Unobserved Heterogeneity in Observational Studies of Political Behavior

Thesis by

Lucas Núñez

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2018
Defended May 4, 2018

Lucas Núñez
ORCID: 0000-0001-5107-6775

ACKNOWLEDGEMENTS

I am deeply indebted to my committee members, R. Michael Alvarez, Philip Hoffman, Jonathan Katz, and Robert Sherman, for their continued encouragement, support, and advice, both academic and professional. I am particularly indebted to my academic advisor and committee chair, R. Michael Alvarez, for his unwavering commitment to research and mentoring, and for his gentle probing that has kept me on course throughout my graduate studies. Additionally, my work in this thesis has significantly benefited from comments and advice from Ben Gillen, Rod Kiewiet, and Jean-Laurent Rosenthal.

I also thank Kapáuhi Stibbard, Sabrina De Jaegher, and especially Laurel Auchampaugh, whose dedication as staff members made studying at Caltech a much easier and enjoyable experience.

This thesis has benefited from conversations and helpful comments from Jun Chen, Laural Doval, Federico Echenique, Michael Ewens, Marcelo Fernández, Michael Gibilisco, Alex Hirsch, Yimeng Li, Gabriel Lopez Moctezuma, Tatiana Mayskaya, Sergio Montero, Alejandro Robinson, Welmar Rosado, Matthew Shum, and Mali Zhang, as well as participants at the HSS Graduate Proseminar.

On a personal level, I am grateful to my parents, Raul and Laura, who have always supported me in my education; and to them and my brothers, Francisco and Agustín, for their encouragement and understanding.

ABSTRACT

This dissertation comprises three chapters dealing with unobserved heterogeneity in observational studies. In Chapters 2 and 3, I develop new estimators that deal with unobserved heterogeneity in the cases in which panel data is not available or the outcome of interest is binary, respectively. In Chapter 4, I analyze the effect of parties' contacting voters on the extent of tactical voting in the 2015 and 2017 United Kingdom General Elections, applying the estimator developed in Chapter 3.

In Chapter 2, I develop a semi-parametric two-step estimator for linear models with unobserved individual level heterogeneity that can be applied on a series of Repeated Cross-Sections, when panel data is unavailable. I show that this estimator provides consistent and asymptotically normal estimates of the parameters of interest. Identification relies on a restriction that requires the conditional expectation of the unobserved individual-level heterogeneity on observed characteristics to be continuous. Using Monte Carlo simulations, I show that this estimator typically outperforms other available alternatives. In particular, it typically has a smaller Root Mean Squared Error, and a relatively small bias that disappears for moderate sample sizes. Furthermore, it is robust to mild violations of the continuity assumption. Finally, I also show that this estimator can recover sensible estimates compared to those from an real panel.

In Chapter 3, 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 of 2016, 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.

In Chapter 4, I apply the PF-CRE estimator developed in Chapter 3 to the study

of tactical voting in the United Kingdom General Elections of 2015 and 2017. In particular, I study the effect that party contacts during the electoral campaigns has on the probability that voters decide to cast a tactical vote for a less preferred party when their most preferred party is out of the race. I show that these effects are of moderate size, but substantively important. For example, during the 2017 election, contact by the most preferred party discouraged tactical voting by 7.02%, while contact by the most preferred viable party encouraged it by 13.41%. Combining counterfactual simulations with Multilevel Regression and Poststratification I estimate the effect that party contact has on the seat distribution in Westminster through tactical voting. My results show that between 9 and 18 seats change hands, depending on the election. Importantly, the Conservative party would have obtained a majority in 2017 had non-viable parties given up contacting their supporters.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

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. While unobserved heterogeneity can have different origins and forms, it always poses the same problem: if ignored and correlated with the covariates of interest, it leads to biased and inconsistent estimates of the quantities of interest. One of the best ways to deal with unobserved heterogeneity is to use panel data. Panel data allows researchers to control for time-invariant unobserved heterogeneity by leveraging the information that comes from observing each individual in multiple time periods. In the following three chapters I deal with this type of unobserved heterogeneity in studies of political behavior, and propose new estimators that deal with it.

In Chapter 2, I deal with the case in which no panel data is available, but researchers have access to a series of Repeated Cross Sections (RCS). In RCS data, a random sample from the same population is taken at several points in time. A large number of social and political research data comes in this form (e.g., General Social Survey, Latin American Public Opinion Project).

To deal with these type of data in the presence of unobserved heterogeneity, I develop a semi-parametric two-step estimator, the *Local Cohorts* (LC) estimator, that allows for the estimation of linear fixed-effects with RCS data. The LC estimator relies on the assumption that, on average, individuals with similar observed time-invariant characteristics have similar (but not necessarily identical) unobserved characteristics. I show that this estimator produces root-n consistent and asymptotically normal estimates of the parameters of interest, under certain assumptions.

The first step of the LC estimator generates a synthetic panel, where each observation corresponds to the average individual with some time-invariant characteristics, as observed in the RCS data. In the second step, the LC estimator applies the within transformation to this synthetic panel to produce estimates of the model parameters.

Using Monte Carlo simulations, I compare the LC estimator with Deaton's cohort average approach (Deaton 1985) and with Moffitt's instrumental variables approach

(Moffitt 1993). My simulations show that the LC estimator typically has a smaller Root Mean Squared Error (RMSE) than the other approaches.

In addition, I study the performance of the LC estimator relative to a fixed-effects estimator from real panel data, by analyzing the effect of partisan news exposure on the favorability ratings of Obama. I find that the LC estimator outperforms alternative RCS data estimators. Moreover, while there is some bias relative to the real panel estimates, the results show that the identification assumptions of the LC estimator are plausible in this case.

In Chapter 3, I deal with a standing problem in the literature: in the case of binary outcome models (like probit or logit), consistent panel data estimators of the model parameters do not permit the estimation of partial effects and probabilities, which are usually the quantities of interest to researchers. My solutions is the *Penalized Flexible Correlated Random Effects* (PF-CRE) estimator for binary outcome models with panel data. This estimator provides consistent and efficient estimates of the model parameters and partial effects, without imposing excessive restrictions on the unobserved heterogeneity. PF-CRE relies on modeling the conditional distribution of the unobserved heterogeneity using a high-dimensional flexible specification based on observed characteristics, combined with a penalized estimation technique to reduce the dimensionality of the model and induce efficiency.

Using Monte Carlo simulations, I show that PF-CRE performs well in small samples, and that it performs better than alternative estimators such as Conditional Maximum Likelihood, pooled Logit, and the traditional Correlated Random Effects estimators. I also provide three applications where I show that the assumptions of PF-CRE hold in each case, and that controlling for unobserved heterogeneity is important in each of the three studies.

In Chapter 4, I study the effect of party contacts during the 2015 and 2017 United Kingdom General Elections on the decision of voters to cast a tactical vote. Tactical voting is defined as casting a vote for a less preferred party when the most preferred party is considered out of the race. Estimating these effects from observational data is challenging because parties may be more likely to contact those voters who are already likely to cast a tactical vote in the first place, artificially inflating the effect of party contacts. To resolve this issue, I use the PF-CRE estimator developed in Chapter 3 to account for unobserved heterogeneity in voter behavior. This unobserved heterogeneity is likely to capture whatever it is parties see in voters that makes them contact them.

My findings show that voters who are contacted by their most preferred party are 7.02 and 2.75 percentage points less likely to cast a tactical vote in 2017 and 2015, respectively. This indicates that party contacts can enforce loyalty or sincerity at the polling booth. However, being contacted by the most preferred viable party increases the probability of a tactical vote by 13.41 and 7.03 percentage points in 2017 and 2015, respectively. Through a combination of counterfactual estimation of Multilevel Regression and Poststratification, I show that these effects influence the distribution of Parliamentary seats. Among other findings, I show that if non-viable parties in each constituency gave up contacting their supporters, the Conservative party would have obtained an outright majority in the 2017 General Election.

Chapter 2

LOCAL COHORTS ESTIMATOR FOR SYNTHETIC PANELS WITH FIXED EFFECTS

2.1 Introduction

In many circumstances, there is a lack of panel data allowing specific individuals to be followed over time. In many of those situations, however, Repeated Cross-Sectional data (RCS) may be available. In RCS data, random samples from the same population are taken at several points in time, using the same (or very similar) sampling techniques and questionnaires. A substantial amount of data available for social research, in fact, comes in the form of RCS data; for example, the Current Population Survey (CPS), the Cooperative Congressional Election Study (CCES), the Latin America Public Opinion Project (LAPOP), data from the Pew Research Center, the General Social Survey (GSS), the American National Election Studies (ANES), among others.

The advantage of panel data over RCS data comes precisely from the fact that a single individual is observed at different points in time. This allows researchers to control for unobserved individual-level heterogeneity (by using fixed-effects) that aids in causal identification, or to estimate dynamic models where individuals' past behavior influences their current behavior. In contrast, RCS data has the drawback that it cannot leverage this information, because each individual is observed only once, meaning that standard estimation techniques cannot be applied.

In light of the relative abundance of RCS data compared to panel data, and the need for reliable estimates under the presence of unobserved individual-level heterogeneity, I develop a semi-parametric two-step estimator, the Local Cohorts estimator (LC), that allows for the estimation of linear fixed-effects models using RCS data. I show that, under some assumptions, this estimator produces root-n consistent and asymptotically normal estimates of the parameters of interest. Furthermore, in Monte Carlo simulations, I show that the LC estimator outperforms other available estimators in small samples.

The main assumption on which the LC estimator relies on is that similar individuals have, on average, similar unobserved individual fixed-effects, where similarity is measured with respect to a set of observable time-invariant individual-level char-

acteristics. Under this assumption, the first step of the LC estimator is to generate a synthetic panel, where each observation corresponds to the average individual with some time-invariant characteristics observed in the RCS data. The second step of the LC estimator is to apply the within transformation, common in fixed-effects models, to the synthetic panel generated in the first step in order to obtain estimates of the parameters of interest.

The LC estimator has advantages over other estimators proposed for fixed-effects models with RCS data. These estimators fall in two categories. The first, and closest to the LC estimator, is the cohort averaging approach developed by Deaton (1985). In this approach, individuals sharing common characteristics (most typically gender and year of birth) are grouped into disjoint cohorts, after which averages within these cohorts are treated as observations in a synthetic panel.[1] This stands in contrast with the LC estimator, which allows observations in the RCS data to belong to multiple groups, being weighted within each group depending on how closely they share observable time-invariant characteristics with other members of the group. This allows the LC estimator to make better use of the information available in the RCS data, compared to the arbitrary splitting of the data in Deaton's approach. The second approach, first developed by Moffitt (1993), relies on observable time-invariant characteristics and functions of time to instrument for variables in the model that are correlated (or suspected to be) with the unobserved fixed effects.[2]

Both these approaches have drawbacks. The cohort averaging approach can produce biased estimates in small samples, even after using bias-correction techniques. This is, in large part, because it cannot resolve the tension that exists between cohort sizes and the number of cohorts: for this estimator's assumptions to be tenable, there must be a large number of observations per cohort, but this necessarily implies that the number of cohorts must be small, which translates into a synthetic panel with few observations.[3] Another shortcoming is that the design of the cohorts is entirely up to the researcher, which can give rise to different estimates from the same data by two researchers seemingly using the same estimation method.

The instrumental variables (IV) approach can suffer from the weak instruments

[1] See also, Verbeek and Nijman (1992, 1992), Collado (1997), Devereux (2007), McKenzie (2004), and Ridder and Moffitt (2007), among others.

[2] See also, Ridder and Moffitt (2007) and Pelser, Eisinga, and Franses (2002, 2004), among others. Verbeek (2008) provides a short review of the literature on the cohort averaging and instrument variable approaches.

[3] Of course, if the data set contains hundreds of thousands of observations this tension is not a serious constraint.

problem, with a first step estimation that has too little variance. This then translates into a large degree of uncertainty around the estimates of the parameters of interest. This approach also requires a correct specification of the instruments' equation which, in general, is unknown. Therefore, a workable version of the IV approach requires the use of a flexible first step specification that allows for an unknown functional form. While this approach can generate consistent estimates of the parameters of interest, using a flexible specification in the first step estimation necessarily leads to larger uncertainty around the estimates of the parameters of interest.

Compared to Deaton's approach, the Local Cohorts estimator is better able to resolve the tension between cohort size and the number of cohorts because it allows observations from the RCS data to belong to multiple cohorts. This translates into smaller bias in small samples but, because of the greater complexity of the first step estimation, it can lead to a larger variance. My simulation results suggest that the reduction in bias is greater than the increase in the variance, leading to a smaller Root Mean Squared Error (RMSE) for the LC estimator. Another advantage of the LC approach is that the non-parametric nature of the first step better captures non-linearities in the relation between the unobserved fixed-effects and the observed time-invariant characteristics that are used to construct the cohorts. This also contributes to reducing the bias of the estimator in small samples.

