistent and efficient.

[31] 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.

[32] 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.

5 Simulations

I conduct three sets of simulation studies to analyze the performance of the PF-CRE estimator in small samples and compare it to that of alternative methods. I use the Oracle estimator as a benchmark for comparison. The Oracle estimator is the traditional (non-penalized) Maximum Likelihood estimate that uses the exact specification of the data generating process (which with real data is unknown). In the first set of simulations I analyze the PF-CRE and CMLE estimates of β relative to the Oracle. In the second set, I compare the estimates of the Partial Effects (PEs) from PF-CRE, the traditional CRE specification from Mundlak (1978), an *un*penalized version of PF-CRE, denoted by UF-CRE, and a pooled logit that ignores the unobserved heterogeneity.[33] I have excluded the Fixed Effects estimator and its bias-corrected alternatives because they are well known to have a significant bias for small-T studies and standard errors in the order of those for CMLE (see Section 3). The final set of simulations studies the specification test for the PF-CRE for different sample sizes.

The data generating process in all simulations is given by:

$$\begin{aligned} Prob(y_{it} = 1 | x_{it}, c_i) = \Lambda(\alpha + x_{it}\beta + c_i),\ with\ x_{it} \in \mathbb{R}^5, \\ \beta = (0.7, 1.3, -0.4, 1.2, -0.2),\ \alpha = 0.2, \\ c_i | \mathbf{x}_i \sim \mathcal{N}(\mu_i, \sigma_c^2), \end{aligned} \tag{15}$$

where $\Lambda(\cdot)$ is the logistic cumulative distribution, $x_{it} \sim \mathcal{N}(0, I_5)$.[34] For each set of simulations I use three different correlation forms for the unobserved heterogeneity:

- *Sparse Specification:*

  $\mu_i = 0.5\overline{x_{i1}} + 0.6\overline{x_{i2}} + 1.2\overline{x_{i1}x_{12}}$

- *Random Effect Specification:*

[33] Mundlak (1978)'s specification of CRE uses the time-means of the covariates to model the unobserved heterogeneity. The UF-CRE uses the same specification as PF-CRE but without the penalized estimation step.

[34] The choice of values for β and α have no particular meaning.

$\mu_i = 0$

- *Complex Specification:*

$$\mu_i = \overline{x_{i1}} + \overline{x_{i2}} + \overline{x_{i3}} + \overline{x_{i4}} + \overline{x_{i5}} + \overline{x_{i1}x_{i2}} + \overline{x_{i1}x_{i3}} + \overline{x_{i2}x_{i3}} + \overline{x_{i2}x_{i4}} + \overline{x_{i3}x_{i4}} + \overline{x_{i3}x_{i5}} + \overline{x_{i4}x_{i5}}$$

where $\overline{x_{ij}}$ denotes the time-mean of x_{itj}, where j denotes the jth variable in x_{it}.

These three data generating processes differ only in terms of the unobserved heterogeneity, all other parameters have the same values across the three of them. This allows for an analysis of PF-CRE under different types of unobserved heterogeneity, holding everything else constant. In the sparse specification, the unobserved heterogeneity is relatively simple and represents one of the best case scenarios for PF-CRE. The sparsity can lead to significant efficiency gains relative to CMLE and helps illustrate the gains from the penalization step. Moreover, because of the inclusion of an interaction term, the traditional CRE approach should be biased. The random effect specification is the simplest form of unobserved heterogeneity possible and PF-CRE should have important efficiency gains relative to CMLE as well. Moreover, this case is included as under this specification, logit can recover the correct partial effects (but not parameters) and the traditional CRE estimator can correctly recover both partial effects and parameters β. Finally the complex specification is useful to study the performance of PF-CRE in a case in which the efficiency gains from the penalization step are significantly reduced.

For all simulations $T = 2$. For the first two sets, n is 1,500, whereas for the specification test simulations I use an n size of 1,000, 2,000, 3,000, and 4,000. All results are based on 1,000 draws from the corresponding data generating process, which is a sufficient number so that the tails of the distributions are appropriately accounted for. In all cases, the penalty parameter λ is selected via 5-fold cross-validation using the Akaike Information criterion (AIC).[35]

[35] As the simulations below show, cross-validation using AIC allows PF-CRE to attain almost the same distribution as the Oracle estimator. Other measures of fit for cross-validation can be used and may have a good performance. However, the margin for improvement relative to AIC is limited.

5.1 Parameter Simulations

The DGP in equation 15 satisfies the assumptions of both the CMLE and PF-CRE estimators, and therefore the estimates of β from both of them are consistent. Table 1 shows the Root Mean Squared Error (RMSE) of the CMLE and PF-CRE estimates of β relative to the RMSE of the Oracle estimator. Because both estimators are consistent, the differences in the relative RMSEs mainly come from the variance of the estimators (see Table A2 in the appendix).[36] As expected, given that the heterogeneity in the DGP is not completely unrestricted, the CMLE estimator produces less efficient estimates than the PF-CRE approach. In fact, the CMLE approach produces RMSEs that are 30% to almost 90% higher than those of the Oracle, depending on the specification of the unobserved heterogeneity. The RMSEs of the PF-CRE approach deviate by at most 3% from those of the Oracle. This illustrates the efficiency gains of this estimator relative to the CMLE estimator, as well as the Oracle properties of PF-CRE (figures A1-A3 show the full distribution of the simulations).

Table 1: *$\widehat{\beta}$ RMSE relative to RMSE of the Oracle Estimator*

	Sparse		RE		Complex	
	CMLE	PF-CRE	CMLE	PF-CRE	CMLE	PF-CRE
β_1	1.32	1.03	1.72	1.00	1.41	1.00
β_2	1.50	1.00	1.70	1.02	1.53	1.00
β_3	1.68	1.00	1.74	1.01	1.35	1.00
β_4	1.87	1.00	1.78	1.02	1.53	1.00
β_5	1.74	1.00	1.81	1.00	1.29	1.00

A value of 1 indicates identical RMSE to the Oracle estimator. Larger (smaller) values indicate a larger (smaller) RMSE than the Oracle's

It is important to note that for the more complex model the efficiency gains of PF-CRE relative to CMLE are smaller relative to the other specifications. This is to be expected. The more complex the unobserved heterogeneity, the less information there is in cross-sectional variation. Therefore, an estimator like CMLE that discards

[36] Both estimators have a small bias in small samples. The simulations show that this bias is typically smaller for the PF-CRE than the CMLE estimator. See Table A1 in the appendix.

cross-sectional variation will have a smaller efficiency loss than in simpler specifications.

Finally, the use of cross-validation to select the penalty parameter λ can generate increase uncertainty (which is not accounted for in the standard error formula in Theorem 1). The simulations presented here strongly suggest that uncertainty over λ does not translate into significant uncertainty over β, as PF-CRE's RMSE and that of the Oracle's are almost identical (the same can be seen in Table A2 for the standard errors). Intuitively, this lack of importance derives from the fact that λ affects β only through the polynomial terms being selected, as β is not itself penalized. The uncertainty over λ could sometimes lead to the inclusion or exclusion of polynomial terms which, by definition, only marginally affect the likelihood, therefore not affecting β significantly.

5.2 Partial Effects Simulations

Here I compare the Partial Effects for the DGP in equation 15 estimated via the PF-CRE approach, the traditional CRE approach, the UF-CRE (i.e., the unpenalized version of PF-CRE), and a pooled logit model.

Table 2 shows the RMSE of the four estimators relative to that of the Oracle estimator for the 5 covariates in the model. Partial effects are calculated for the mean value of the covariates. The RMSE of the PF-CRE approach is the lowest, and is at most 3% deviated from that of the Oracle's. The traditional CRE approach, in turn, produces estimates with RMSEs that can be more than 550% higher than the Oracle's. This is because the CRE approach includes terms that do not belong in the data generating process for the unobserved heterogeneity (as in the Sparse and RE specifications), while it excludes terms that do belong there (as in the Sparse and Complex specifications). This leads to both inconsistent and inefficient estimates. The UF-CRE approach also produces estimates with a RMSE that can be 80% higher than the Oracle's. This reflects the inefficiency of the unpenalized approach, as it includes many more parameters than there are in the DGP in all three specifications. However, the efficiency loss is smaller for the Complex specification, as this specification contains more terms. Finally, the pooled logit approach, which ignores the unobserved heterogeneity produces

RMSEs that can be 500% higher than the Oracle's. This high RMSE is a consequence of the logit approach completely ignoring the unobserved heterogeneity. Most of the error in this case comes from the bias of the logit approach (see Tables A3 and A4 for the bias and standard deviations of the estimators). The only case in which the pooled logit performs well is for random effects. This is expected, as the unobserved heterogeneity in this case is independent of x.

Table 2: *$\widehat{PE}$ RMSE relative to RMSE of the Oracle Estimator*

	Sparse				RE				Complex			
	PF-CRE	Logit	UF-CRE	CRE	PF-CRE	Logit	UF-CRE	CRE	PF-CRE	Logit	UF-CRE	CRE
β_1	1.02	2.40	1.08	3.43	1.00	0.99	1.53	2.37	1.00	2.29	1.17	3.76
β_2	1.00	5.05	1.23	5.57	1.00	0.97	1.83	3.37	1.00	1.66	1.29	4.91
β_3	1.02	1.23	1.39	2.93	1.00	1.00	1.48	1.86	1.00	5.77	1.07	2.42
β_4	1.00	1.86	1.50	6.46	1.00	0.97	1.68	3.04	1.00	1.77	1.28	4.88
β_5	1.03	1.02	1.38	1.90	1.00	1.00	1.38	1.48	1.00	5.41	1.02	1.48

A value of 1 indicates identical RMSE to the Oracle estimator. Larger (smaller) values indicate a larger (smaller) RMSE than the Oracle's

5.3 Specification Test Simulations

Using the same setting as for the previous simulations, I calculate the rejection rate of the model specification test in equation 14 for four different sample sizes (1,000, 2,000, 3,000, and 4,000) at the 90% and 95% level. For each sample size, I draw 1,000 samples of the data generating process for the Sparse and Complex Specifications.[37] In both specifications, the null hypothesis that PF-CRE is consistent and more efficient than CMLE is true. Table 3 shows that the size of the test is close to theoretical expectations; that is, it rejects about 5% at the 95% level and about 10% at the 90% level, for all sample sizes.