In addition, I study the performance of the LC estimator relative to a fixed-effects estimator from real panel data. The effect of interest is the impact of partisan news exposure on the favorability ratings of Obama. In this study, I find that the LC estimator outperforms alternative RCS data estimators. Moreover, while there is some amount of bias relative to the real panel estimates, the results show that the identification assumptions of the LC estimator are plausible in this case.

The rest of this chapter is organized as follows. In Section 2.2 I present the Local Cohorts estimator and I show that it is root-n consistent and asymptotically normally distributed. In Section 2.3 I describe Deaton's and Moffitt's approaches to the estimation of fixed-effects models with RCS data. In Section 2.4 I compare the performance of the LC estimator with alternative estimators under different specifications, using Monte Carlo simulations. In Section 2.5 I estimate the effects of partisan news exposure on voters' perceptions of Obama using the LC estimator, finding them comparable to those derived from real panel data. Finally, in Section 2.6 I conclude and suggest avenues for further research.

2.2 Unobserved Heterogeneity and the Local Cohorts Estimator

Before presenting the estimation of models with unobserved individual-level heterogeneity using RCS data, it is useful to review the estimation of these models when panel data is available, as this serves as the basis for the Local Cohorts estimator developed in this chapter. Consider the following model with unobserved individual-level heterogeneity (fixed-effects):

$$y_{it} = x_{it}\beta_0 + f_i + \varepsilon_{it}, \quad t = 1, ..., T, \quad i = 1, ..., n, \tag{2.1}$$

where x_{it} denotes a k-dimensional vector of real-valued explanatory variables, $\beta_0 \in \mathbb{R}^k$ are the parameters of interest, $f_i \in \mathbb{R}$ are individual-level unobserved effects that may be corelated with x_{it}, and ε_{it} is an error term with mean 0 and variance σ^2 such that $E(\varepsilon|x, f) = 0$. Because x_{it} and f_i are (potentially) correlated, estimating β_0 via an OLS regression of y_{it} on x_{it} ignoring f_i yields biased and inconsistent estimates. When panel data is available, this problem can be solved by applying the within transformation to the model in equation 2.1, which demeans the variables for each individual. Since the unobserved individual-level heterogeneity is assumed to be constant in time, demeaning eliminates the unobserved term f_i from equation 2.1, and β_0 can be then consistently estimated via OLS on the transformed equation.

When only RCS data is available, however, each individual in the sample is only observed at a single time period (or cross-section). That is, letting i_t denote that individual i is observed at time t only, the data generating process for the model with unobserved individual-level heterogeneity can be re-written as:

$$y_{i_t t} = x_{i_t t}\beta_0 + f_{i_t} + \varepsilon_{i_t t}. \tag{2.2}$$

Precisely because in RCS data each individual is only observed at only one time period, the within transformation cannot be applied to eliminate the individual-level heterogeneity, f_{i_t}.

Local Cohorts Estimator

Consider the model in equation 2.2 and a RCS sample of $(y_{i_t t}, x_{i_t t}, z_{i_t})$, where $y_{i_t t} \in \mathbb{R}$ is the outcome of interest, $x_{i_t t} \in \mathbb{R}^k$ are the explanatory variables, and $z_{i_t} \in \mathcal{Z} \subset \mathbb{R}^d$ are additional observable individual-level characteristics. Calculating the expectation of equation 2.2 conditional on a given value z_c of the observed time-invariant characteristics results in:

$$E(y_{i_t t}|Z = z_c) = E(x_{i_t t}|Z = z_c)\beta_0 + E(f_{i_t}|Z = z_c) + E(\varepsilon_{i_t t}|Z = z_c), \tag{2.3}$$

which, by denoting $r_{ct} = E(r_{i_t t}|Z = z_c)$ for any variable r, can be written as:

$$y_{ct} = x_{ct}\beta_0 + f_{ct} + \varepsilon_{ct}. \tag{2.4}$$

Under the assumption that conditional on a value of z, the expected fixed-effects from different cross-sections are the same, that is, $f_{ct} = f_c \ \forall t$, equation 2.4 reduces to:

$$y_{ct} = x_{ct}\beta_0 + f_c + \varepsilon_{ct}. \tag{2.5}$$

By using multiple values of the time-invariant characteristics, it is possible to generate multiple (synthetic) observations as the one in equation 2.5. That is, computing equation 2.3 for values $(z_1, ..., z_C)$, with $z_c \in \mathcal{Z}$, results in a synthetic panel of size $C \times T$, where I refer to each $c = 1, ..., C$ as a local cohort.

Based on the synthetic panel generated as in equation 2.3, it is possible to estimate β_0 by applying OLS on the within transformation of equation 2.4. Thus, I define the Local Cohorts estimate of β_0 as:

$$\hat{\beta}_{LC} = (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{Y}, \tag{2.6}$$

where $\widetilde{Y}$ is the matrix form of $\widetilde{y}_{ct} = y_{ct} - \sum_{t=1}^{T} \frac{1}{T} y_{ct}$, and similarly for $\widetilde{X}$.

The challenge of the procedure described in equation 2.3 is to obtain sample estimates of the conditional expectations that are used to create the synthetic panel. They key assumption required for identification is that $E(f_{i_t}|Z = z_c) \equiv f_{ct} = f_c, \ \forall t$; that is, individuals who are similar on observables, have on average the same unobserved individual-level heterogeneity. For this assumption to hold in practice, it is necessary for each local cohort to have a sufficiently large number of observations. In that case, it is reasonable to say that the average individual with the a particular value of the observed time-invariant characteristics has the same average fixed-effect at different time periods. Given that the idea is to generate a local cohort for each individual observed in the first cross-section of the RCS data, each local cohort at each point in time will typically contain only one individual such that $z_{i_t} = z_c$.[4] Thus, using simple averaging to approximate the conditional expectations will not work, as it is unjustifiable to say that the average fixed-effect for each cohort will be constant across time with only one observation in each local cohort.

To solve this issue, I use non-parametric estimation to generate the synthetic panel. The idea is that individuals in a neighborhood of z_c also are similar in terms of

[4] Except when z takes on discrete values.

fixed-effects. Hence, in order to capture the average fixed-effect for the local cohort defined by z_c, I average the observations in the neighborhood of z_c. For this approach to be valid, it requires a slight modification of the fixed-effects stability assumption described earlier. In particular, it requires that $E(f_{i_t}|Z = z_c) = \psi(z_c)$, where $\psi(\cdot)$ is a continuous function of z_c. That is, it requires that, on average, the fixed-effect of individuals who are similar in terms of their observable time-invariant characteristics, be also similar, with the relation being stable across time.

More formally, I propose a two-step semi-parametric estimator of the parameter of interest, β_0, which I call the Local Cohorts estimator (LC). In the first (non-parametric) step, the conditional expectations of $y_{i_t t}$ and $x_{i_t t}$ on z_c (denoted y_{ct} and x_{ct}) are obtained using kernel estimators:

$$y_{ct} = \frac{\sum_{i_t=1}^{n_t} \mathcal{K}\big(S^{-1/2}(z_{i_t} - z_c)/h\big) y_{i_t t}}{\sum_{i_t=1}^{n_t} \mathcal{K}\big(S^{-1/2}(z_{i_t} - z_c)/h\big)}$$

and

$$x_{ct} = \frac{\sum_{i_t=1}^{n_t} \mathcal{K}\big(S^{-1/2}(z_{i_t} - z_c)/h\big) x_{i_t t}}{\sum_{i_t=1}^{n_t} \mathcal{K}\big(S^{-1/2}(z_{i_t} - z_c)/h\big)},$$

where $\mathcal{K}(u)$ denotes a kernel function, h is a bandwidth, and S is the variance-covariance matrix of the time-invariant covariates z.[5]

The second (parametric) step uses the estimated values y_{ct} and x_{ct} for multiple values of z_c as the input in a within estimator for fixed-effects. That is, I estimate the finite dimensional parameter vector β_0 by:

$$\widehat{\beta}_{LC} = \arg\min_{\beta \in \mathcal{B}} \sum_{c=1}^{C} \sum_{t=1}^{T} \frac{1}{2} (\widetilde{y}_{ct} - \beta \widetilde{x}_{ct})^2, \tag{2.7}$$

where $\widetilde{y}_{ct} = y_{ct} - \frac{1}{T} \sum_{t=1}^{T} y_{ct}$ is the within transformation of y_{ct}, and similarly for $\widetilde{x}_{ct}$.

While any choice of z_c values is valid, I advocate the use of those z actually observed in the first-cross section of the RCS data to generate the synthetic panel. This has two practical advantages relative to arbitrary user-specified values. First, using values of z actually observed in the sample ensures that the estimates incorporate the sample

[5] This is exactly the same specification as the multidimensional Naradaya-Watson non-parametric regression estimator.

distribution of these covariates, ensuring a representative sample. Second, it reduces the amount of extrapolation necessary in the first step estimates, as it guarantees that there exists at least one observation with said value.

Asymptotic Properties of the Local Cohorts Estimator

To establish the asymptotic properties of the Local Cohorts estimator, it is useful to define some notation first. Let $(\widetilde{Y}, \widetilde{X})$ denote the matrix form of the within transformation of the estimates from the first step, $(\widetilde{Y}_0, \widetilde{X}_0)$ their population values; finally, $G_n(\beta, \widetilde{Y}, \widetilde{X}) = \sum_{c=1}^{C} \sum_{t=1}^{T} g(\beta, \widetilde{y}_{ct}, \widetilde{x}_{ct})$, where $g(\beta, \widetilde{y}_{ct}, \widetilde{x}_{ct}) = \frac{1}{2}(\widetilde{y}_{ct} - \beta\widetilde{x}_{ct})$, the squared loss function. Notice that both $(\widetilde{Y}, \widetilde{X})$ and $(\widetilde{Y}_0, \widetilde{X}_0)$ are functions of z.

In what follows, I first formally define the assumptions I use to establish the asymptotic properties of the estimator and then discuss their meaning and the role they play in the proofs.

A1. $E(\varepsilon|z, x) = 0$, with ε i.i.d. with mean 0 and variance 1.

A2. For all t, $E(f_{i_t}|Z = z_c) = \psi(z_c)$, where $\psi(\cdot)$ is a continuous function of z_c.

A3. $|\mathcal{K}(u)| < \infty$, and $\int_{\mathbb{R}^d} |\mathcal{K}(u)|du < \infty$, and $\mathcal{K}(u)$ is symmetric, and $\int |u|^2|\mathcal{K}(u)|du < \infty$.

A4. For some $\Lambda_1 < \infty$ and $L < \infty$, either $\mathcal{K}(u) = 0$ for $||u|| > L$ and for all $u, u' \in \mathbb{R}^d$, $|\mathcal{K}(u) - \mathcal{K}(u')| \leq \Lambda_1||u - u'||$, or $\mathcal{K}(u)$ is differentiable, $|\frac{\partial}{\partial u}\mathcal{K}(u)| \leq \Lambda_1$, and for some $\nu > 1$, $|\frac{\partial}{\partial u}\mathcal{K}(u)| \leq \Lambda_1||u||^{-\nu}$, for $||u|| > L$.

A5. $f(z)$, $f(z) \times E(y|Z = z)$, and $f(z) \times E(x|Z = z)$ are uniformly continuous and bounded, where $f(z)$ is the density of z.

Assumption A1 is simply a strict exogeneity assumption, common in regression analysis, and I assume the variance of ε to be one to simplify notation.[6] Assumption A2 is the key identifying assumption. It states that, on average, the fixed-effects of individuals who are similar to each other in terms of their observable time-invariant characteristics, is similar across those individuals. Assumption A3 imposes some mild restrictions on the kernel $\mathcal{K}(\cdot)$. Assumption A4 requires that the kernel $\mathcal{K}(\cdot)$ be sufficiently smooth, in this case that it either has truncated support and is Lipschitz

[6] This could be weakened to $E(\varepsilon[x\ z]) = 0$.

continuous, or that it has a bounded derivative with an integrable tail.[7] Assumptions like A4 are common in the literature of semi-parametric two-step M-estimators.[8] Finally, assumption A5 requires that the expectations of y and x conditional on z be continuous.

To show the consistency of the Local Cohorts estimator, it is first useful to establish the following uniform convergence in probability result:

Proposition 1 (Uniform Convergence) *Under assumptions A1-A5, and* $h \to 0$ *"fast enough,"*

$$\sup_{\beta \in \mathcal{B}} |G_n(\beta, \widetilde{Y}, \widetilde{X}) - E(g(\beta, \widetilde{Y}_0), \widetilde{X}_0)| \xrightarrow{p} 0$$

Proof. See Appendix.