For the Complex Specification of the unobserved heterogeneity, the specification test tends to over-reject the null hypothesis that the PF-CRE estimator is efficient and

[37]The main point of interest in this section is to study the performance of the specification test for different complexities of the unobserved heterogeneity. The Random Effect specification is excluded from this simulations as the performance of the specification test can be sufficiently inferred from the other two specifications.

consistent. This implies that the test will provide a conservative recommendation when the unobserved heterogeneity is more complex. However, this over-rejection disappears with larger sample sizes.

Table 3: *Simulations: Specification Test*

	Rejection Rate			
	Sparse		Complex	
n	10%	5%	10%	5%
1,000	0.096	0.053	0.133	0.087
2,000	0.094	0.049	0.113	0.065
3,000	0.097	0.048	0.107	0.059
4,000	0.103	0.050	0.106	0.056

Rejection rate calculated as the percentage of p-values smaller than 5% or 10% from 1,000 simulations for each sample size.

Figures A7 and A8 in the appendix show quantile-quantile plots, where the horizontal axis represents the quantiles from the simulations, and the vertical axis the quantiles from the theoretical distribution of the test (in this case, a $\chi^2_{(5)}$). The quantile-quantile plots for the Sparse Specification test show that the empirical quantiles of the test statistic are similar to their theoretical counterparts.[38] In the case of the Complex Specification, the plots show that the specification test tends to generate larger statistics than it should, but that this tendency diminishes and disappears for larger samples sizes.

Overall, the simulations show that the asymptotic properties of PF-CRE travel well to small samples. The PF-CRE estimator produces estimates of the model parameters that are more efficient than those of the CMLE estimator when the data generating process for the unobserved heterogeneity satisfies the assumptions of PF-CRE. In addition, the simulations also illustrate the advantages of the PF-CRE estimator in the estimation of partial effects. They show that the flexibility of its specification gives it a significant advantage over the traditional correlated random effects, and that the

[38]Deviations for the larger values are expected as many more simulations would be necessary for an accurate representation of the tail of the distribution, as larger values occur with very small probability

penalization step can help to significantly reduce the uncertainty around the estimated quantities. Finally, the simulations show that the specification test has rejection rates that are close to theoretical levels, when sample sizes are large enough.

6 Application: Tactical Voting in the 2015 U.K. General Election

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).[39]

The literature on tactical voting has generally focused on measuring its extent, but less on why some voters behave tactically while others do not. In this application I focus on the effect that being contacted by political parties has on voters' propensity to cast a tactical vote. The empirical challenge lies in correctly identifying the effect of party contact itself, independent of the effect of unobserved confounders. In particular, parties possibly contact the voters that they believe are more likely to respond to the parties' message or appeals. However, researchers do not observe how parties decide which voters to contact. Thus, from the researchers' point of view, this constitutes unobserved heterogeneity in voters' behavior that is also correlated with the observed covariates (in this case, being contacted by a party).[40]

To address this challenge, I use a panel data survey collected prior to the 2015 United Kingdom General Election. Controlling for the unobserved heterogeneity using PF-CRE allows to reduce or eliminate the concerns outlined in the previous paragraph. In particular, the unobserved heterogeneity modeled by PF-CRE captures voters' overall

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

[40] Ideally, disentangling the effects of party contacts from the fact that parties parties choose whom to contact can be done by relying on field experiments, in the spirit of Gerber et al. (2008) for voter turnout. However, while an experimental intervention in a real election aimed at increasing voter turnout may be relatively uncontroversial, one aimed at altering voters' choices faces significant ethical dilemmas.

characteristics and tendencies, which will reflect the fact that parties choose to contact some voters but not others.

6.1 Data and Model Specification

To study the effect of party contact on the probability of casting a tactical vote I use data from three waves of the British Election Study Online Panel. These data were collected prior to the 2015 United Kingdom General Election.[41] 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. 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.[42] 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. Finally, I include a number of time-invariant characteristics that serve as control variables in pooled logit estimates and also as additional terms to model the conditional distribution of the unobserved heterogeneity in the PF-CRE estimator. Among these, I include employment status, retirement status, student status, education

[41]The study covers England, Scotland, and Wales, but excludes Northern Ireland because of its different party system.

[42]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.

level, gender, ethnicity, age, and home ownership.

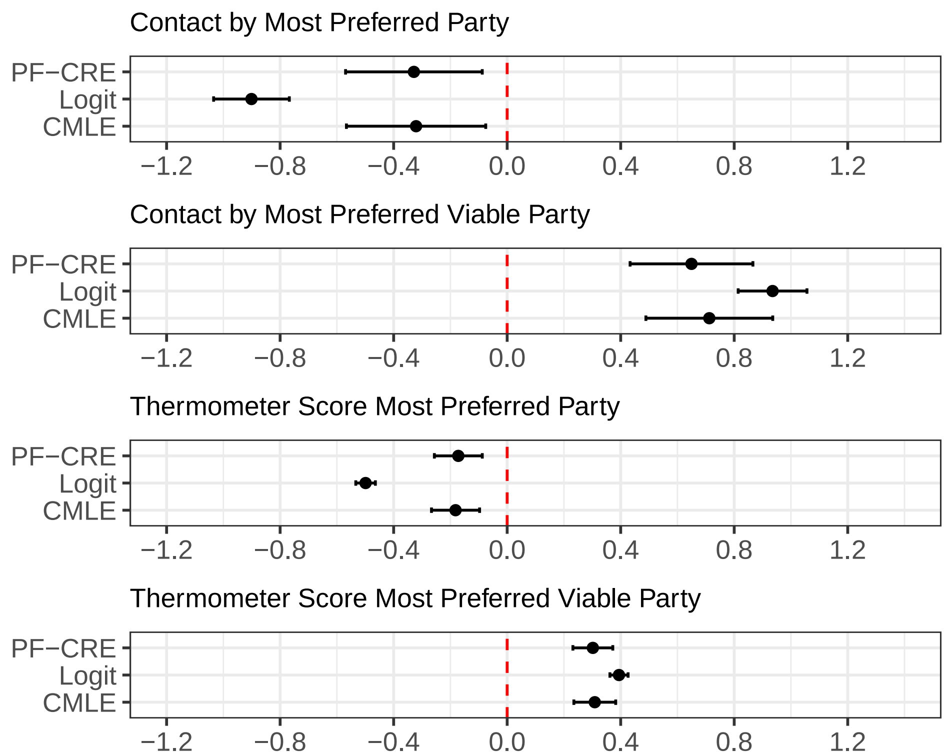
To model the correlation between the unobserved heterogeneity and the covariates of interest in the PF-CRE estimator, I use the time-means of the covariates of interest, plus the time-invariant characteristics, and two-way interactions among them, for a total of 230 terms. Given that I use the logistic distribution in this application, I compare the coefficient estimates from the PF-CRE estimator with those of the Conditional Maximum Likelihood estimator (CMLE). While both PF-CRE and CMLE account for unobserved heterogeneity, only PF-CRE allows for the estimation of partial effects. Additionally, I estimate a pooled logit that includes the time-invariant characteristics as controls.[43] Despite the inclusion of additional controls, the logit model does not account for the unobserved heterogeneity. I compare coefficient and partial effect estimates from the pooled logit and PF-CRE estimator to show the discrepancies that arise from ignoring the unobserved heterogeneity.

6.2 Results

Figure 1 shows that the coefficient estimates of PF-CRE and CMLE are very similar to one another.[44] Indeed, the specification test does not reject the null hypothesis that PF-CRE is consistent and more efficient than CMLE, with a p-value of 0.29. This clearly establishes the validity of the PF-CRE approach in this case. Importantly, PF-CRE allows me to estimate partial effects that CMLE cannot estimate. It is also clear from Figure 1 that the pooled logit model overestimates the effects of being contacted by the most preferred and most preferred viable parties on the decision to cast a tactical vote. Estimates for the thermometer scores also show overestimation by the logistic model.

[43] I do not include a traditional CRE estimator here because the CRE estimator is nested in PF-CRE.
[44] See Table B1 in the appendix for details with the estimates from the three models.

Figure 1: *Coefficient Estimates, Tactical Voting 2015 U.K. Election*

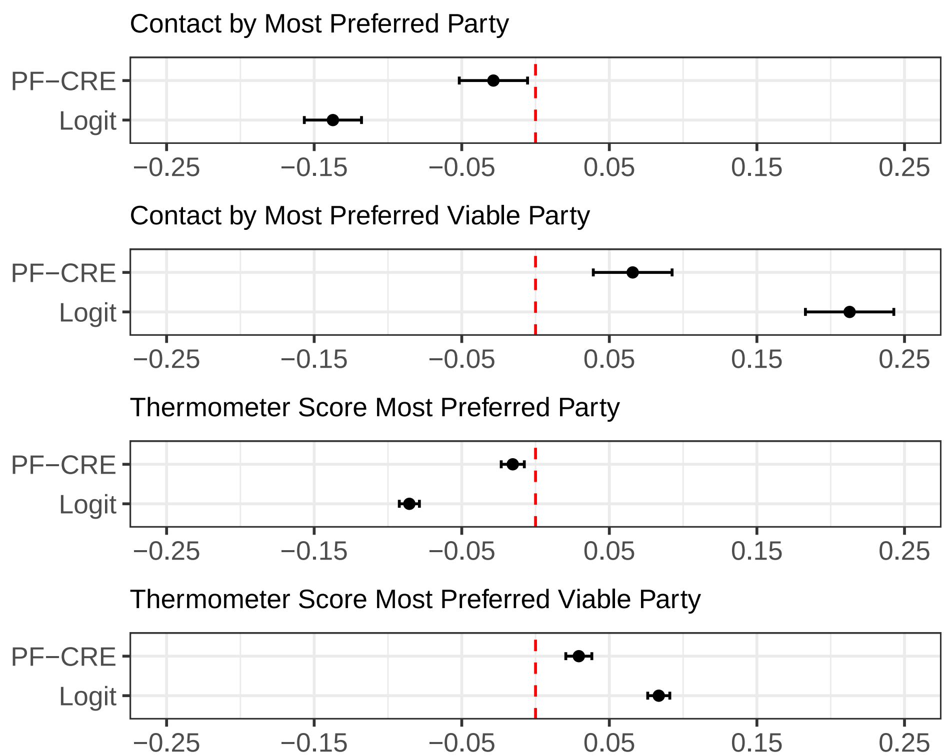


The tuning parameter for the SCAD penalty in PF-CRE was obtained through 10-fold cross validation using the Akaike information criterion. Logit standard errors are clustered by respondent.