The proof of this proposition relies on using the triangle inequality to separate $|G_n(\beta, \widetilde{Y}, \widetilde{X}) - E(g(\beta, \widetilde{Y}_0), \widetilde{X}_0)|$ into a term that depends on the estimated functions of the conditional expectation of y and x with respect to z, $(\tilde{Y}, \tilde{X})$, and a term that depends on the population functions of the conditional expectations, $(\tilde{Y}_0, \tilde{X}_0)$. The latter term can be shown to converge in probability to zero by standard results for M-estimators (see, for example, Newey and McFadden 1994) as it does not depend on the kernel estimates. The former term relies on uniform convergence of kernel estimators coupled with the smoothness of $g(\cdot)$ to prove that it also converges in probability to zero.[9] The requirement that $h \to 0$ "fast enough" means that $O_p((\frac{\ln(n)}{nh^d})^{1/2} + h^2) = o_p(1)$; this requirement is not particularly stringent, and it is satisfied by the optimal bandwidth for the Naradaya-Watson non-parametric regression estimator, $h \propto n^{-1/(d+4)}$.[10]

Proposition 2 (Consistency) *Under assumptions A1-A5, and* $h \to 0$ *"fast enough," the Local Cohorts estimator* $\widehat{\beta}_{LC}$ *is consistent:*

$$\widehat{\beta}_{LC} \xrightarrow{p} \beta_0$$

[7] Hansen (2008)) notes that most commonly used kernels satisfy this assumption, including the polynomial kernel class, the higher order polynomial kernels of Muller (1984) and Granovsky and Muller (1991), the normal kernel, and the higher order Gaussian kernels of Wand and Schucany (1990) and Marron and Wand (1992).

[8] See, for example, Newey (1994), Escanciano, Jacho-Chávez, and Lewbel (2012, 2014), Mammen, Rothe, and Schienle (2012), and Hahn and Ridder (2013).

[9] See, for example, Andrews (1995), Fan and Yao (2003), and Hansen (2008).

[10] See, for example, Pagan and Ullah (1999), Chapter 3.

Proof. See Appendix.

The proof of this proposition relies on the consistency of M-estimators (see, for example, Theorem 2.1 in Newey and McFadden (1994). Assumptions A1 and A2 provide the identification conditions, so that the population parameter vector β_0 maximizes the population objective function $E(g(\beta, \widetilde{Y}_0, \widetilde{X}_0))$. Proposition 1 establishes that the sample analog of the objective function is asymptotically arbitrarily close to its population version, so that the maximum of $G_n(\beta, \widetilde{Y}, \widetilde{X})$ converges to the maximum of $E(g(\beta, \widetilde{Y}_0, \widetilde{X}_0))$, which is β_0.

Proposition 3 (Asymptotic Normality) *Under assumptions A1-A5, and $h \to 0$ "fast enough,"*

$$\sqrt{nT}\left(\hat{\beta}_{LC} - \beta_0\right) \xrightarrow{d} \mathcal{N}(0, V_0^{-1}\Sigma_0 V_0^{-1}),$$

where

$$V_0 = \frac{\partial^2}{\partial \beta^2} g(\beta_0, \widetilde{Y}_0, \widetilde{X}_0) = \widetilde{X}_0' \widetilde{X}_0$$

and

$$\Sigma_0 = E(\Gamma_0' \Gamma_0),$$

where

$$\Gamma_0(z) = \left[\frac{\partial}{\partial \beta} g(\beta_0, \widetilde{Y}_0, \widetilde{X}_0)\right] + \sum_{v=\widetilde{y},\widetilde{x}_1,..,\widetilde{x}_d} D_v\left[\frac{\partial}{\partial \beta} g(\beta_0, \widetilde{Y}_0, \widetilde{X}_0)\right]\left[v - v_0\right],$$

where D_v denotes the first-order derivative with respect to $v = \widetilde{y}, \widetilde{x}_1, ..., \widetilde{x}_k$.

Proof. See Appendix.

The proof of this proposition comes from verifying the conditions for Theorem 3.2 in Ichimura and Lee (2010). V_0^{-1} is the standard variance estimator for the case in which the regressors used in the second step are known. Σ_0 incorporates the effects of the first (non-parametric) step estimation in the final estimate for β_0. Notice that if there were no first step estimation, the second term in $\Gamma_0(z)$ would be zero, and then Σ_0 would simplify to $\widetilde{X}_0' \widetilde{X}_0$. In that case, the overall variance of the estimator simplifies to $(\widetilde{X}_0' \widetilde{X}_0)^{-1}$, which is the standard variance for the fixed-effects estimator when the regressors are known (instead of being generated by the first step of the estimation).

The proofs of consistency and normality of the LC estimator rely on the time-invariant characteristics being continuous. However, discrete variables are often

available (for example, gender). Incorporating these types of variables to the LC estimator is nonetheless straightforward. The estimation in the first step must simply be conducted for each value of the discrete variable or variables. Moreover, the availability of discrete variables for use in the first step can reduce the small sample bias of the estimator by reducing the size of the bandwidth required for the continuous time-invariant characteristics.

Intuitively, the first stage of the LC estimator aggregates similar individuals (in terms of their observed time-invariant characteristics) to create the synthetic panel used in the second stage. The time-invariant characteristics, z, operate similarly to instruments; that is, the time-varying elements of the model, (y, x), are projected to z as in the first stage of instrumental variables estimation. Just like IV estimators, the better the prediction of the first (nonparametric) step in LC, the better its performance. Better prediction in the first step implies that the local cohorts capture a significant proportion of the variation in the original data, reducing information loss due to aggregation.

2.3 Alternative Estimators

Cohort Averaging Approach

The first approach to estimating fixed-effects models with RCS data is the cohort averaging approach of Deaton (1985). This approach uses cohorts to generate a synthetic panel that is then used to estimate β_0, similarly to the Local Cohorts estimator. In Deaton's approach cohorts are defined as a partition of space of observed time-invariant characteristics $\mathcal{Z} = \bigcup_{c=1}^{C} Z_c$, where $Z_c \cap Z_{c'} = \emptyset$ for all c and c'. For example, if $\mathcal{Z}$ includes gender and year of birth, cohorts in this approach may be defined as disjoint groups by gender and decade of birth.[11] Deaton then averages the RCS observations by cohort and time, obtaining the following equation based on equation 2.2:

$$\bar{y}_{ct} = \bar{x}_{ct}\beta_0 + \bar{f}_{ct} + \bar{\varepsilon}_{ct}, \tag{2.8}$$

where $\bar{y}_{ct} = \sum_{i_t=1}^{n} y_{i_t t}\mathbb{I}(z_{i_t} \in Z_c) / \sum_{i_t=1}^{n} \mathbb{I}(z_{i_t} \in Z_c)$, the average of the $y_{i_t t}$s that belong to cohort c observed in cross-section t, and similarly for $x_{i_t t}$, f_{i_r} and $\varepsilon_{i_t t}$. This results in a synthetic panel with observations on C cohorts across T periods. To estimate

[11] Browning, Deaton, and Irish (1985) use cohorts of households defined on the basis of five year bands and whether the heads of the household is a manual or non-manual worker; Blundell, Duncan, and Meghir (1998) use year of birth intervals of 10 years, interacted with two education groups; Banks, Blundell, and Preston (1994) use five year age groups; Propper, Rees, and Green (2001) use seven birth groups and ten regions to build their cohorts.

β_0 from equation 2.8, Deaton assumes that $\bar{f}_{ct} = f_c$ for all c, which implicitly states that the unobserved individual-level heterogeneity actually has the form of group effects (with some noise around each group). This assumption is related to assumption A2 in the Local Cohorts estimator, but it uses a coarser aggregation. While the continuity assumption in the Local Cohorts estimator establishes a local restriction, Deaton's assumption is about larger sets. Under that assumption, the within transformation can be applied on equation 2.8 to obtain an estimate of β_0 by OLS:

$$\widehat{\beta}_D = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{x}_{ct} - \bar{x}_c)' \right)^{-1} \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{y}_{ct} - \bar{y}_c) \right), \tag{2.9}$$

where $\bar{x}_c = \sum_{t=1}^{T} \frac{1}{T} \bar{x}_{ct}$, and similarly for $\bar{y}_c$.

The asymptotic behavior of this estimator can be obtained using several alternative asymptotic sequences; this is because in addition to the two dimensions in panel data (n and T), there are two other dimensions in cohort models: the number of cohorts C and the size of the cohorts n_c.[12]

Deaton (1985) also proposes a related estimator that performs better in finite samples with smaller cohort sizes (although cohort sizes must still tend to infinity for consistency, to ensure that $\bar{f}_{ct} = f_c$). Considering the cohort averages $\bar{y}_{ct}$ and $\bar{x}_{ct}$ to be measurements of population values $y^{\star}_{ct}$ and $x^{\star}_{ct}$ with errors, he proposes to use an errors-in-variables model, in which the measurement errors are distributed with mean zero and independently of the true values; that is:

$$\begin{pmatrix} \bar{y}_{ct} \\ \bar{x}_{ct} \end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} y^{\star}_{ct} \\ x^{\star}_{ct} \end{pmatrix}; \begin{pmatrix} \sigma_{00} & \sigma' \\ \sigma & \Sigma \end{pmatrix} \right). \tag{2.10}$$

With estimates of Σ and σ in equation 2.10, the within estimator can be adjusted to eliminate the variation due to measurement error:

$$\hat{\beta}_{DE} = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{x}_{ct} - \bar{x}_c)' - \tau\hat{\Sigma} \right)^{-1} \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{y}_{ct} - \bar{y}_c) - \tau\hat{\sigma} \right), \tag{2.11}$$

[12] Deaton (1985), Verbeek and Nijman (1993), and Collado (1997) assume that the number of cohorts tends to infinity, with cohort sizes held (roughly) constant (which implies that the number of individuals tends to infinity). Moffitt (1993) and Verbeek and Vella (2005) consider the case in which the number of individuals tends to infinity while the number of cohorts is held fixed (thus cohort sizes tend to infinity). McKenzie (2004) considers the case in which $T \to \infty$ and the cohort sizes tend to infinity.

where $\hat{\Sigma}$ and $\hat{\sigma}$ are estimators of Σ and σ that can be obtained from the data, and $\tau = 1$ in Deaton (1985), but Verbeek and Nijman (1993) show that $\tau = (T-1)/T$ has better small sample properties.[13]

The cohort averaging approach to estimating fixed-effects models with RCS data relies on the assumption that, on average, the fixed-effect for a given cohort is the same for all cross-sections under analysis. In practical terms, this requires that in each cohort there is a sufficient number of observations such that the average is somewhat close to the population value, even when using the estimator in equation 2.11. That is, each cohort needs to average across a reasonably large number of observations from the the RCS data. In empirical applications, Browning, Deaton, and Irish (1985) use cohort sizes of about 190 individuals, while Blundell, Browning, and Meghir (1994) use cohort sizes of around 500. More recently, Devereux (2007) argues that cohort sizes should be much larger than that, possibly over 2,000 individuals. The need for large cohorts puts a strain on the data, as larger cohorts necessarily imply fewer cohorts overall, which translates into a synthetic panel with few observations. This may not be a very strong restriction for the CPS data used in many of the empirical applications of this method, as cross-sectional sizes are around 150,000 individuals. But for smaller sample sizes, it creates a tension between the bias generated by small cohort sizes and the uncertainty over the estimates generated by having few observations in the synthetic panel.

Another significant shortcoming of the cohort averaging approach is that the way cohorts are constructed is important. A priori, there is no guidance as to how to define a cohort, which can lead to different researchers working with the same data to use completely different cohorts and obtain different estimates of the quantities of interest. Besides the arbitrary definition of cohorts, Deaton's approach can also lead to some groupings that are not entirely reasonable. For example, suppose cohorts are defined by gender and decade of birth. One such cohort might comprise men born between 1960 and 1969, and another men born between 1970 and 1979. This way of constructing cohorts implies that a man born in 1969 has more in common (in terms of his unobserved individual-level characteristics) with a man born in 1960 than with one born just a year after him, in 1970. The use of disjoint cohorts always allows for unreasonable groupings such as this one, regardless of the specifics of how cohorts are defined.

[13] For more details on this estimator, please see Deaton (1985), Verbeek and Nijman (1993), and Ridder and Moffitt (2007).

The Local Cohorts estimator developed in this chapter significantly reduces the shortcomings of the cohort averaging approach. First, by using a local definition of cohort and by allowing observations to belong to multiple cohorts, it avoids the tension between cohort size and the number of cohorts that is inherent to Deaton's approach. This implies that the Local Cohorts estimator is better able to capture the underlying unobserved individual-level heterogeneity, especially in smaller samples where this tension is more apparent. Second, by defining cohorts based on neighborhoods whose size depends on the characteristics of the data and can overlap with each other, the Local Cohorts estimator avoids the arbitrariness that cohorts in Deaton's approach have.

Instrumental Variables Approach

Moffitt (1993) proposes to use instruments to solve the omitted variable bias that comes with the unobserved fixed-effects. As Deaton's approach, it relies on a vector of observed time-invariant individual-level characteristics $z_{i_t} \in \mathcal{Z} \subset \mathbb{R}^d$. Moffitt also considers a vector $w_{i_t t} \in \mathcal{W} \subset \mathbb{R}^m$ of time-varying variables that are uncorrelated with the fixed effects f_{i_t}. These variables $w_{i_t t}$ may simply consist of functions of t. Moffitt's IV approach is based on the following two equations:

$$x_{i_t t} = \delta_1 w_{i_t t} + \delta_2 z_{i_t} + e_{i_t t} \tag{2.12}$$

and

$$f_{i_t} = z_{it}\gamma + \nu_{i_t}. \tag{2.13}$$

To estimate β_0 with this approach, one first obtains the predicted values $\widehat{x}_{i_t t}$ from equation 2.12, as in any IV approach and the estimates the following equation via OLS:

$$y_{i_t t} = \widehat{x}_{i_t t}\beta + z_{i_t}\gamma + \nu_{i_t} + \varepsilon_{i_t t}. \tag{2.14}$$

Letting y, X, Z, and ν be the stacked $n \times T$ vectors for all i and t, and defining $U = [X\ Z]$ and $\widehat{U} = [\widehat{X}\ Z]$, the IV estimator of β_0 is consistent if $\text{plim}\frac{1}{nT}\widehat{U}'\nu = 0$ (and $\widehat{U}$ is of full rank, $d + m$), which can be achieved as n goes to infinity, holding T fixed. Notice that the assumption that $\text{plim}\frac{1}{nT}Z'\nu = 0$ is similar to Deaton's assumption that the average of the individual fixed-effects is time invariant. In fact, if z is defined as cohort dummies interacted with time, Moffitt's and Deaton's estimators are identical (see Verbeek 2008).

A clear shortcoming of Moffitt's approach is that it requires a correct specification of the first stage model. For example, if the fixed-effect is related to the observed time-

invariant characteristics via a quadratic function, but the model is specified with a linear function, the assumption that $Z'\nu$ approaches zero as n goes to infinity will not hold. This problem can potentially be addressed by using a flexible functional form in the first stage estimation, but this incurs in efficiency costs. Moreover, despite the attractiveness of using time-varying covariates, $w_{i_t t}$, that are unrelated to the fixed-effects, finding variables that are unrelated to the fixed-effect is hard to do.

2.4 Monte Carlo Simulations

To study the small sample properties of the Local Cohorts estimator, I conduct a series of Monte Carlo simulations and analyze the bias, standard error, and root mean squared error (RMSE) of the Local Cohorts estimator, Moffitt's IV estimator, and two versions of Deaton's cohort averaging approach with different cohort sizes.[14]

For these simulations I consider the following data generating process:

- $y_{it} = 2x_{it} + f_i + \varepsilon_{it},\ \varepsilon_{it} \sim \mathcal{N}(0, 10)$
- $x_{it} = \frac{1}{2}f_i^2 + f_i + \nu_{it},\ \nu_{it} \sim \mathcal{N}(0, 5)$, so that x and f have a non-linear relation.

I consider three cases in terms of the relation between the time-invariant characteristics z and the fixed-effect f:

- **Base Case:** $\mathcal{Z} \subset \mathbb{R}^2,\ z_1 \sim \mathcal{U}(-15, 15),\ z_2 \sim \mathcal{N}(0, 2)$, and $f_i = z_{1i}\sin(z_{1i}/6) + z_{2i} + \eta_i,\ \eta_i \sim \mathcal{N}(0, 2)$.
- **Group Effects Case:** $\mathcal{Z} \subset \mathbb{R}^2,\ z_1 \sim \mathcal{U}(-15, 15),\ z_2 \sim \mathcal{N}(0, 2)$, and $f_i = \frac{1}{2}\left(\sum_{q=1}^{5}\mathbb{I}(z_{1i} \in q(z_1)) + \sum_{q=1}^{5}\mathbb{I}(z_{2i} \in q(z_2))\right)$, where q denotes the quintiles of z and $q(z)$ is the set that indicates that an observation belongs to the qth quintile. Therefore, in this case, f is simply the average of the quintiles to which the observation belongs to in terms of z_1 and z_2, taking values on the set $\{1, \frac{3}{2}.2, \frac{5}{2}, 3, \frac{7}{2}, 4, \frac{9}{2}, 5\}$.

[14] For Moffitt's estimator I use a polynomial in z for equations 2.12 and 2.13 to capture the non-linearities in the data generating process for both x and f with respect to z. For Deaton's estimator I use versions that have cohorts of approximate size 50 and 200. These cohorts are created by splitting the data along percentiles of the observable time-invariant covariates such that each cohort has (approximately) the same size.

- **Underspecified Case:** $\mathcal{Z} \subset \mathbb{R}^3$, $z_1, z_2, z_3 \sim \mathcal{N}(0, 2)$, with $corr(z_1, z_2) = corr(z_1, z_3) = 0.2$ and $corr(z_2, z_3) = -0.3$, with z_3 unobserved, and $f_i = z_{1i} \sin(z_{1i}/6) + z_{2i} + z_{3i} + \eta_i$, $\eta_i \sim \mathcal{N}(0, 2)$.[15]

The Base Case satisfies all the assumptions for consistency and asymptotic normality of the Local Cohorts estimator. The Group Effects Case satisfies the assumptions in Deaton's estimator, but represents a mild violation of the continuity assumption (A2) of the Local Cohorts Estimator. The violation is mild because it only occurs at a finite number of points (16). This case is included to show that, even when the continuity assumption does not hold, the Local Cohorts estimator can still perform relatively well. The Underspecified Case incorporates an additional time-invariant characteristic that affects the fixed-effect and is correlated with the other time-invariant characteristics but that is not observed. This case is designed to show that an incomplete specification of the first step model does not affect the Local Cohorts estimator's main properties, but it compromises the validity of Moffitt's approach.[16]

For all cases, I draw samples for two time periods, $T = 2$, with cross-sectional sizes between 250 and 5,000, in increments of 250. Unless otherwise noted, for each sample size and case, I draw a total of 200 Monte Carlo samples.

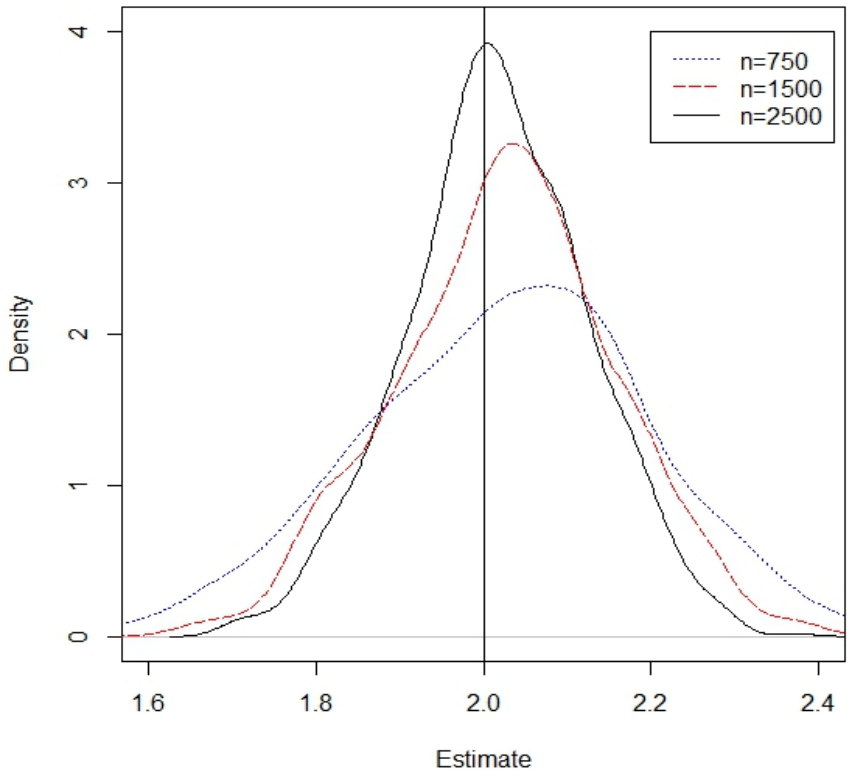
Figure 2.1 shows 1,000 estimates from the Local Cohorts estimator for the Base Case with three different cross-sectional sample sizes: 750, 1,500, and 2,500. As can be seen from the figure, when the sample size is relatively small the distribution of the estimates is not centered around the true value of the coefficient (in this case, β_0=2), and the distribution is skewed and is slightly platykurtic. When the sample size is larger, like in the case with 2,500 observations per cross-section, the distribution of the estimates is centered at the true value of the coefficient, and has almost no skewness and is mesokurtic, meaning that for moderate samples sizes a normal distribution describes the distribution of the estimates sufficiently well.

Figure 2.2 (and Table 2.B1 in the Appendix) shows the Monte Carlo simulation results for the Base Case. As can be seen from panel (a), the Local Cohorts estimator has a smaller RMSE than the alternative estimators for all sample sizes.

[15]The distribution of z_1 in the Underspecified case is different from the distribution used for this variable in the other cases. This was done to simplify the creation of non-independent variables. If the simulations for the Base Case and the Group Effects Case were run with z_1 distributed as in the Underspecificed Case, the results would remain qualitatively the same, although with smaller biases for Deaton's and the Local Cohorts estimators.

[16]However, because of increased unexplained variability, the Local Cohorts estimator will have a higher variance and may need more observations to sufficiently reduce the small sample bias.

Figure 2.1: Distribution of LC for Different Sample Sizes: Base Case Simulations



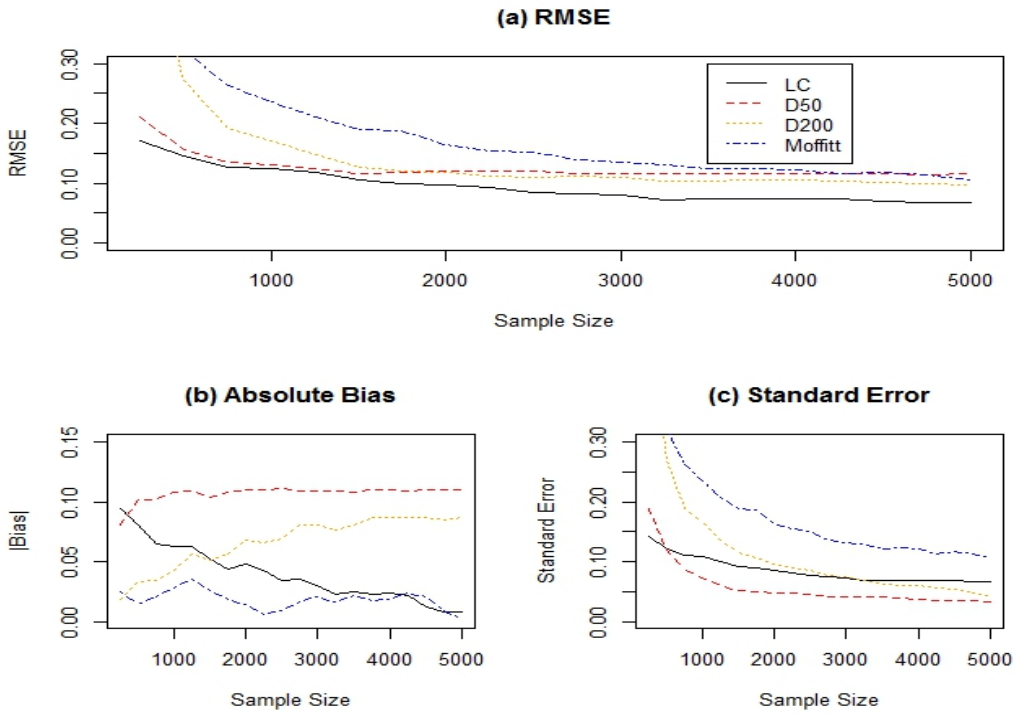
The data generating process in the Base Case satisfies all the assumptions in the Local Cohorts estimator. The vertical line at 2 indicates the true value of the parameters β_0.

Deaton's estimator with cohort size 50 is the next best one in terms of RMSE in the smaller samples, but it is defeated by Deaton's estimator with cohort size 200 for moderate to larger samples. Moffitt's IV estimator is the one with the largest RMSE, except for the larger sample sizes where it is better than Deaton's with cohort size 50.

In terms of the bias, shown in panel (b), Moffitt's IV estimator is clearly superior to all others in small and moderate samples, and performs similarly well to the Local Cohorts estimator for the larger samples.[17] Both versions of Deaton's estimator have

[17]The large variability in the bias of Moffitt's IV estimator is due to the rather small number of Monte Carlo simulations coupled with the large standard error of this estimator. If the Monte Carlo simulations included a significantly larger number of simulations, this bias should be a flat line, approximately at zero. The use of a relatively small number of simulations is due to the computational demands of the LC estimator.