Why does pooled logit overestimate the effects of party contacts? In principle, unobserved heterogeneity is in fact unobserved, and researchers can only speculate as to its sources. In the case of party contacts, it is possible that candidates (and their campaigns) from viable parties in a given constituency tend to contact supporters of non-viable parties that they believe are likely to defect their preferred party and vote tactically. At the same time, candidates from non-viable parties may be more likely to contact potential defectors from among their supporters as a way to prevent their numbers from dropping. This implies that the voters that parties contact are those who are more likely to cast a tactical vote in the first place. Therefore, when ignoring the heterogeneity (like the pooled logit does) the coefficient estimates for party contact capture both the effects of contact itself plus the selection effects just described.

Accounting for the unobserved heterogeneity, as the PF-CRE estimator does, controls for the selection effect introduced by the way parties choose voters for contact. This reduces or eliminates the bias introduced by this selection effect, as it captures voters' overall characteristics, which are likely related to how parties decide which voters to contact.

Figure 2: *Partial Effects, Tactical Voting 2015 U.K. Election*



Partial effects are calculated for a baseline individual. Baseline values for the conditional mean equation in PF-CRE were chosen to be consistent with those of the observed characteristics in the baseline individual. Logit standard errors are clustered by respondent.

Figure 2 presents the partial effects for the pooled logit and PF-CRE estimators. While the CMLE and PF-CRE coefficient estimates are indistinguishable from one another, only the PF-CRE estimator provides estimates of probabilities and partial effects. To calculate the partial effects I use a baseline individual who is a man between 40 and 50 years of age, who works full time, owns his home outright, and finished high school, with all other variables set at the median for an individual with these characteristics.

The PF-CRE estimates show that when the baseline respondent is contacted by his most preferred party, he is 2.9% less likely to cast a tactical vote for a less preferred party, suggesting that party contact enforces party loyalty or sincerity in voters. Logit estimates this quantity at 13.7%, almost five times the effect. 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 6.6%. Logit also overestimates this effect, in this case at 21.3%.

The results presented here show that unobserved heterogeneity is an important confounder in the study of tactical voting during the 2015 U.K. General Election. This is evidenced by the significant overestimation of different effects when the heterogeneity is ignored. The PF-CRE estimator allows for the estimation of partial effects when accounting for the unobserved heterogeneity that other estimators cannot, and the results show that parties' efforts to contact voters during the pre-election season have significant effects on the probability that voters cast a tactical vote. These results are important, because they show that parties can benefit from contacting voters as a way to encourage or discourage them from voting tactically.

7 Additional Applications

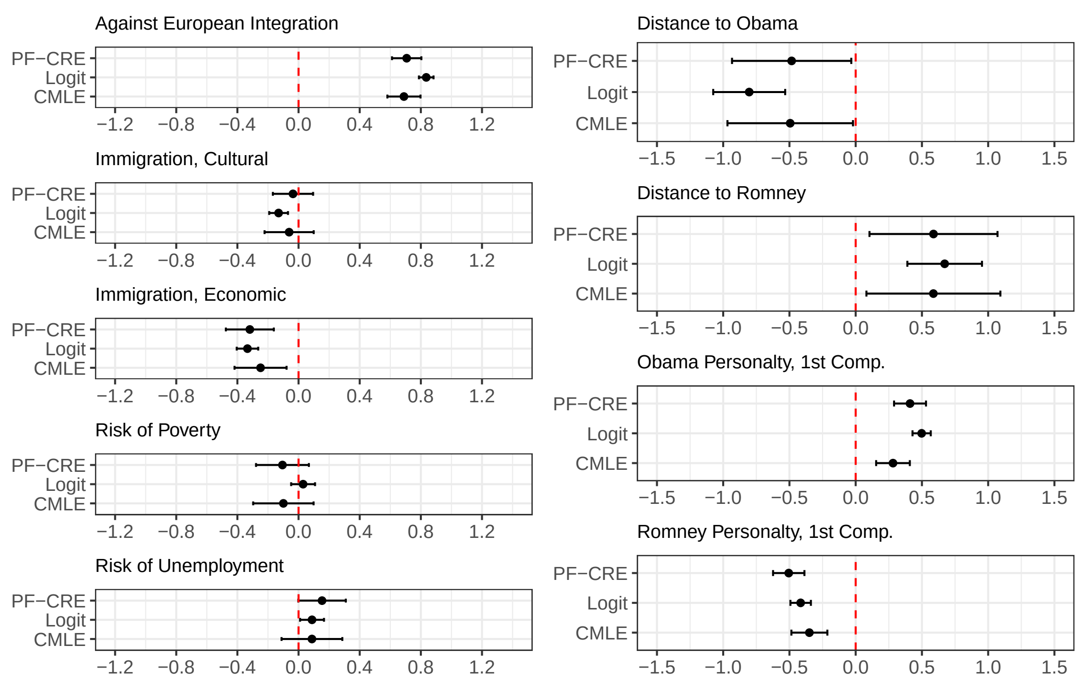
In this section I present a very brief discussion of two additional applications: (1) the effect of preferences for immigration and economic fears on voting decisions in the 2016 Brexit Referendum in the U.K.; and (2) the effect of ideological preferences and candidate personality perceptions on vote choice during the 2012 U.S. presidential election. The goal of this section is to show that the unobserved heterogeneity matters in these contexts and that PF-CRE provides consistent estimates of the model parameters. Further details and discussion of both these applications are available in appendices C and D.

In the Brexit Referendum case, the outcome of interest is voting in favor of Brexit.

The covariates of interest are preferences against European integration, views on immigration as it relates to British culture and the economy, and fears of falling into poverty or unemployment in the coming year. I model the conditional distribution of the unobserved heterogeneity with a total of 231 terms (of which 10 are selected by the penalized estimation). The specification test for PF-CRE returns a p-value of 0.16, which provides statistical evidence for its validity. As Figure 3a shows, coefficient estimates from PF-CRE are similar to those of CMLE, and have smaller standard errors. Importantly, pooled logit overestimates some effects and provides excessively small confidence intervals for other variables.

Figure 3: *Coefficient Estimates, Additional Applications*

(a) Brexit Referendum **(b)** 2012 U.S. Presidential Election



The tuning parameters for the SCAD penalty in the PF-CRE estimator was obtained through 5-fold cross validation using the Akaike information criterion. Logit standard errors are clustered by respondent.

In the case of the 2012 U.S. Presidential Election, the outcome of interest is voting for Obama. The covariates are respondents' ideological distances to Obama and Romney, and personality evaluations about the candidates. I model the conditional distribution of the unobserved heterogeneity in PF-CRE with 136 terms. The specification test supports the PF-CRE specification, with a p-value of 0.34. As Figure 3b shows, this is reflected in the similar coefficient estimates from PF-CRE and CMLE, with PF-CRE estimates generally having a slightly smaller variance. Pooled logit coefficients, on the other hand, overestimate the effects of personality evaluations and distance to Obama.

Put together, the main application to tactical voting in Britain, plus the two applications briefly described in this section show that PF-CRE is a valid alternative to estimating binary outcome models with unobserved heterogeneity. PF-CRE's value is two-fold: (1) it provides consistent estimates of the model parameters *and* partial effects, which estimators like CMLE cannot estimate, and (2) it provides more efficient estimates (albeit sometimes only slightly more efficient).

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 chapter, 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, and the penalization step induces a parsimonious model that results in efficiency gains. Using a model specification test, I show that these assumptions hold in three different applications to political behavior.

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.

The main application I provide for the PF-CRE estimator is to tactical voting during the 2015 U.K. General Election. I show that ignoring unobserved heterogeneity leads to overestimation of the effect of being contacted by the most preferred and most preferred viable parties on the probability of casting a tactical vote, by as much as a factor of five. The intuition behind this overestimation is that parties possibly know something about voters, that researchers do not observe, that makes them more attractive for proselytizing. This makes party contacts correlated with these unobserved factors, leading to biased estimates.

I also provide two additional applications on electoral behavior, one to vote choice during the 2012 U.S. Presidential Election, and the other to vote choice during the 2016 Brexit Referendum in the U.K. In both these cases, the assumptions of the PF-CRE estimator hold, and alternative estimators produce upward or downward biased estimates of the partial effects of interest. While the validity of PF-CRE must be determined on a case by case basis, these results suggest that it is feasible in a number of applications.

PF-CRE can be applied in other areas of social science beyond political behavior. An area where PF-CRE can be useful is to the study of comparative political institutions and international relations. In these types of environments, most of the variation in the data is usually across units; within unit variation is typically much smaller. For this reason, methods like CMLE and Fixed Effects tend to discard almost all of the information in the data, leading to mostly statistically non significant results. The alternative is to ignore unobserved heterogeneity in these environments, which is also not desirable. 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. This is accomplished via the penalization step: 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 and partial effects. However, further research is necessary to determine the gains that PF-CRE can achieve with these types of data.

A number of extensions to PF-CRE are possible. The most natural ones are extensions to discrete outcome models other than binary ones. Commonly used multinomial and ordered response models (like Conditional Logit and Ordered Probit) can incorporate unobserved heterogeneity in the form of correlated random effects (Wooldridge, 2010b). However, the penalization step in these cases requires some refining. As in the binary case, allowing for a flexible specification with a penalization step can help these models realistically capture the unobserved heterogeneity, without leading to inefficient estimates or very restrictive assumptions.