Figure 2.2: Base Case Simulations



The data generating process in the Base Case satisfies all the assumptions in the Local Cohorts estimator. For reference, the true value of the parameter of interest is $\beta_0 = 2$. RMSE, Absolute Bias, and Standard errors are in units (not as percentages of β_0). LC refers to the Local Cohorts estimator; D50 and D200 refer to Deaton's estimator with cohort sizes of approximately 50 and 200 individuals, respectively; and Moffitt refers to Moffitt's estimator implemented with a flexible polynomial to account for the unknown data generating process' functional form.

rather large biases, and these biases do not decrease in time. The reason for this poor performance in terms of bias is partly due to the inflexibility of Deaton's approach, which requires either the fixed-effects for individuals who are very different in terms of time-invariant characteristics to be the similar (the partition of $\mathcal{Z}$ is too coarse) or, when it avoids that, it uses too few observations in each cohort that cannot ensure that the estimates of the fixed-effect for a given cohort are stable across time (the partition of $\mathcal{Z}$ is too fine-graded).

Panel (c) of Figure 2.2 shows the standard error of each of the estimators for the different sample sizes. Moffitt's IV estimator has the largest standard error of all the estimators considered, which explains why, despite having the smallest bias, it performs poorly in terms of RMSE. The Local Cohorts estimator is somewhere in between the two versions of Deaton's estimator, and has a larger standard error

than them for the larger samples. To get a better idea of what these standard errors mean, it is useful to compare them to the standard error from the estimates of an real panel. For a sample size of 5,000 observations per time period, the Local Cohorts estimator has a standard error that is 32% larger than that of the real panel (with the same data generating process), whereas Moffitt's IV estimator has a standard error that is 108% larger than the panel's.

Overall, the Local Cohorts estimator performs better than the other estimators considered. Although Deaton's estimators can have smaller variances, they are biased; and while Moffitt's IV estimator is the one with the smallest bias, it has a much larger variance.

Figure 2.3 (and Table 2.B2 in the Appendix) show the Monte Carlo results for the Group Effects Case, which satisfies the assumptions in Deaton's approach but violates the continuity assumption (A2) of the local cohorts estimator at 16 distinct points.[18] As expected, Deaton's estimator with cohort size 50 outperform the other estimators in terms of RMSE.

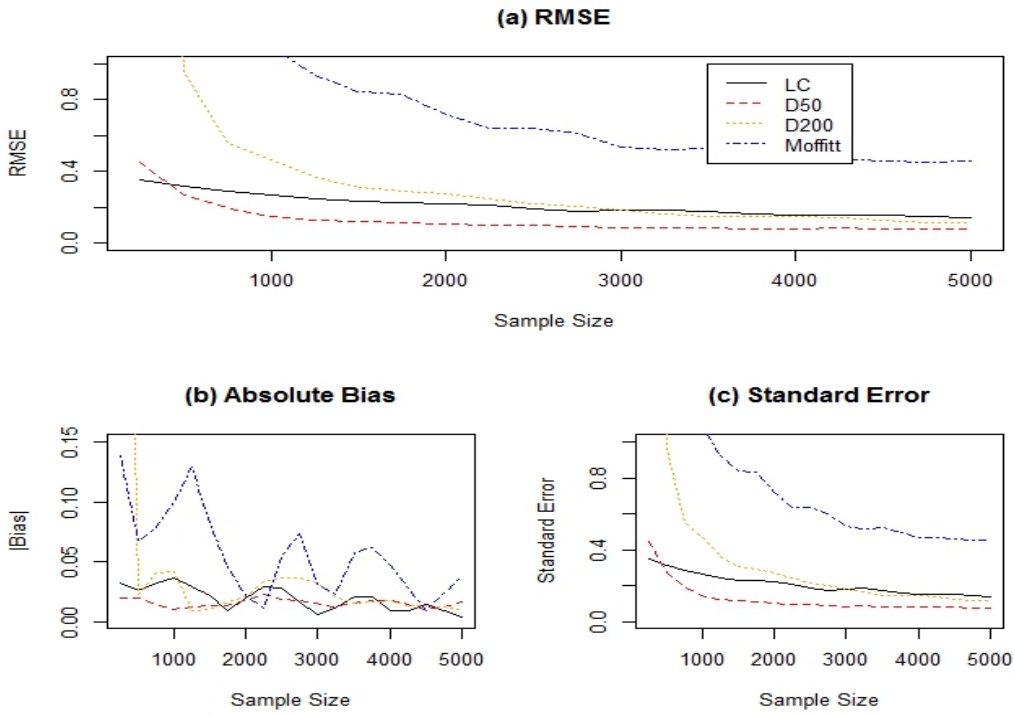
With respect to the bias, both the Local Cohort estimator and Deaton's perform similarly well, with very small biases. Moffitt's estimator has the largest bias of all estimators in this case. This bias is due to the fact that the flexible polynomial used in the first step of this estimator is not adequate to deal with the discontinuities present in the relation between the unobserved individual-level heterogeneity (f) and the observed individual-level characteristics (z). Panel (c) shows that it is in terms of the standard error that Deaton's estimators outperform the Local Cohorts estimator. The main reason for the better performance of Deaton's estimator in this case is that the relation between f and z follows a step function. This significantly helps Deaton's approach, since it assumes that the fixed-effects are actually group effects, which matches exactly the data generating process in this case. Contrary to this, the Local Cohorts estimator does not assume that the fixed-effects follow a step function, meaning that it loses efficiency in determining that this.[19] That is, in this case, the more parsimonious nature of Deaton's approach allows for a more precise estimation, with standard errors of about half the size of the Local Cohorts estimator.

Figure 2.4 (and Table 2.B3 in the Appendix) shows the Monte Carlo simulation

[18] These points are defined by combinations of the first to fourth quintiles of z_1 and z_2.

[19] This is similar to comparing the OLS estimates of a linear regression model with nonparametric estimation of the same model.

Figure 2.3: Group Effects Case Simulations



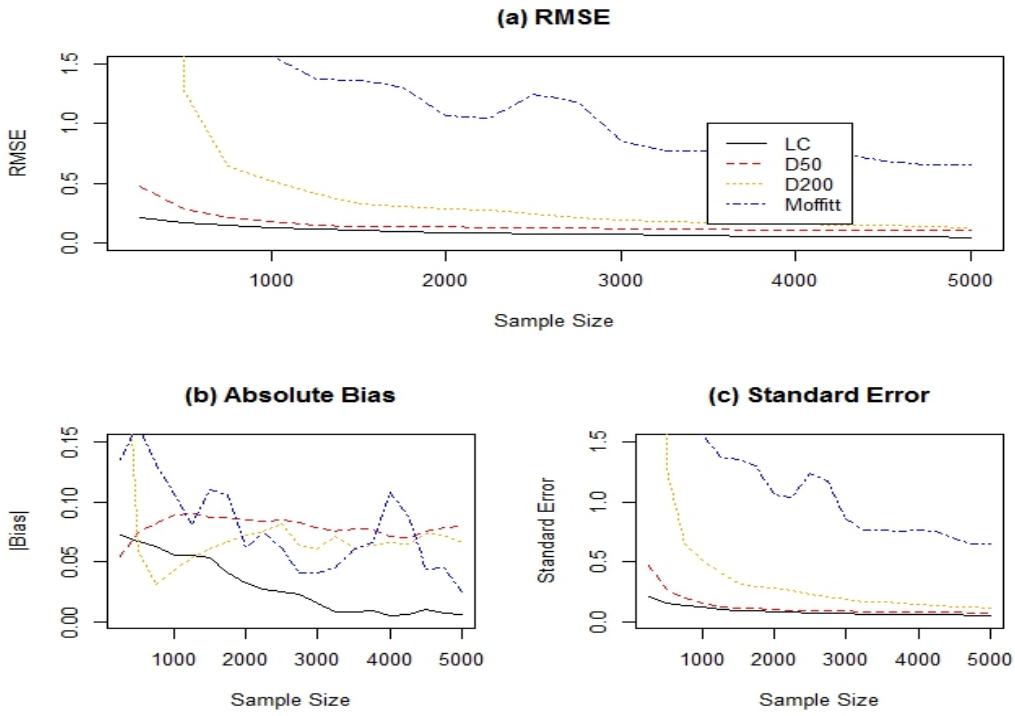
The data generating process in the Group Effect case assumes that the unobserved individual-level heterogeneity is constant (up to random noise) within groups defined by the quintiles of the observable time-invariant characteristics (which implies a mild violation of assumption A2 of the Local Cohorts estimator). For reference, the true value of the parameter of interest is $\beta_0 = 2$. *RMSE, Absolute Bias, and Standard errors are in units (not as percentages of* β_0. *LC refers to the Local Cohorts estimator; D50 and D200 refer to Deaton's estimator with cohort sizes of approximately 50 and 200 individuals, respectively; and Moffitt refers to Moffitt's estimator implemented with a flexible polynomial to account for the unknown data generating process' functional form.*

results for the Underspecified Case, in which the fixed-effects depend on three time-invariant characteristics that are correlated with one another, but only two of them are Observable. As can be seen from the simulations, the performance of Deaton's estimators and the LC estiamtor are qualitatively similar to their performance for the Base Case.[20]

The most significant change between the Base Case and the Underspecified Case is the performance of Moffitt's IV estimator. As Panel (b) of Figure 2.4 shows, this estimator is no longer unbiased. The source of its bias is the misspecification

[20]There are differences in the sizes of the bias and variance term, which are in part due to the different distribution used for z_1.

Figure 2.4: Underspecified Case Simulations



The data generating process in the Underspecified Case assumes that the unobserved individual-level heterogeneity depends on three time-invariant characteristics (and noise), one of which is unobserved. For reference, the true value of the parameter of interest is $\beta_0 = 2$. RMSE, Absolute Bias, and Standard errors are in units (not as percentages of β_0. LC refers to the Local Cohorts estimator; D50 and D200 refer to Deaton's estimator with cohort sizes of approximately 50 and 200 individuals, respectively; and Moffitt refers to Moffitt's estimator implemented with a flexible polynomial to account for the unknown data generating process' functional form.

in the first step of the estimator, derived from the unobservability of z_3. This implies that in the second step, the predicted values of x remain correlated with the unobserved heterogeneity from z_3. Both the LC and Deaton's estimator do not suffer significantly from this problem, as in both, the first step estimation integrates over the unobserved time-invariant characteristics.[21]

To summarize, the Local Cohorts estimator has some bias in the smaller samples, but this bias disappears in larger samples, even when the identification assumption (A2) is mildly violated or when there is an unobserved systematic component among the time-invariant characteristics. Deaton's estimator performs better than the Local

[21]This integration over unobserved time-invariant characteristics helps reduce the bias from misspecification, but it increases the variance of the estimates.

Cohorts estimator only when the fixed-effects are indeed group effects. However, when this is not the case, it becomes biased, even for the larger sample sizes. Moffitt's estimator generally has no bias, except when its first step is misspecified. This estimator is also very inefficient, always resulting in the highest variance of all the estimators analyzed. Taken together, the Local Cohorts estimator typically outperforms the alternative estimators analyzed here.

2.5 Comparison to Real Panel Data

In this section I compare the performance of the Local Cohorts estimator developed in this paper with fixed-effects estimates from actual panel data. To do this, I use the National Annenberg Election Study (NAES), 2008 Online Panel. The NAES panel is composed of five waves of interviews conducted over the curse of 16 months, including the pre-primary season of the 2008 Presidential Campaign, early primaries, late primaries and party conventions, general election campaign, and post election period. I limit the analysis to the 10,742 respondents who participated in all five NAES waves. I also limit the analysis to the last three waves as these are the ones that include all the variables necessary for the analysis.

The model examined in this section studies the effect of partisan news exposure on respondents' favorability ratings for the Democratic Presidential Nominee, Obama. Favorability of Obama is measured via a feeing thermometer on a scale from 1 (unfavorable views) to 100 (favorable views). To measure exposure to partisan news on television, I use a question that asks respondents to report which programs they watch regularly on television, out of multiple lists adding up to the 45 most frequently watched news programs, according to Nielsen Ratings. The ideological bias of the 45 TV programs is taken from Dilliplane (2014), who classifies them as liberal, conservative, and neutral.[22] I use several time-invariant characteristics to obtain the LC estimates. These characteristics include age, education, income, party ID, and race, and were obtained from a panel profile wave in NAES. Formally, the model of interest is:[23]

$$\text{ObamaTherm}_{it} = \beta_1 \#\text{LiberalTV}_{it} + \beta_2 \#\text{ConservativeTV}_{it} + \beta_3 \#\text{NeutralTV}_{it} + f_i + \varepsilon_{it} \tag{2.15}$$

[22] Table 2.C1 shows the classification of TV programs.

[23] Note that the model specification used in this paper is different from that used in Dilliplane (2014). That paper estimates a fixed effects model using as explanatory variables the average exposure to each type of media for each individual throughout the panel interacted with the panel waves. Thus, the model estimates by Dilliplane (2014) is more a study of individual trends, than it is a traditional fixed-effects analysis.

To study the performance of the LC estimator, I generate a sample of repeated cross-sections from the NAES panel in the following way: (1) I draw a random sample of 3,000 respondents from the first wave; (2) draw a random sample of 3,000 respondents from the second wave from among those respondents not sampled in the first wave; (3) draw a random sample of 3,000 respondents from the third wave from among those not sampled in the first or second wave. This ensures that none of the respondents in each of the cross-sections is included in the other ones, so that the resulting dataset resembles real RCS data.

The first, non-parametric, step of the LC estimator requires choosing a kernel. For this application I use a product kernel composed of (1) the identity function for party ID and race, and (2) Gaussian 4th order kernels for age, education, and income. The bandwidths for age, education, and income were selected via cross-validation on the feeling thermometer for Obama.

I compare the results from the LC estimator with those of Moffitt's IV approach, where I use the same time-invariant characteristics plus time dummies in the first stage equation, and to a linear regression that includes the time-invariant characteristics as controls. As ground truth, I use the estimates from a subpanel defined by the individuals in the first repeated cross-section described above.[24]

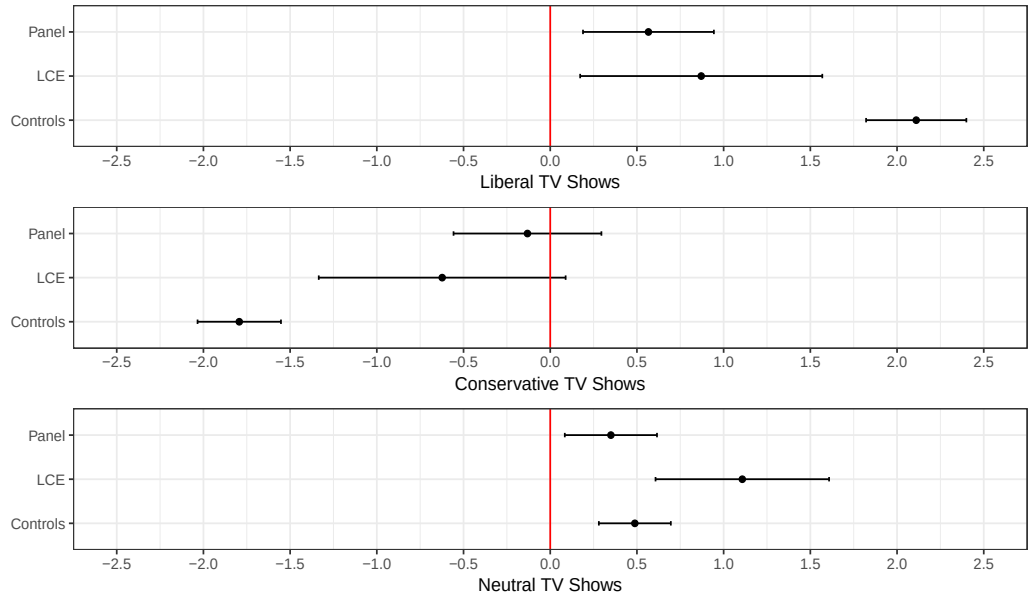
Figure 2.5 shows the estimated coefficients for the effect of the number of liberal, conservative, and neutral TV shows regularly watched by respondents on their favorability ratings of Obama estimated with the real panel, the LC estimator, and a simple OLS regression with controls. The estimates from Moffitt's IV approach are very large and off target, and are therefore reported separately in Table 2.C2 in the appendix.[25] As the figure shows, the LC estimator generally overestimates the effect of watching all three types of TV programs on the favorability ratings of Obama. While the fixed-effects estimates using the real panel predict that regularly watching an extra liberal TV show leads to an increase of 0.57 in Obama's favorability ratings, the LC estimates predict an increase of 0.87 points.[26] However, the panel and LC confidence intervals have sufficient overlap so that both estimates are statistically

[24] I do not include a version of Deaton's estimator since this estimator can be though of as a special case of the LC estimator, but where cohorts are arbitrarily defined by the researcher.

[25] The most likely reason why Moffitt's IV approach estimates are off target by such a large margin is that the number of TV shows regularly watched in each category resembles an exponential distribution with a small rate. This hampers the ability of the first stage estimates to generate good predictions of the independent variables.

[26] Note that these quantities are actually quite small. Given Obama's average rating of 51.6, it implies and increase of 1.1% or 1.68% according to the panel and LC estimates, respectively.

Figure 2.5: Coefficient Estimates from a Real Panel, LCE, and Linear Regression with Controls



Panel estimates are fixed effects estimates from a real panel of size 3000, based on the same cross-section at time 1 as the LCE and Controls estimates. Controls estimates refer to a linear model of the pooled cross-sections, including the time-invariant characteristics as controls. The variance of the LC estimator was obtained via bootstrap.

indistinguishable from one another in this case. The estimate from the linear regression with controls, which ignores the unobserved heterogeneity, predicts an increase of 2.1 points in Obama's favorability rating, which is almost 4 times the estimate from the panel. Moffitt's IV estimator instead predicts an increase of 26.9 points.

The panel estimates for the number of conservative TV programs watched shows a non-significant effect of -0.13 points, while the LC estimator shows a larger, but still non-significant estimate of -0.6 points. The linear regression with controls and Moffitt's IV estimator, on the other hand, predict a decrease in Obama's rating of 1.8 and 11.6 points, respectively, and are statistically significant.

Finally, the panel fixed-effects estimator predicts an increase of 0.35 points in Obama's rating from watching an extra neutral TV program, while the LC estimator significantly overestimates this quantity, at 1.1 points. The linear regression with controls provides an estimate that is closer to the panel estimate, at 0.49 points. Moffitt's IV estimator is, as with the other variables, significantly misestimating the

effect, at -5.7 points.

Overall, the results for the LC estimator are encouraging. While the LC estimator overestimates the effects of different TV media slants on feeling thermometers for Obama relative to the estimates from a real panel, this overestimation is not severe. Importantly, the estimates obtained with LC have smaller bias from the ground truth than those obtained with alternative estimators that do not rely on real panel data. This indicates that the identification assumptions of the LC estimator plausibly hold in this case, notwithstanding the relatively small bias encountered.

2.6 Discussion

In this chapter I develop a semi-parametric two-step estimator procedure, the Local Cohorts estimator, for estimating linear models with individual-level unobserved heterogeneity using repeated cross-sections. I provide identification conditions and derive asymptotic properties of the estimator, establishing its root-n consistency and asymptotic normality. The identification conditions require that similar individuals have, on average, similar individual-level unobserved heterogeneity. This assumption is typically weaker than the ones used by other estimators, which require that the unobserved individual-level heterogeneity be in fact a group effect with no systematic variation within each group.[27] Strictly speaking, the identification assumptions for the Local Cohorts estimator do not include group effects as a special case, although when these groups are known it can be easily modified to account for them. However, even under unknown groups, this restriction does not represent a significant shortcoming for the LC estimator relative to others, as estimators that rely on group effects have severe problems if the groups are unknown.[28]

In Monte Carlo simulations, I show that the Local Cohorts estimator performs well, with some bias in small samples that disappears in moderate sample sizes, even when the unobserved individual-level heterogeneity comes in the form of group effects. Furthermore, compared to other available estimators, it typically has a smaller RMSE, being only larger than models that assume the presence of group effects when this is indeed the case.

Beyond the Monte Carlo simulations, I compare the Local Cohorts estimator to a fixed-effects estimator derived from real panel data. The model being estimated seeks to determine the effects of partisan news exposure on the favorability ratings

[27] See, for example, Deaton (1985) and Inoue (2008).

[28] Deaton (1985) implicitly assumes that there are group effects, but does not establish what these groups are, how many there are, or how to discover them from the data.

of then presidential candidate Obama. I find that the LC estimator typically outperforms alternative models for RCS data. Moreover, while there is some amount of bias relative to the real panel estimates, the results show that the identification assumptions of the LC estimator plausibly hold in this case.

It is possible that the Local Cohorts estimator can prove a useful alternative to the fixed-effects estimators if the efficiency loses are not too large (i.e., when the RCS data is large). This is because RCS data will generally not suffer from attrition as panel data does, and it is cheaper to collect, allowing for significantly larger sample sizes that could more than compensate the efficiency loses relative to the fixed-effects estimator for panel data. Finally, the Local Cohorts could be developed into a complement of fixed-effects panel estimators to help deal with attrition and non-response at certain waves of a panel, as well as boosting the estimates by including in the estimation individuals sampled in cross-sections complementary to the panel design.

References

Andrews, Donald W. K. 1995. "Nonparametric Kernel Estimation for Semiparametric Models". *Econometric Theory* 11:560–596.

Banks, James, Richard Blundell, and Ian Preston. 1994. "Life Cycle Expenditure Allocations and the Consumption Costs of Children". *European Economic Review* 76:598–606.

Blundell, Richard, Martin Browning, and Costas Meghir. 1994. "Consumer Demand and the Life-Cycle Allocation of Household Expenditures". *Review of Economic Studies* 61 (1): 57–80.

Blundell, Richard, Alan Duncan, and Costas Meghir. 1998. "Estimating Labor Supply Responses Using Tax Reforms". *Econometrica* 66 (4): 827–861.

Browning, Martin, Angus Deaton, and Margaret Irish. 1985. "A Profitable Approach to Labor Supply and Commodity Demands over the Life Cycle". *Econometrics* 53:503–543.

Collado, M. Dolores. 1997. "Estimating Dynamic Models from Time Series of Independent Cross-Sections". *Journal of Econometrics* 82:37–67.

Deaton, Angus. 1985. "Panel Data from Time Series of Cross-Sections". *Journal of Econometrics* 30:109–126.

Devereux, Paul J. 2007. "Small Sample Bias in Synthetic Cohort Models of Labor Supply". *Journal of Applied Econometrics* 22:839–848.

Dilliplane, Susanna. 2014. "Activation, Conversion, or Reinforcement? The Impact of Partisan News Exposure on Vote Choice". *American Journal of Political Science* 58 (1): 79–94.

Escanciano, Juan Carlos, David Jacho-Chávez, and Arthur Lewbel. 2012. "Identification and Estimation of Semiparametric Two Step Models". Mimeo.

– . 2014. "Uniform Convergence of Weighted Sums of Non and Semiparametric Residuals for Estimation and Testing". *Journal of Econometrics* 178:426–433.

Fan, Jianqing, and Qiwei Yao. 2003. *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag.

Granovsky, Boris L., and Hans-Georg Muller. 1991. "Optimizing Kernel Methods: A Unifying Variational Principle". *International Statistical Review* 59:373–388.

Hahn, Jingyong, and Geert Ridder. 2013. "Asymptotic Variance of Semiparametric Estimators with Generated Regressors". *Econometrica* 81 (1): 315–340.

Hansen, Bruce E. 2008. "Uniform Convergence Rates for Kernel Estimation with Dependent Data". *Econometric Theory* 24:726–748.

Ichimura, Hidehiko, and Sokbae Lee. 2010. "Characterization of the Asymptotic Distribution of Semiparametric M-estimators". *Journal of Econometrics* 159 (2): 252–266.

Inoue, Atsushi. 2008. "Efficient Estimation and Inference in Linear Pseudo-Panel Data Models". *Journal of Econometrics* 142:449–466.

Mammen, Enno, Christoph Rothe, and Melanie Schienle. 2012. "Nonparametric Regression with Nonparametrically Generated Covariates". *Annals of Statistics* 40 (2): 1132–1170.

Marron, J. Steve, and Matthew Paul Wand. 1992. "Exact Mean Integrated Squared Errors". *Annals of Statistics* 20:712–736.

McKenzie, David J. 2004. "Asymptotic Theory for Heterogeneous Dynamic Pseudo-Panels". *Journal of Econometrics* 120:235–262.

Moffitt, Robert. 1993. "Identification and Estimation of Dynamic Models with a Time Series of Repeated Cross-Sections". *Journal of Econometrics* 59:99–123.

Muller, Hans-Georg. 1984. "Smooth Optimum Kernel Estimator of Densities, Regression Curves and Modes". *Annals of Statistics* 12:766–774.

Newey, Whitney K. 1994. "Kernel Estimation of Partial Means and a General Variance Estimator". *Econometric Theory* 10 (2): 233–253.

Newey, Whitney K., and Daniel McFadden. 1994. "Large Sample Estimation and Hypothesis Testing". In *Handbook of Econometrics, Volume 4*, ed. by R Engle and D McFadden, 2111–3155.

Pelser, Ben, Rob Eisinga, and Philip Hans Franses. 2004. "Ecological Panel Inference from Repeated Cross Sections". In *Ecological Inference: New Methodological Strategies*, ed. by Gary King, Martin A. Tanner, and Ori Rosen. New York, NY: Cambridge University Press.

– . 2002. "Inferring Transition Probabilities from Repeated Cross Sections". *Political Analysis* 10:113–133.