Another extension is to allow for the model coefficients and the coefficients in the conditional distribution of the unobserved heterogeneity to vary by individual in the form of random coefficients. Random coefficients can be powerful tools to capture unobserved heterogeneity (independent of the covariates). An extension in this direction can exploit recent developments in penalized estimation of generalized linear mixed models (see Hui et al., 2017).

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.

Aldrich, J. H. and McKelvey, R. D. (1977). A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections. *American Political Science Review*, 71(1):111–130.

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. and Brehm, J. (2002). *Hard Choices, Easy Answers: Values, Information, and American Public Opinion*. Princeton University Press, New Haven.

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).

Ansolabehere, S., Rodden, J., and Snyder, J. M. (2008). The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting. *American Politial Science Review*, 102:215–232.

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.

Becker, S. O., Fetzer, T., and Novy, D. (2017). Who Voted for Brexit? A Comprehensive District-Level Analysis. CEP Discussion Paper No. 1480.

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.

Brinegar, A., Jolly, S., and Kitschelt, H. (2004). Varieties of Capitalism and Political Divides over European Integration. In Marks, G. and Steenbergen, M., editors, *Eu-*

ropean Integration and Political Conflict, pages 62–89. Cambridge University Press, Cambridge.

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.

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

Curtice, J. (2016). A Question of Culture of Economics? Public Attitudes to the European Union in Britain. *The Political Quarterly*, 87(2):209–218.

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

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

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.

Feldman, S. (1988). Structure and Consistency in Public Opinion: the Role of Core Beliefs and Values. *American Journal of Political Science*, 32(2):416–440.

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.

Garry, J. and Tilley, J. (2015). Inequality, State Ownership and the EU: How Economic Context and Economic Ideology Shape Support for the EU. *European Union Politics*, 16(1):139–154.

Gerber, A. S., Green, D. P., and Larimer, C. W. (2008). Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment. *American Political Science Review*, 102(1):33–48.

Goren, P. (2005). Party Identification and Core Political Values. *American Journal of Political Science*, 49(4):881–896.

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

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

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.

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.

Inglehart, R. F. and Norris, P. (2016). Trump, Brexit, and the Rise of Populism: Economic Have-Nots and Cultural Backlash. HKS Faculty Research Working Paper Series, RWP16-026.

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

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

McCann, J. A. (1997). Electoral Choices and Core Value Change: The 1992 Presidential Campaign. *American Journal of Political Science*, 41(2):564–583.

McLaren, L. M. (2002). Public Support for the European Union: Cost/Benefit Analysis or Perceived Cultural Threat? *Journal of Politics*, 87(2):209–218.

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 Consistent Observations. *Econometrica*, 16:1–32.

Poole, K., Rosenthal, H., Lewis, J., Lo, J., and Carroll, R. (2013). `basicspace`: A Package to Recover a Basic Space from Issue Scales.

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

Shipman, T. (2016). *All Out War: The Full Story of How Brexit Sank Britain's Political Class.* William Collins, London.

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

Tucker, J. A., Pacek, A. C., and Berinsky, A. J. (2002). Transitional Winners and Losers: Attitudes Toward EU Membership in Post-Communist Countries. *American Journal of Political Sciene*, 46(3):557–571.

Wooldridge, J. M. (2010a). Correlated random effects models with unbalanced panels. Mimeo, Michigan State University.

Wooldridge, J. M. (2010b). *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

California Institute of Technology

A Additional Figures and Tables from Simulations

Table A1: *$\widehat{\beta}$ Bias relative to true β*

	Sparse		RE		Complex	
	CMLE	PF-CRE	CMLE	PF-CRE	CMLE	PF-CRE
β_1	1.37%	−0.60%	0.30%	−0.82%	2.76%	0.32%
β_2	2.15%	−0.22%	1.09%	−0.86%	2.40%	0.36%
β_3	0.43%	−1.30%	−0.15%	−1.99%	3.63%	1.37%
β_4	1.87%	−0.50%	1.65%	−0.71%	2.66%	0.69%
β_5	2.53%	−1.11%	0.71%	−2.11%	4.64%	1.82%

The quantities in this table are calculated as: $(\widehat{\beta}/\beta - 1) \times 100$.

Table A2: *$\widehat{\beta}$ Standard Deviation relative the standard deviation of the Oracle*

	Sparse		RE		Complex	
	CMLE	PF-CRE	CMLE	PF-CRE	CMLE	PF-CRE
β_1	1.319	1.026	1.727	1.000	1.394	1.000
β_2	1.467	1.001	1.711	1.023	1.497	0.995
β_3	1.694	1.004	1.766	1.011	1.336	1.001
β_4	1.845	1.002	1.772	1.017	1.494	1.001
β_5	1.744	0.998	1.815	1.001	1.289	1.000

A value of 1 indicates identical standard error to the Oracle estimator. Larger (smaller) values indicate a larger (smaller) standard error than the Oracle's.

Table A3: *$\widehat{PE}$ Bias relative to true PE*

	Sparse				RE				Complex			
	PF-CRE	Logit	UF-CRE	CRE	PF-CRE	Logit	UF-CRE	CRE	PF-CRE	Logit	UF-CRE	CRE
β_1	−2.48%	23.45%	−2.08%	−14.89%	−1.56%	−0.91%	−1.34%	−3.48%	−1.26%	25.39%	−1.77%	−23.35%
β_2	−1.88%	−30.81%	−1.47%	−14.40%	−1.65%	−1.01%	−1.39%	−3.39%	−0.86%	2.84%	−1.38%	−23.21%
β_3	−2.29%	−9.84%	−2.01%	−14.84%	−1.05%	−0.39%	−1.37%	−3.47%	−1.32%	−108.04%	−2.12%	−23.66%
β_4	−1.45%	−8.66%	−0.90%	−13.84%	−1.72%	−1.09%	−1.12%	−3.23%	−1.09%	3.90%	−1.62%	−23.24%
β_5	−1.84%	−8.41%	−1.38%	−14.05%	−1.50%	−0.87%	−0.96%	−3.06%	0.00%	−189.86%	−0.55%	−21.88%

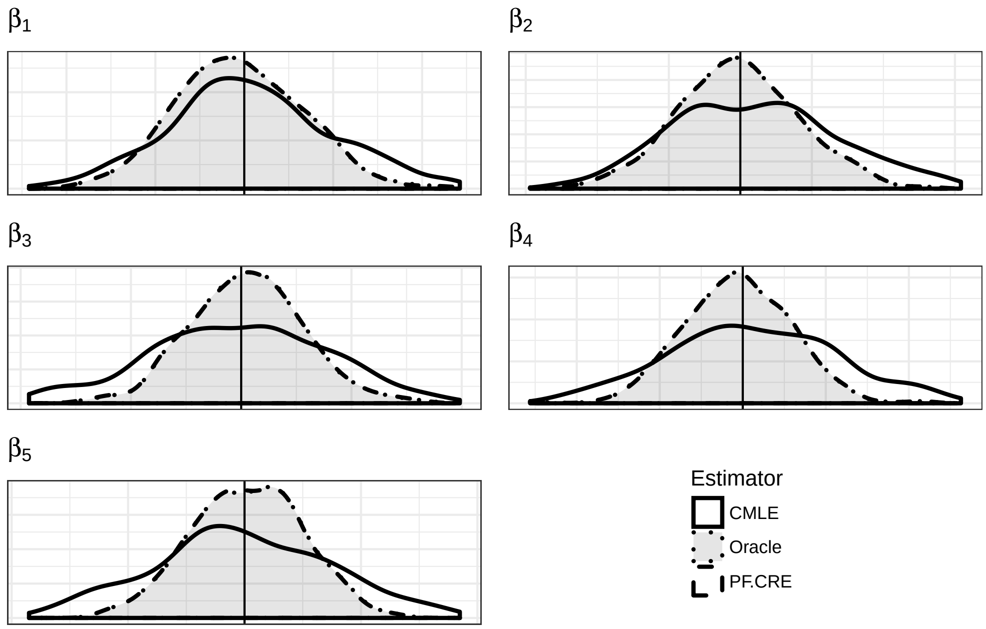
The quantities in this table are calculated as: $(\widehat{P}E/PE - 1) \times 100$.

Table A4: *$\widehat{PE}$ Standard error relative to the standard error of the Oracle estimator*

	Sparse				RE				Complex			
	PF-CRE	Logit	UF-CRE	CRE	PF-CRE	Logit	UF-CRE	CRE	PF-CRE	Logit	UF-CRE	CRE
β_1	0.973	0.726	1.070	3.125	1.000	1.010	1.559	2.382	0.994	0.673	1.165	3.189
β_2	1.000	0.738	1.275	5.336	0.998	1.003	1.928	3.513	0.996	0.578	1.283	4.166
β_3	0.980	0.956	1.379	2.682	1.001	1.007	1.481	1.841	0.997	0.618	1.064	2.071
β_4	0.990	0.934	1.549	6.129	1.000	1.003	1.767	3.153	0.991	0.636	1.273	4.133
β_5	0.974	0.931	1.347	1.765	0.999	1.006	1.377	1.477	1.000	0.703	1.021	1.345

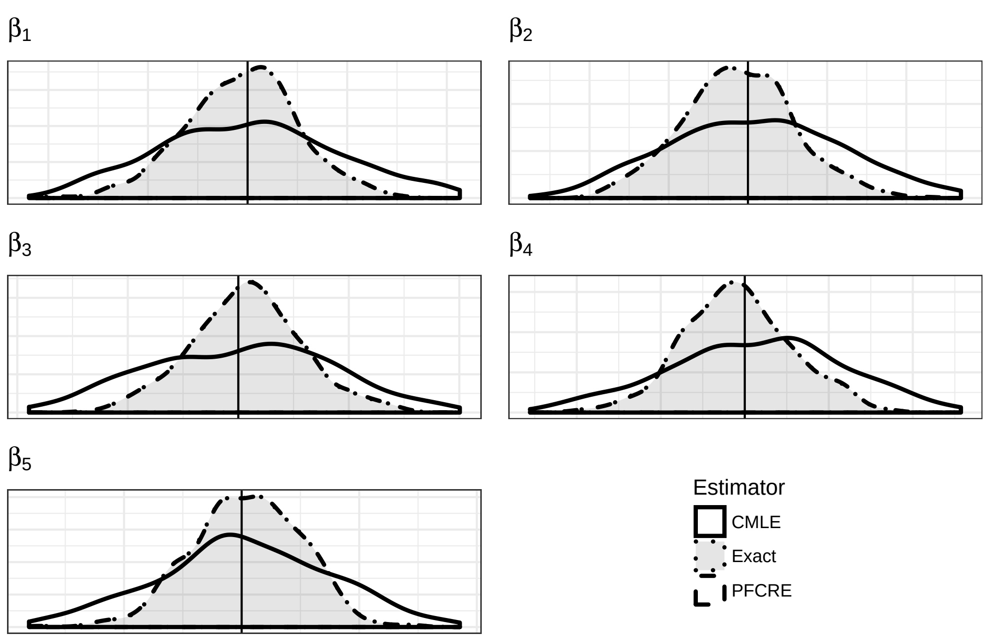
A value of 1 indicates identical standard error to the Oracle estimator. Larger (smaller) values indicate a larger (smaller) standard error than the Oracle's.

Figure A1: *$\widehat{\beta}$ Distributions: PF-CRE v. CMLE, Sparse Specification*



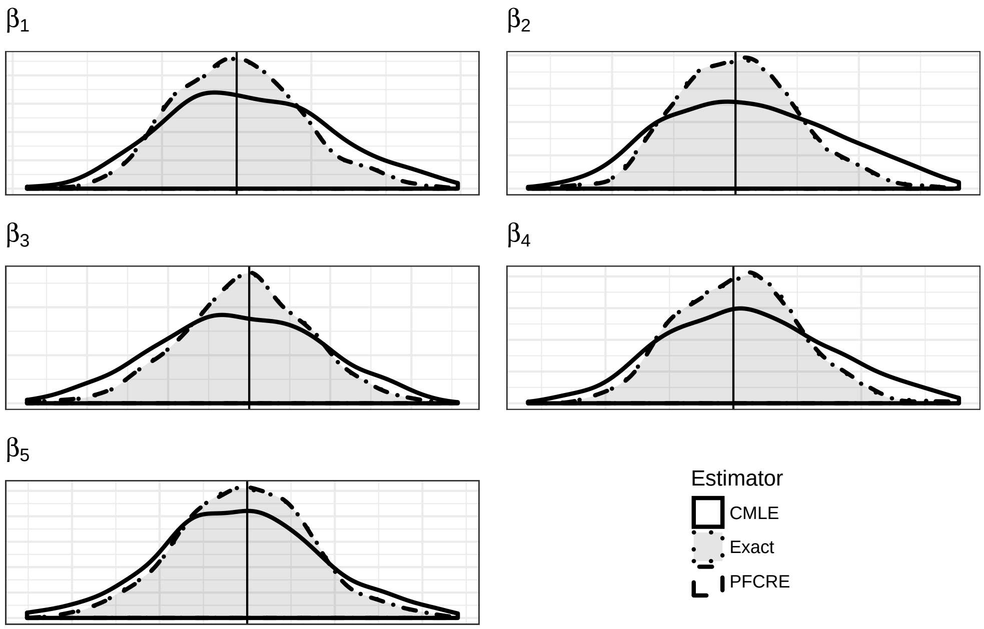
Vertical lines represent the true value of the parameters. The distributions correspond to the estimates for each parameter and estimator.

Figure A2: *$\widehat{\beta}$ Distributions: PF-CRE v. CMLE, Random Effect Specification*



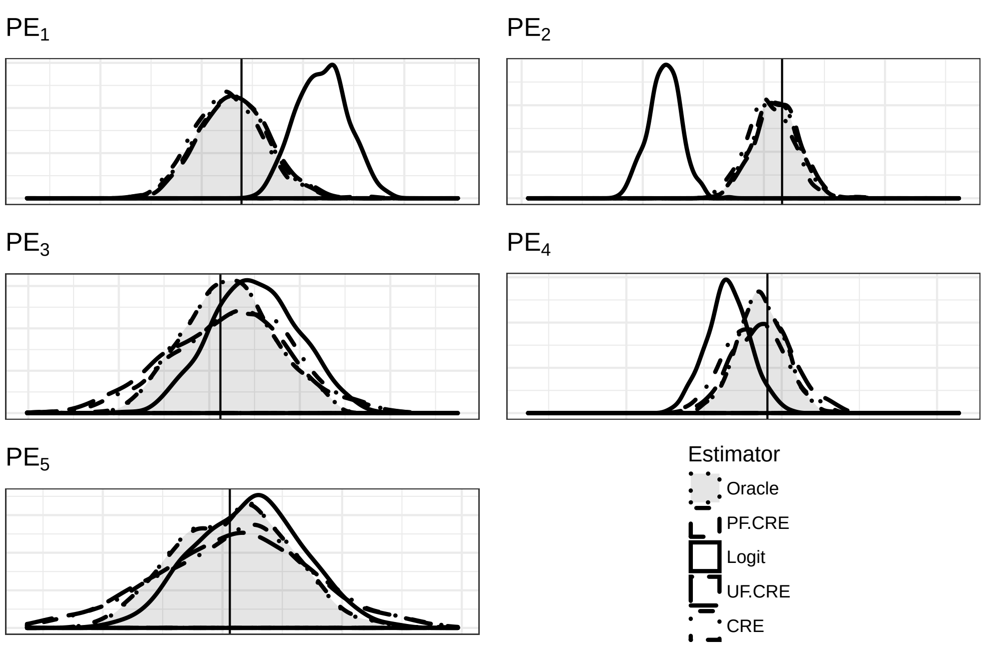
Vertical lines represent the true value of the parameters. The distributions correspond to the estimates for each parameter and estimator.

Figure A3: *$\widehat{\beta}$ Distributions: PF-CRE v. CMLE, Complex Specification*



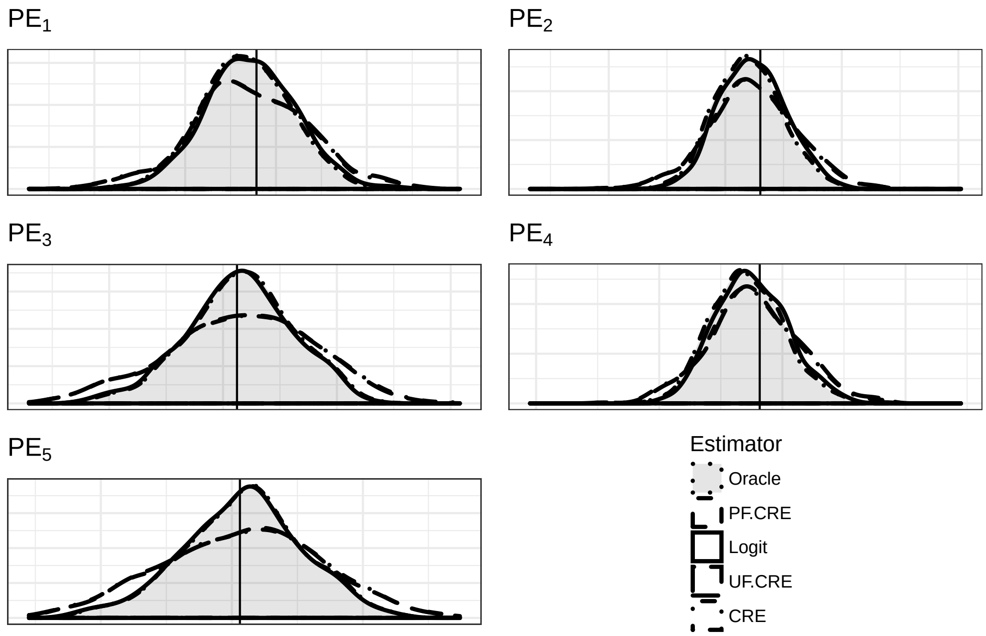
Vertical lines represent the true value of the parameters. The distributions correspond to the estimates for each parameter and estimator.

Figure A4: *$\widehat{PE}$ Distributions, Sparse Specification Specification*



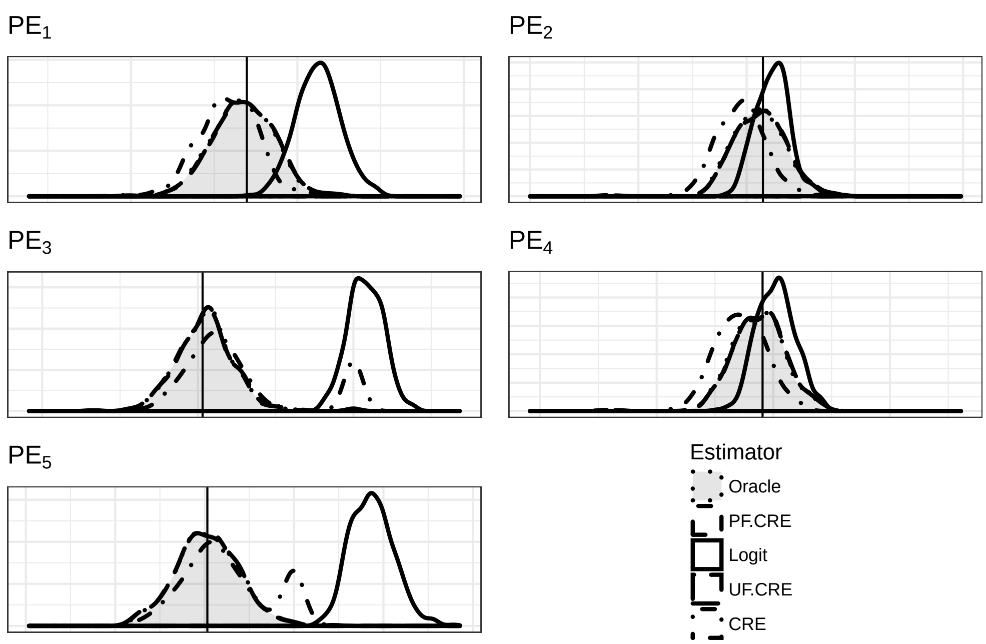
Vertical lines represent the true value of the parameters. The distributions correspond to the estimates for each parameter and estimator.