Propper, Carol, Hedley Rees, and Katherine Green. 2001. "The Demand for Private Medical Insurance in the UK: A Cohort Analysis". *The Economic Journal* 111:C180–C200.

Ridder, Geert, and Robert Moffitt. 2007. "The Econometrics of Data Combination". In *Handbook of Econometrics, Volume 6B*, ed. by James J. Heckman and Edward E. Leamer. North-Holland: Elsevier Science.

Verbeek, Marno. 2008. "Pseudo-Panels and Repeated Cross-Sections". In *The Econometrics of Panel Data*, ed. by Patrick Sevestre. Berlin: Springer-Verlag.

Verbeek, Marno, and Theo E. Nijman. 1992. "Can Cohort Data Be Treated as Genuine Panel Data?" *Empirical Economics* 17:9–23.

– . 1993. "Minimum MSE Estimation of a Regression Model with Fixed Effects from a Series of Cross-Sections". *Journal of Econometrics* 59:125–136.

Verbeek, Marno, and Francis Vella. 2005. "Estimating Dynamic Models from Repeated Cross-Sections". *Journal of Econometrics* 127:83–102.

Wand, Matthew Paul, and William R. Schucany. 1990. “Gaussian-based Kernels”. *Canadian Journal of Statistics* 18:197–204.

2.A Proofs

Proof of Proposition 1 (Uniform Convergence)

First, notice that by the Triangle Inequality $\sup_{\beta\in\mathcal{B}}\left|G_n(\beta,\widetilde{Y},\widetilde{X}) - E(g(\beta,\widetilde{Y}_0,\widetilde{X}_0))\right| \leq$

$$\sup_{\beta\in\mathcal{B}}\left|G_n(\beta,\widetilde{Y}_0,\widetilde{X}_0) - E(g(\beta,\widetilde{Y}_0,\widetilde{X}_0))\right| + \sup_{\beta\in\mathcal{B}}\left|(nT)^{-1}\sum\sum\left[g(\beta,\widetilde{Y},\widetilde{X}) - g(\beta,\widetilde{Y}_0,\widetilde{X}_0)\right]\right|.$$

Notice that the first term does not depend on the nonparametric first step estimator (but on its true functional form). Therefore, the first term is simply an M-estimator. Since $\mathcal{B}$ is assumed to be compact, the function $g(\cdot)$ is continuous in β, and $E(\sup_{\beta\in\mathcal{B}} g(\beta,\widetilde{Y}_0,\widetilde{X}_0)) < \infty$, this term is $o_p(1)$ by standard results of the Uniform Law of Large Numbers for M-estimators (Newey and McFadden, 1994).

Next, notice that the function $g(\cdot)$ is continuously differentiable (since it is simply a square function). Therefore, it is Lipschitz continuous. This means that $\exists\ \kappa$ such that $\left|g(\beta,\widetilde{Y},\widetilde{X}) - g(\beta,\widetilde{Y}_0,\widetilde{X}_0)\right| \leq \kappa||(\widetilde{Y},\widetilde{X}) - (\widetilde{Y}_0,\widetilde{X}_0)||$, and this holds true for all z (remember that $\widetilde{Y}$ and $\widetilde{X}$ are functions of z). Since this holds for all z, it also holds for the supremum over z. Thus:

$$\left|(nT)^{-1}\sum\sum\left[g(\beta,\widetilde{Y},\widetilde{X}) - g(\beta,\widetilde{Y}_0,\widetilde{X}_0)\right]\right| \leq \kappa \sup_{z\in\mathcal{Z}}||(\widetilde{Y},\widetilde{X}) - (\widetilde{Y}_0,\widetilde{X}_0)||.$$

Now, notice that

$$\sup_{z\in\mathcal{Z}}||(\widetilde{Y},\widetilde{X}) - (\widetilde{Y}_0,\widetilde{X}_0)|| \leq \sup_{z\in\mathcal{Z}}|\widetilde{Y} - \widetilde{Y}_0| + \sup_{z\in\mathcal{Z}}|\widetilde{x}^1 - \widetilde{x}^1_0| + ... + \sup_{z\in\mathcal{Z}}|\widetilde{x}^k - \widetilde{x}^k_0|,$$

where the superscripts in $\widetilde{x}^m$ indicate the mth column of the matrix $\widetilde{X}$

Now, each of the terms in the right hand side of the last inequality can be bounded by:

$$\sup_{z\in\mathcal{Z}}|\bar{r}_{ct} - \bar{r}_{0ct}| + \frac{1}{T}\sup_{z\in\mathcal{Z}}|\bar{r}_{ct} - \bar{r}_{0cs}|,$$

where r here stands for y and each of the k dimensions of x. Under assumptions **A3-A5**, each of the terms in the last equation is of order $O_p\left(\left(\frac{\ln(n)}{nh^d}\right)^{1/2} + h^2\right)$, which under appropriate conditions for h is $o_p(1)$ (Hansen, 2008).[29] For example, taking h to be the optimal bandwidth for the Naradaya-Watson estimator (the one that minimizes the mean integrated squared error), $h \propto n^{-1/(4+d)}$, is sufficient for

[29] Newey (1994a,1994b) also provide similar results but under stronger conditions that require z to have bounded support.

obtaining the $o_p(1)$ rate. Then, given that T is fixed, these are finite sums of $o_p(1)$ terms, and therefore:

$$\sup_{z \in \mathcal{Z}} ||(\widetilde{Y}, \widetilde{X}) - (\widetilde{Y}_0, \widetilde{X}_0)|| = o_p(1).$$

So, putting all of it together, we have that:

$$\sup_{\beta \in \mathcal{B}} |G_n(\beta, \widetilde{Y}, \widetilde{X}) - E(g(\beta, \widetilde{Y}_0, \widetilde{X}_0))| \xrightarrow{p} 0.$$

Proof of Proposition 2 (Consistency)

The consistency of the LC estimator follows from the consistency of M-estimators (see, for example, Theorem 2.1 in Newey and McFadden, 1994). The identification assumption A2, plus the exogeneity assumption A1, ensure that the function $g(\beta, \widetilde{Y}_0, \widetilde{X}_0)$ has a well-separated maximum at β_0. The uniform convergence result from Proposition 1 fulfills the other requirement for consistency.

Proof of Proposition 3 (Asymptotic Normality)

The assumptions in Theorem 3.2 in Ichimura and Lee (2010) can be verified, from which the asymptotic normality follows.

- **Asumption 3.1** This assumption requires identification and consistency of the estimator. It is satisfied by the identification restrictions and Proposition 2 (Consistency).
- **Asumption 3.2** This assumption requires the existence a linear approximation of the objective function with a bounded error. It is satisfied by the smoothness of $g(\cdot)$.
- **Asumption 3.3** This assumption requires that a second-order Taylor expansion of $E(g(\cdot))$ be well defined. This assumption is satisfied, again, by the smoothness of $g(\cdot)$ and the continuity assumptions in A5.
- **Asumption 3.4** This assumption imposes a series of smoothness conditions on the first step estimation. Condition (a) in this assumption is satisfied from the identification assumption. Condition (b) is satisfied by the smoothness of the kernels. Condition (c) is satisfied from the uniform convergence results for Kernel estimators (Hansen, 2008, Newey, 1994a and 1994b). Conditions (d) and (e) are satisfied as the first stage estimation does not depend on β_0.

- **Asumption 3.5** This assumption ensures that the remainder term of the Taylor Series expansion is negligible. It can be verified by applying proposition 3.1 of the same paper. Conditions (a) and (b) are satisfied by choosing

$$\omega(\cdot) = \begin{bmatrix} 2 & -2\beta \\ -2\beta & 2\beta^2 \end{bmatrix}$$

and condition (c) is satisfied as the first step estimation does not depend on β_0.

- **Asumption 3.6** This assumption requires that the effect of the first stage estimation on the final precision of the estimates of β_0 be representable as the sum of zero-mean and finite-variance random variables. It can be verified by using the results from remark 3.3 of Ichimura and Lee (2010), and defining:

$$g(z,\theta) = \begin{bmatrix} E(2\tilde{y} - \beta\widetilde{x}) \\ E(-2\beta(\widetilde{y} - \beta\widetilde{x})) \end{bmatrix}$$

and

$$E\big(\varphi(z,\theta)|\nu(Z) = \nu(z)\big) = \begin{bmatrix} E(y|Z = z) \\ E(x|Z = z) \end{bmatrix}.$$

2.B Tables from Simulations

Table 2.B1: Bias, Standard Error, and Root Mean Square Error for Simulations of the Base Case

Sample	RMSE				Absolute Bias				Standard Error			
Size	LC	D50	D200	M	LC	D50	D200	M	LC	D50	D200	M
250	0.17	0.21	0.54	0.43	0.10	0.08	0.02	0.02	0.13	0.19	0.53	0.43
500	0.15	0.16	0.27	0.32	0.08	0.10	0.03	0.01	0.12	0.12	0.27	0.32
750	0.13	0.13	0.19	0.26	0.06	0.10	0.03	0.02	0.11	0.09	0.19	0.26
1000	0.12	0.13	0.17	0.24	0.06	0.11	0.04	0.03	0.11	0.07	0.16	0.23
1250	0.12	0.12	0.15	0.21	0.06	0.11	0.06	0.04	0.10	0.06	0.14	0.21
1500	0.11	0.12	0.13	0.19	0.05	0.10	0.05	0.03	0.09	0.05	0.12	0.19
1750	0.10	0.12	0.12	0.19	0.04	0.11	0.06	0.02	0.09	0.05	0.11	0.19
2000	0.10	0.12	0.12	0.17	0.05	0.11	0.07	0.01	0.08	0.05	0.10	0.16
2250	0.09	0.12	0.11	0.15	0.04	0.11	0.07	0.01	0.08	0.05	0.09	0.15
2500	0.08	0.12	0.11	0.15	0.03	0.11	0.07	0.01	0.08	0.05	0.08	0.15
2750	0.08	0.12	0.11	0.14	0.04	0.11	0.08	0.02	0.07	0.04	0.08	0.14
3000	0.08	0.12	0.11	0.13	0.03	0.11	0.08	0.02	0.07	0.04	0.08	0.13
3250	0.07	0.12	0.10	0.13	0.02	0.11	0.08	0.02	0.07	0.04	0.07	0.13
3500	0.07	0.12	0.10	0.12	0.02	0.11	0.08	0.02	0.07	0.04	0.06	0.12
3750	0.07	0.12	0.11	0.12	0.02	0.11	0.09	0.02	0.07	0.04	0.06	0.12
4000	0.07	0.12	0.11	0.12	0.02	0.11	0.09	0.02	0.07	0.04	0.06	0.12
4250	0.07	0.11	0.10	0.12	0.02	0.11	0.09	0.02	0.07	0.04	0.06	0.11
4500	0.07	0.12	0.10	0.12	0.01	0.11	0.09	0.02	0.07	0.04	0.05	0.12
4750	0.07	0.11	0.10	0.11	0.01	0.11	0.09	0.01	0.07	0.03	0.05	0.11
5000	0.07	0.11	0.10	0.11	0.01	0.11	0.09	0.00	0.07	0.03	0.04	0.11

The data generating process in the Base Case satisfies all the assumptions in the Local Cohorts estimator. LC refers to the Local Cohorts estimator, D50 and D200 refer to Deaton's estimator with cohort sizes of 50 and 200 individuals, and M refers to Moffitt's estimator using a flexible polynomial to account for the unknown data generating process. For reference, the true value of the parameter of interest is $\beta_0 = 2$.