Figure A5: $\widehat{PE}$ *Distributions, Random Effect Specification*



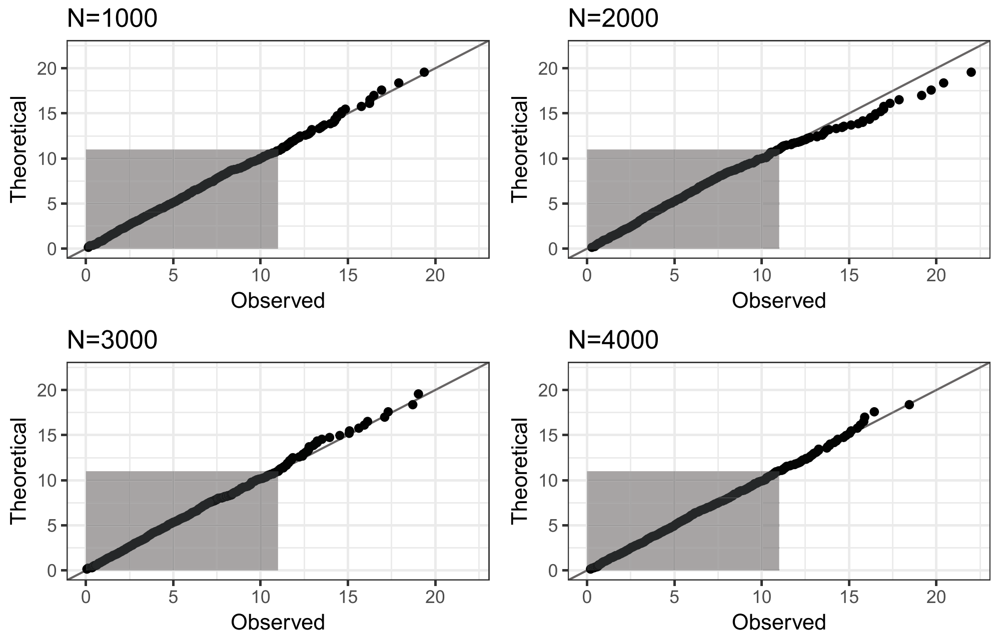
Vertical lines represent the true value of the parameters. The distributions correspond to the estimates for each parameter and estimator.

Figure A6: $\widehat{PE}$ *Distributions, Complex Specification*



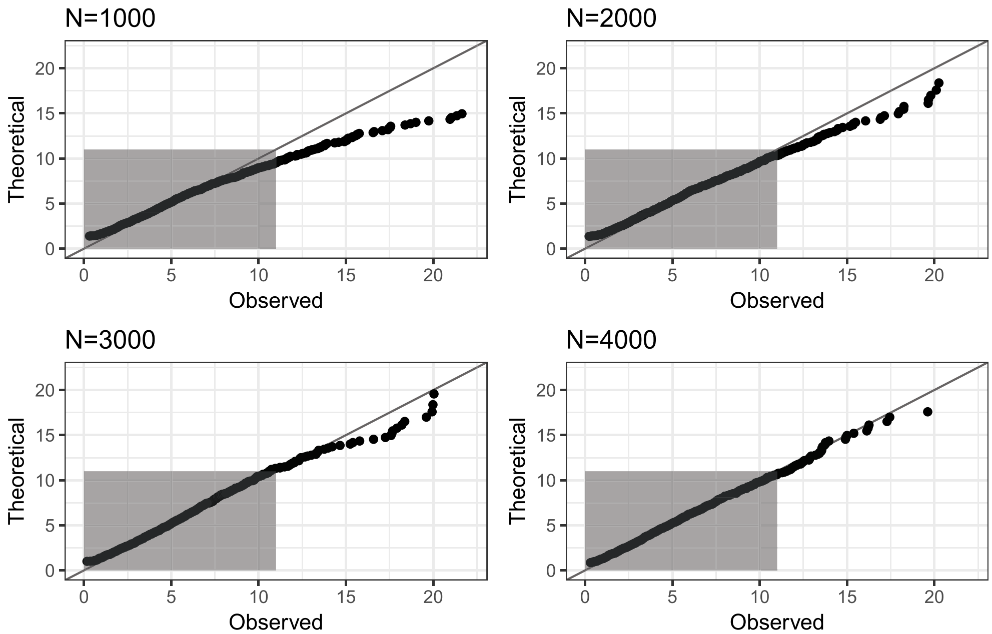
Vertical lines represent the true value of the parameters. The distributions correspond to the estimates for each parameter and estimator.

Figure A7: *Specification Test, Quantile-Quantile Plots, Sparse Specification*



Observed are the sample quantiles from simulations of the PF-CRE specification test. Theoretical are the theoretical quantiles from a $\chi^2_{(5)}$. *The shaded area represents the 95% theoretical quantile.*

Figure A8: *Specification Test, Quantile-Quantile Plots, Complex Specification*



Observed are the sample quantiles from simulations of the PF-CRE specification test. Theoretical are the theoretical quantiles from a $\chi^2_{(5)}$. *The shaded area represents the 95% theoretical quantile.*

B Additional Figures and Tables from Tactical Voting Application

Table B1: *Coefficient Estimates, Tactical Voting 2015 U.K. Election*

	PF-CRE			CMLE			Logit		
	β	Low	High	β	Low	High	β	Low	High
Contact Preferred	-0.33	-0.57	-0.09	-0.32	-0.57	-0.08	-0.90	-1.03	-0.77
Contact Viable	0.65	0.43	0.87	0.71	0.49	0.94	0.94	0.81	1.06
Therm. Preferred	-0.17	-0.26	-0.09	-0.18	-0.27	-0.10	-0.50	-0.53	-0.46
Therm. Viable	0.30	0.23	0.37	0.31	0.24	0.38	0.39	0.36	0.43
Controls	No			No			Yes		
N^o γ terms	230			-			-		
Selected γs	42			-			-		
n	3,824			3,824			3,824		
Effective n	3,824			1,164			3,824		
Observations ($n \times T_i$)	10,378			10,378			10,378		
Effective Obs.	10,378			3,263			10,378		
$\chi^2_{(4)}$	4.96			-			-100.21		
p-value	0.29			-			NA		

All confidence intervals are at the 95% level. The tuning parameters for the SCAD penalty in PF-CRE was obtained through 5-fold cross-validation. 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.

C Application: Brexit Referendum

During the 2015 British General Election, internal struggles within the Conservative Party lead Prime Minister David Cameron to promise a referendum on E.U. membership (Becker et al., 2017). In the run up to the Brexit Referendum, held on June 23rd, 2016, many arguments were presented for leaving the European Union. Some of them had to do with ensuring British independence from bureaucrats in Brussels or with preventing U.K. taxpayer money from lining up the Euro-coffers. In fact, the Leave campaign stressed that by leaving the EU, the U.K. would save £350 million each week.[45] Other arguments had to do with immigration (both from the E.U. and from other countries as a result of E.U. policy) and its effects on the British economy and culture. Research suggests that hostility towards the European Union has been fueled by the perception that E.U. membership represents a cultural threat (McLaren, 2002; Curtice, 2016; Inglehart and Norris, 2016). On the economic side, voters with fears of losing employment or of their economic well-being being negatively affected by E.U. policy were expected to be more favorable towards the U.K. exit from the European Union.[46]

In this application, I focus on whether fears of falling into poverty or unemployment affected voters' decision to support or oppose Brexit. Estimating these effects in a causal manner is not trivial, however. Notably, these economic fears may be more prevalent among certain groups of the population that, at the same time, are more (or less) likely to support Brexit for other reasons, some of which may be unobserved.

C.1 Data and Model Specification

I use a panel data survey from the British Election Study Online Panel, collected prior to the Brexit Referendum. These data allow me to study how variation in individuals' economic fears and immigration concerns played out in their referendum vote decisions.

[45] For an account of the Brexit Referendum campaign, see Shipman (2016).

[46] There is a relatively large literature that focuses on an utilitarian approach to European integration. See, for example, Tucker et al. (2002); Brinegar et al. (2004); Garry and Tilley (2015).

The main variables of interest indicate respondents' beliefs that in the next 12 months they will fall into poverty or unemployment, both on a scale from 1 to 5. I also include respondents' overall preferences against European integration, on a scale from 0 (unite fully with the European Union) to 10 (protect our independence).[47] I also include two additional questions about attitudes towards immigration. The first one measures respondents' beliefs on whether immigration is good for Britain's economy, on a scale from 1 (bad) to 7 (good); the second measures respondents' beliefs on whether immigration enriches Britain's cultural life, on a scale from 1 (undermines cultural life) to 7 (enriches cultural life). I also include a number of time-invariant characteristics: identification as middle or working class, age, education, household income, race, employment status, gender, and personality traits.

I model the conditional distribution of the unobserved heterogeneity using PF-CRE with the time-means of the covariates of interest, plus the time-invariant characteristics, and up to two-way interactions among these terms, for a total of 231 terms. I compare the estimates from PF-CRE to those of Conditional Maximum Likelihood (CMLE), a pooled logit estimator that includes the time-invariant characteristics as controls, and a traditional CRE approach that only uses the time-means of the covariates to model the conditional distribution of the unobserved heterogeneity.

C.2 Results

Table C1 presents the coefficient estimates from the three methods considered. The coefficient estimates from PF-CRE and CMLE are similar, with PF-CRE having smaller confidence intervals. The specification test of the null hypothesis that PF-CRE is consistent and more efficient than CMLE returns a p-value of 0.16, which implies the validity of PF-CRE. The pooled logit, which ignores the unobserved heterogeneity, overestimates some effects, particularly the coefficient on preferences against European

[47]Some respondents were assigned to a different version of this question, on a scale from 0 (unification has already gone too far) to 10 (unification should be pushed further). Results change slightly when using this version of the question. However, qualitative (and to a large extent) quantitative results remain the same.

integration. Pooled logit also estimates a significant effect of believing that immigration enriches Britain's cultural life. However, this effect disappears when accounting for the unobserved heterogeneity as in PF-CRE and CMLE.

The estimates of being at risk of unemployment highlight the efficiency advantage of PF-CRE relative to CMLE. While PF-CRE and CMLE provide very similar point estimates, the inefficiency of CMLE would incorrectly lead to the conclusion that there is no statistically significant effect of fears of unemployment on voting for Brexit. PF-CRE, one the other hand, shows that this effect is statistically significant.

Table C1: *Coefficient Estimates for Brexit Referendum*

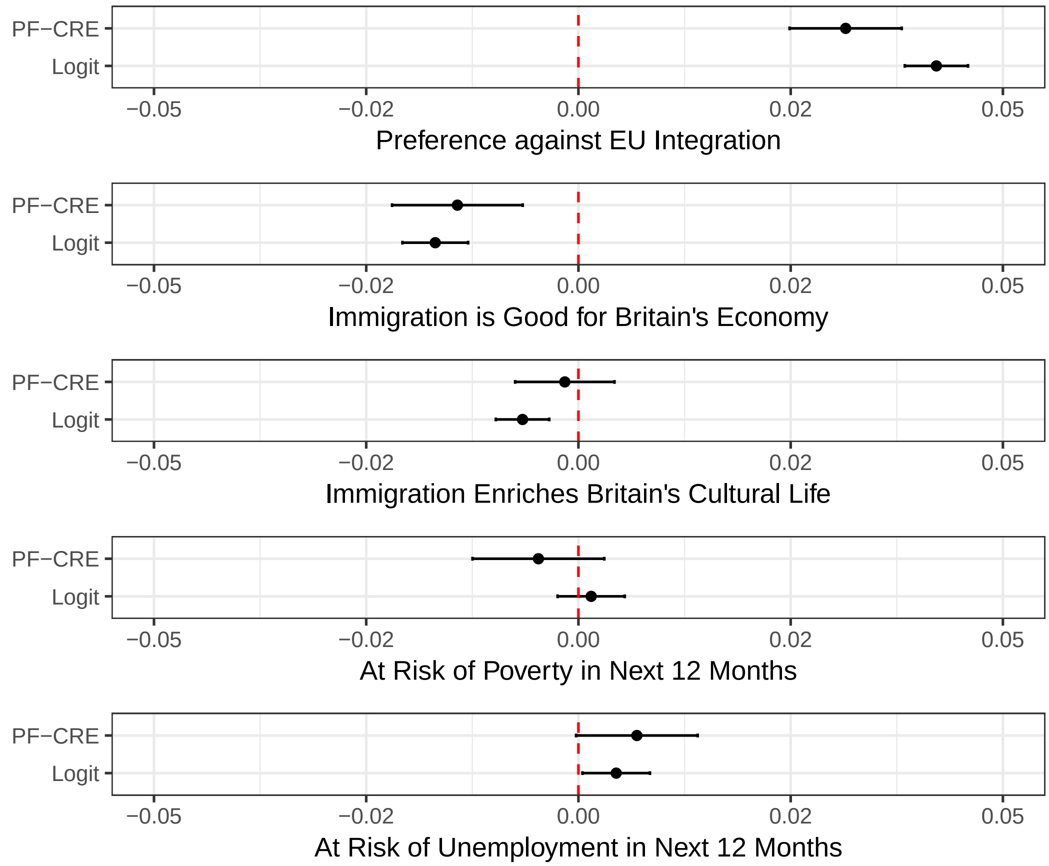
	PF-CRE			CMLE			Logit		
	β	Low	High	β	Low	High	β	Low	High
Against Integration	0.71	0.61	0.80	0.69	0.58	0.80	0.84	0.79	0.88
Immigration, Cultural	-0.04	-0.17	0.10	-0.06	-0.22	0.10	-0.13	-0.19	-0.07
Immigration, Economic	-0.32	-0.48	-0.16	-0.25	-0.42	-0.08	-0.33	-0.40	-0.26
Risk Poverty	-0.11	-0.28	0.07	-0.10	-0.30	0.10	0.03	-0.05	0.11
Risk Unemployment	0.15	0.00	0.31	0.09	-0.11	0.29	0.09	0.01	0.17
Controls	No			No			Yes		
N^o γ terms	231			-			-		
Selected γs	10			-			-		
Observations	8,033			8,033			8,033		
Effective Obs	8,033			1,132			8,033		
$\chi^2_{(5)}$	7.94			-			-15.18		
p-value	0.16			-			-		

All confidence intervals are at the 95% level. The tuning parameters for the SCAD penalty in PF-CRE was obtained through 5-fold cross-validation. Logit standard errors are clustered at the individual level. The effective number of observations refers to the number of actual observations used in estimation for CMLE. There is no χ^2 test reported for CMLE since this estimator is the basis for that test.

Figure C9 presents the partial effects estimated for a baseline individual.[48] The PF-CRE estimates show that an increase in preferences against European integration are associated with a 3.15% increase in the probability of voting in favor of Brexit; logit overestimates this effect by a third. In terms of respondents' views on immigration, the results show that those who find that immigration is good for the economy are

[48] The baseline individual is a 45 year old white male, who is employed full time and has some college education. All other variables were set to the average value for an individual with those characteristics, as observed in the sample.

Figure C9: *Partial Effects from Brexit Referendum*



The tuning parameter for the SCAD penalty in PF-CRE was obtained through 5-fold cross validation using the Akaike information criterion. Logit standard errors are clustered by respondent. Baseline values for the conditional mean equation for PF-CRE in partial effects were chosen to be consistent with those of the observed characteristics of the baseline individual.

1.43 percentage points less likely to support Brexit, whereas there is no effect on the cultural side. Logit overestimates both these effects, at 1.69 and 0.66 percentage points, respectively. Finally, both estimators show that respondents who consider themselves at risk of unemployment in the near future are more likely to support Brexit, by about 0.45 percentage points.

Overall, the results show that accounting for unobserved heterogeneity in the Brexit context has some important implications for our understanding of voting behavior. Beyond the overestimation of various effects, there is no evidence that cultural fears actually drive support for Brexit. On the other hand, even after controlling for unobserved heterogeneity, the evidence shows that those voters with fears of losing their jobs in the near future are more likely to support the U.K.'s exit from the European Union. These results suggest that materialist concerns were the prime drivers of the referendum results, and that values related to Britain's culture did not play a significant role.

D Application: 2012 U.S. Presidential Election

The study of how voters make choices in elections has generally focused on two main axes: (1) ideological preferences, and (2) valence issues. The first axis is typically represented by the ideological distance between voters and the candidates, usually measured as part of standard political surveys.[49] The second axis is measured in surveys through questions, or batteries of questions, aimed at determining voters' opinion on different personal characteristics of the candidates, beyond their political positions: whether they think the candidates are moral, experienced, care about regular people, among others.

Unobserved heterogeneity is usually present in observational studies of ideology and valence and how they relate to individuals' vote choices. Important variables are not measured, are hard to measure, or are simply not available in the data at hand. For example, core values, which are hard to accurately capture in surveys, can be important motivators behind vote choices. The challenge they pose is that they are generally correlated with voters' ideological and personality evaluations about the candidates (Alvarez and Brehm, 2002; Feldman, 1988). Therefore, ignoring them leads to biased inferences about these variables. Core values are generally thought of as fixed, at least in the short and near term (Feldman, 1988; McCann, 1997).[50] Therefore, treating them as unobserved heterogeneity during the course of an election campaign is an appropriate course of action when they are unobserved or unmeasured.

Beyond the omitted variable bias, there are other challenges that accounting for un-

[49] Other focus on particular issue positions, sometimes in combination with overall ideological positions.

[50] Goren (2005) challenges that core values are largely fixed, and posits that they are influenced by partisan identification.

observed heterogeneity can help ameliorate in the context of vote choice. For example, positive evaluations of a candidate are usually associated with a higher probability of casting a vote for that candidate. However, a voter who has decided to cast a vote for a given candidate may then begin viewing that candidate's personality under a kinder light (even if just to diminish cognitive dissonance). Unobserved heterogeneity can alleviate this problem by accounting for individuals' general tendency to have positive (or negative) views about a candidate; the remaining variation in the data is more likely to reflect how changes in individuals' views about the candidates affect vote choices, than the other way around.

D.1 Data and Model Specification

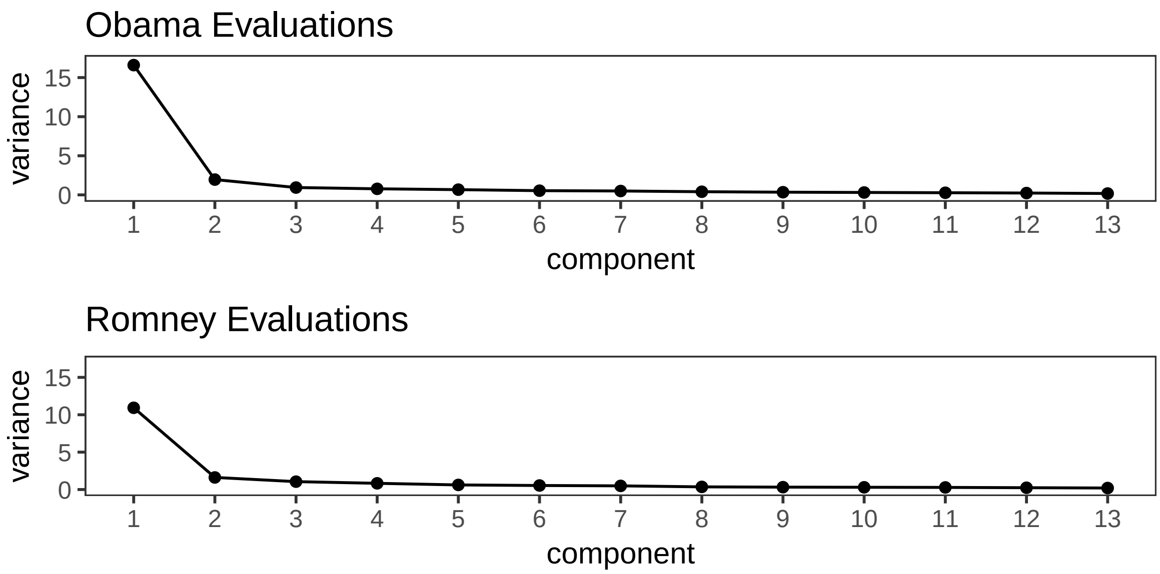
To study the effect of ideological distance and candidate personality evaluations on vote choice, I use data from three waves of The American Panel Study (TAPS) from the 2012 U.S. Presidential Election. The outcome of interest is whether a respondent intends to vote for Obama during the General Election (Romney voters, non-voters, and third party voters are grouped together for the analysis).

The variables of interest are the ideological distance of each respondent to Obama and Romney, and individuals' perceptions about the candidates' personalities. I construct ideological distance as the absolute distance between the respondents' self-reported ideological position and their perceptions about the candidates' positions. Given the well-known problems of differential item functioning, self placements were adjusted using the Aldrich-McKelvey rescaling (Aldrich and McKelvey, 1977), as implemented in the `basicspace` package in `R` (Poole et al., 2013).

Voters' perceptions of the candidates' personalities are based on a battery of 10 questions.[51] These evaluations are very highly correlated with each other, and using them all together in a model introduces more noise than explanatory power (Ansolabehere et al., 2008, make a similar argument for the case of issue positions). For this reason,

[51] Respondents are asked to rate the following statements for each candidate: He is optimistic, He is partisan, He is fair, He is a strong leader, He is trustworthy, He is experienced, He is knowledgeable, He is inspiring, He is decisive, He cares about people like me, He is moral, He has a bad temper.

Figure D1: *Principal Components 2012 Presidential Election*



Principal components of personality evaluations were calculated separately for each candidate. The evaluations for each candidate consist of 10 items, each ranging from 1 (disagree) to 7 (agree).

I simplify the personality evaluations by replacing them with their first two principal components for each candidate, as additional dimensions do not contribute significantly to explaining the variance in each candidate evaluations (see Figure D1).

To model the conditional distribution of the unobserved heterogeneity in PF-CRE, I use the time-means of the covariates of interest, plus time-invariant characteristics, with up to two-way interactions, for a total of 136 terms. The time-invariant characteristics I include are race, income, year of birth, education, gender, and party identification from the first wave of the panel.[52] I compare the estimates from PF-CRE to those of CMLE, and a pooled logit estimator that includes the time-invariant characteristics as control variables.

D.2 Results

Table D1 shows the coefficient estimates for the main variables of interest in the model: ideological distance and the first two components of the candidate personality evalua-

[52]Party identification in the TAPS data shows some variation across panel waves for some individuals. However, I choose to use the responses from the first wave, as subsequent variation is possibly a reflection of measurement error rather than actual changes in party identification.

tions for Obama and Romney. The point estimates for PF-CRE and CMLE are similar to each other, with PF-CRE estimates generally having a slightly smaller variance. In fact, the specification test does not reject the null hypothesis that PF-CRE is consistent and more efficient than CMLE, with a p-value of 0.34. The pooled logit model, which does not account for the unobserved heterogeneity, significantly overestimates the effect of ideological distance to Obama, by around 66%. Logit also provides incorrect estimates of the other parameters, but with a smaller bias than ideological distance to Obama. These differences highlight the importance of controlling for unobserved heterogeneity in the estimation of vote choice.

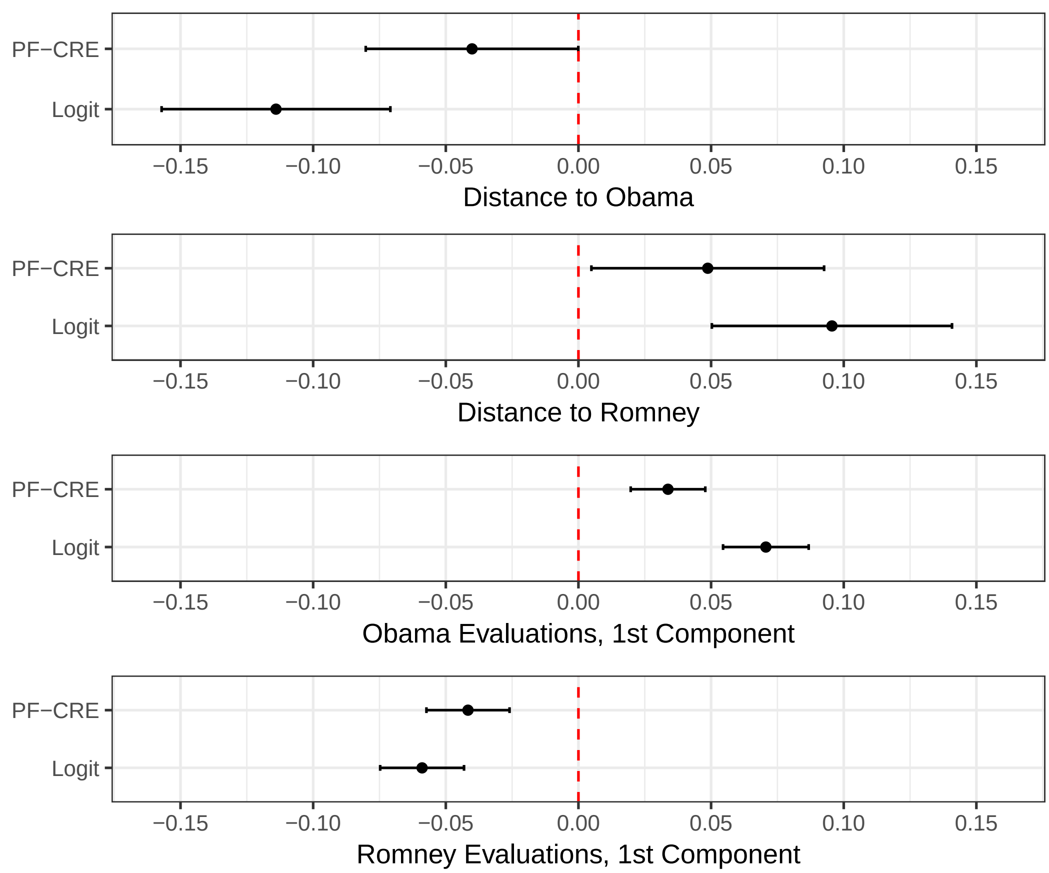
Table D1: *Coefficient Estimates for 2012 U.S. Presidential Election*

	PF-CRE			CMLE			Logit		
	β	Low	High	β	Low	High	β	Low	High
Distance BO	-0.48	-0.93	-0.03	-0.49	-0.97	-0.02	-0.80	-1.08	-0.53
Distance MR	0.59	0.10	1.07	0.59	0.08	1.09	0.67	0.39	0.95
BO Eval, 1st	0.41	0.29	0.53	0.28	0.15	0.41	0.50	0.43	0.57
MR Eval, 1st	-0.51	-0.62	-0.39	-0.35	-0.49	-0.21	-0.42	-0.49	-0.34
BO Eval, 2nd	-0.30	-0.55	-0.04	-0.25	-0.55	0.05	-0.28	-0.44	-0.12
MR Eval, 2nd	0.16	-0.10	0.42	0.16	-0.10	0.43	0.22	0.06	0.37
Controls	No			No			Yes		
N^o γ terms	136			-			-		
Selected γs	22			-			-		
Observations	2,253			2,253			2,253		
Effective Obs	2,253			626			2,253		
$\chi^2_{(8}$	6.85			-			23.76		
p-value	0.34			-			0.00		

All confidence intervals are at the 95% level. The tuning parameters for the SCAD penalty in PF-CRE was obtained through 10-fold cross-validation. Logit standard errors are clustered at the individual level. The effective number of observations refer to the number of actual observations used in estimation for CMLE. There is no χ^2 test reported for CMLE since this estimator is the basis for that test.

Figure D2 shows the partial effects estimated from PF-CRE and the pooled logit estimators. The baseline individual for these partial effects is a 45 years old white man, neither a democrat nor republican, with all other variables set a the mean for a person with these characteristics.

Figure D2: *Partial Effects from 2012 U.S. Presidential Election*



The tuning parameter for the SCAD penalty in PF-CRE was obtained through 10-fold cross validation using the Akaike information criterion. Logit standard errors are clustered by respondent. Baseline values for the conditional distribution of the unobserved heterogeneity in the partial effects where chosen to be consistent with those of the observed characteristics of the baseline individual.

The partial effects from PF-CRE show that, for the baseline individual, increasing the ideological distance to Obama is associated with a 4.0 percentage points decrease in the probability of voting for him. An increase in the ideological distance to Romney increases the probability of voting for Obama by about 4.8 percentage points. Pooled logit, which ignores the unobserved heterogeneity, overestimates these effects by about 175 and 88 percent, respectively. A similar picture arises from personality evaluations. PF-CRE estimates that a more positive evaluation of Obama is associated with a 3.4 percentage points increase in the probability of voting for him, whereas better personality evaluations of Romney are associated with a decrease of about 4.1 percentage point in the probability of voting for Obama. Pooled logit overestimates these effets by 110 and 42 percent, respectively.

D.3 Discussion

Overall, the partial effects from PF-CRE show that voters' perceptions of personality characteristics and ideological distance for both candidates have effects of similar size. While ideological distance to the candidates is an important predictor of vote choice, after controlling for unobserved heterogeneity, its partial effect is of comparable size to that of personality evaluations. Furthermore, the partial effects for ideological distance have a large degree of uncertainty relative to those of personality evaluations. These results suggest that ideological considerations are not the dominant axis along which vote intentions move, at least within the time-frame of an election year. Instead, candidate personality evaluations are of similar importance, and have a stronger statistical association with vote choice.

The difference between pooled logit and PF-CRE estimates of the partial effects for ideological distance and personality evaluations point to two related conclusions, one methodological and the other substantive. On the methodological side, this difference is illustrative of the perils of ignoring unobserved heterogeneity. As the pooled logit shows, this leads to partial effects that can be twice as large as those of a model that controls for the unobserved heterogeneity. On the substantive side, the smaller partial effects of the PF-CRE model, and specifically those for ideological distance, are possibly an indication of the effects of political polarization, as they point to choices that are weakly responsive to changes in voters' perceptions to ideology during the campaign than would otherwise be expected.