Table 2.B2: Bias, Standard Error, and Root Mean Square Error for Simulations of the Group Effects Case

Sample	RMSE				Absolute Bias				Standard Error			
Size	LC	D50	D200	M	LC	D50	D200	M	LC	D50	D200	M
250	0.35	0.45	4.42	1.83	0.03	0.02	0.68	0.14	0.35	0.45	4.36	1.82
500	0.31	0.27	0.96	1.28	0.03	0.02	0.02	0.07	0.31	0.27	0.96	1.27
750	0.29	0.20	0.55	1.15	0.03	0.01	0.04	0.08	0.29	0.20	0.55	1.15
1000	0.27	0.15	0.47	1.09	0.04	0.01	0.04	0.10	0.26	0.15	0.46	1.08
1250	0.24	0.13	0.36	0.93	0.03	0.01	0.01	0.13	0.24	0.12	0.36	0.93
1500	0.23	0.12	0.31	0.84	0.02	0.01	0.01	0.08	0.23	0.12	0.31	0.84
1750	0.23	0.11	0.29	0.83	0.01	0.01	0.02	0.05	0.23	0.11	0.29	0.83
2000	0.22	0.11	0.27	0.72	0.02	0.02	0.02	0.02	0.22	0.10	0.27	0.72
2250	0.21	0.10	0.24	0.64	0.03	0.02	0.03	0.01	0.21	0.10	0.24	0.64
2500	0.19	0.10	0.22	0.64	0.03	0.02	0.04	0.05	0.19	0.09	0.21	0.64
2750	0.17	0.09	0.20	0.61	0.02	0.02	0.04	0.07	0.17	0.09	0.20	0.60
3000	0.18	0.08	0.18	0.53	0.01	0.02	0.03	0.03	0.18	0.08	0.18	0.53
3250	0.19	0.09	0.16	0.52	0.01	0.01	0.02	0.02	0.18	0.09	0.16	0.52
3500	0.17	0.08	0.15	0.52	0.02	0.01	0.01	0.06	0.17	0.08	0.15	0.52
3750	0.16	0.08	0.15	0.50	0.02	0.02	0.02	0.06	0.16	0.08	0.14	0.49
4000	0.15	0.08	0.15	0.47	0.01	0.02	0.02	0.05	0.15	0.08	0.15	0.47
4250	0.15	0.08	0.14	0.47	0.01	0.01	0.01	0.03	0.15	0.08	0.14	0.47
4500	0.15	0.08	0.13	0.46	0.01	0.01	0.01	0.01	0.15	0.08	0.12	0.46
4750	0.15	0.08	0.12	0.45	0.01	0.01	0.01	0.02	0.15	0.07	0.11	0.45
5000	0.14	0.07	0.12	0.45	0.00	0.02	0.01	0.04	0.14	0.07	0.12	0.45

The data generating process in the Group Effects Case assumes that the unobserved individual-level heterogeneity is constant (up to random noise) within groups defined by the quintiles of the observable time-invariant characteristics (which implies a mild violation of assumption A2 of the Local Cohorts estimator). LC refers to the Local Cohorts estimator, D50 and D200 refer to Deaton's estimator with cohort sizes of 50 and 200 individuals, and M refers to Moffitt's estimator using a flexible polynomial to account for the unknown data generating process. For reference, the true value of the parameter of interest is $\beta_0 = 2$.

Table 2.B3: Bias, Standard Error, and Root Mean Square Error for Simulations for the Underspecified Case

Sample	RMSE				Absolute Bias				Standard Error			
Size	LC	D50	D200	M	LC	D50	D200	M	LC	D50	D200	M
250	0.22	0.47	18.07	2.15	0.07	0.05	0.33	0.13	0.20	0.47	18.06	2.14
500	0.17	0.28	1.27	2.00	0.07	0.07	0.06	0.17	0.15	0.27	1.27	1.99
750	0.15	0.21	0.64	1.81	0.06	0.08	0.03	0.13	0.13	0.20	0.64	1.80
1000	0.13	0.18	0.51	1.57	0.06	0.09	0.04	0.11	0.12	0.16	0.51	1.56
1250	0.11	0.15	0.41	1.37	0.06	0.09	0.05	0.08	0.10	0.12	0.40	1.37
1500	0.11	0.14	0.33	1.36	0.05	0.09	0.06	0.11	0.09	0.11	0.32	1.35
1750	0.10	0.14	0.30	1.31	0.04	0.09	0.07	0.11	0.09	0.11	0.29	1.30
2000	0.09	0.13	0.29	1.06	0.03	0.09	0.07	0.06	0.08	0.10	0.28	1.06
2250	0.08	0.13	0.27	1.04	0.03	0.08	0.08	0.07	0.08	0.09	0.26	1.04
2500	0.08	0.12	0.24	1.24	0.03	0.08	0.08	0.06	0.07	0.09	0.23	1.24
2750	0.07	0.12	0.21	1.18	0.02	0.08	0.06	0.04	0.07	0.09	0.20	1.18
3000	0.07	0.12	0.19	0.85	0.01	0.08	0.06	0.04	0.07	0.09	0.18	0.85
3250	0.06	0.11	0.18	0.77	0.01	0.07	0.07	0.05	0.06	0.08	0.17	0.76
3500	0.06	0.11	0.17	0.76	0.01	0.08	0.06	0.06	0.06	0.08	0.16	0.76
3750	0.06	0.11	0.17	0.76	0.01	0.08	0.06	0.07	0.06	0.08	0.15	0.75
4000	0.05	0.10	0.16	0.77	0.01	0.07	0.07	0.11	0.05	0.08	0.14	0.76
4250	0.05	0.10	0.15	0.76	0.01	0.07	0.06	0.09	0.05	0.07	0.13	0.75
4500	0.05	0.11	0.15	0.69	0.01	0.07	0.07	0.04	0.05	0.07	0.13	0.69
4750	0.05	0.11	0.14	0.65	0.01	0.08	0.07	0.04	0.05	0.07	0.12	0.65
5000	0.05	0.11	0.13	0.65	0.01	0.08	0.07	0.02	0.05	0.07	0.11	0.65

The data generating process in the Underspecified Case assumes that the unobserved individual-level heterogeneity depends on three time-invariant characteristics (and noise), one of which is unobserved. LC refers to the Local Cohorts estimator, D50 and D200 refer to Deaton's estimator with cohort sizes of 50 and 200 individuals, and M refers to Moffitt's estimator using a flexible polynomial to account for the unknown data generating process. For reference, the true value of the parameter of interest is $\beta_0 = 2$.

2.C Tables from Application

Table 2.C1: Classification of TV Programs

Liberal	Conservative	Neutral
Anderson Cooper 360	Fox News	ABC World News
Countdown with Keith Olbermann	Fox Report with Shepard Smith	CBS Evening News
Hardball with Chris Matthews	Hannity and Colmes	Lou Dobbs
CNN Headline News/Newsroom	Hannity's America	NBC Nightly News
ABC News Nightline	The O'Reilly Factor	Meet the Press
Situation Room w. Wolf Blitzer	The Beltway Boys	Today Show
The Daily Show w. Jon Stewart	Studio B w. Shepard Smith	The NewsHour w. Jim Lehrer
Good Morning America	Geraldo at Large	Larry King Live
This Week w. G. Stephanopoulos	Your World with Neil Cavuto	60 Minutes
The View	Fox and Friends	Face the Nation
The Colbert Report	Special Report w. Brit Hume	Reliable Sources
Late Edition w. Wolf Blitzer		The Early Show
MSNBC Live		Frontline
Out in the Open		CBS Sunday Morning
BET News		20/20
		Dateline NBC
		The McLaughin Group
		CBS Morning News
		America This Morning

The classification of TV programs comes from Dilliplane (2014).

Table 2.C2: Estimates from NAES 2008 Data

	Lib TV	L	U	Con TV	L	U	Neu TV	L	U
Panel	0.57	0.19	0.94	-0.13	-0.56	0.30	0.35	0.08	0.62
LCE	0.87	0.17	1.57	-0.62	-1.33	0.09	1.11	0.61	1.61
Controls	2.11	1.82	2.40	-1.79	-2.03	-1.55	0.49	0.28	0.70
Moffitt's	26.93	23.00	30.86	-11.63	-13.19	-10.07	-5.72	-8.30	-3.15

Panel refers to estimates obtained from real panel data. Controls refers to a linear regression that includes the observed time-invariant characteristics as controls. L and U denote the lower and upper bounds of 95% confidence intervals for the corresponding coefficients.

Chapter 3

PARTIAL EFFECTS FOR BINARY OUTCOME MODELS WITH UNOBSERVED HETEROGENEITY

3.1 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 individual heterogeneity relates to the observed covariates in the model.[1]

In this chapter I develop an estimator that deals with unobserved heterogeneity in binary outcome models, the *Penalized Flexible Correlated Random Effects* (PF-

[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.

CRE) estimator. In the PF-CRE estimator, I explicitly account for the correlation between the observed and unobserved components of the model, using a large flexible specification (more details below). Moreover, I include a penalization step for variable selection to induce efficiency. This estimator addresses the 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, I estimate the model 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 heterogeneity, but not to the covariates of interest. Like other penalized estimation

[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.

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.

3.2 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.

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:

$$y_{it} = \mathbb{I}[\alpha + x_{it}\beta + c_i - \varepsilon_{it} > 0], \quad i = 1, ..., n, \; t = 1, ..., T, \tag{3.1}$$

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.[4]

When the error terms are independently and identically distributed according to a known cumulative distribution $G(\cdot)$, equation 3.1 can be alternatively written as:

$$Prob(y_{it} = 1 | x_{it}, c_i) = G(\alpha + x_{it}\beta + c_i). \tag{3.2}$$

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

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

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

are calculated by taking expectations over c.[5] The partial effects for the model in equation 3.2 are defined as:

$$PE_j(x) = E\left[\frac{\partial}{\partial x_j}G(\alpha + x\beta + c)\Big|x\right], \quad j = 1, ..., k \tag{3.3}$$

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

$$APE_j = E\left[\frac{\partial}{\partial x_j}G(\alpha + x\beta + c)\right], \quad j = 1, ..., k \tag{3.4}$$

where the last expectation is taken with respect to both x and c.[6]

Assumptions for Identification and Estimation

The identification challenge in the model of equation 3.2 lies in c_i being unobserved *and* correlated with x_{it}.[7] The identification strategy I use in this chapter 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 polynomial on $z_i^1, ..., z_i^T$, where $z_i^t = \sum_{s=1}^{T}(x_{is})^t$ (Altonji and Matzkin 2005, and

[5] 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.

[6] Note that some authors refer to equation 3.3 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 3.4. I reserve the term average partial effect for equation 3.4.

[7] 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.

references therein for further details).[8] Note that when divided by T, $(z_i^1, ..., z_i^T)$ are in fact the first T non-central moments of $(x_{i1}, ..., x_{iT})$.

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.

Under Assumption 2, I restrict attention to a linear index of the first τ moments of $(x_{i1}, ..., x_{iT})$, observed time-invariant characteristics, and interaction terms (notice that this 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 for the unobserved heterogeneity, hence Assumption 3. However, other distributions are possible, as long as they have finite moments.[9]

[8] 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.

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

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

$$\begin{aligned} c_i &= z_i\gamma + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma^2) \\ f(c_i|x_{i1}, ..., x_{iT}) &= \mathcal{N}(z_i\gamma, \sigma^2) \end{aligned} \tag{3.5}$$

Estimation

Imposing Assumptions 1, 2, and 3 to the model in equation 3.2 results in the following specification:

$$Prob(y_{it} = 1|x_{it}, c_i) = G(\alpha + x_{it}\beta + z_i\gamma + \eta_i), \ with\ \eta_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \tag{3.6}$$

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.[10]

In principle, the parameters β in equation 3.6 can be estimated via Maximum Likelihood. The log-Likelihood function for this model is:

$$\log L(\beta, \alpha, \gamma, \sigma) = \sum_{t=1}^{T}\sum_{i=1}^{n}[y_{it}\log(p_{it}) + (1 - y_{it})\log(1 - p_{it})] \tag{3.7}$$

with

$$p_{it} \equiv Prob(y_{it} = 1|x_{it}) = \int_{-\infty}^{\infty} G(\alpha + x_{it}\beta + z_i\gamma + \eta_i)\frac{1}{\sigma}\phi(\eta_i/\sigma)d\eta_i, \tag{3.8}$$

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

The model in equation 3.6 represents a *flexible* specification of a Correlated Random Effects (CRE) model. It is a CRE-type model as it assumes a specific correlation form between the unobserved heterogeneity and the covariates in the model (represented by $z_i\gamma$). It is flexible because, under Assumptions 1 and 2, it can accommodate a wide range of correlation forms.

The flexible specification derived from Assumptions 1 and 2 requires the estimation of additional coefficients (γ). When the number of covariates is small, γ is relatively low dimensional. 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

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

requires the estimation of 20 additional parameters.[11] 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 it does 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. I estimate β using *Penalized Flexible Correlated Random Effects* (PF-CRE), which is defined by:

$$(\widehat{\beta}, \widehat{\alpha}, \widehat{\gamma}, \widehat{\sigma}) = \arg \max_{(\beta,\alpha,\gamma,\sigma)} \log L(\beta, \alpha, \gamma, \sigma) - \Pi_\lambda(\gamma), \tag{3.9}$$

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 and Li 2001), defined as:

$$\Pi_\lambda(\gamma) = \begin{cases} \lambda|\gamma| & \text{if } |\gamma| \leq \lambda, \\ -\frac{|\gamma|^2 - 2a\lambda|\gamma| + \lambda^2}{2(a-1)} & \text{if } \lambda < |\gamma| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\gamma| > a\lambda, \end{cases} \tag{3.10}$$

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.[12]

[11] 3 time-means, 6 two-way interactions, 10 three-way interactions, and the variance of the random effect.

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

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,*

$$\sqrt{nT}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, I(\beta)^{-1}), \tag{3.11}$$

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

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 (Ibrahim et al. 2011; Hui, Muller, and Welsh 2017). Consequently, it has the same asymptotic properties of the Maximum Likelihood estimator of the data generating process. Consistency, efficiency, and normality of the PF-CRE estimator thus follow from the properties of Maximum Likelihood estimators.[13]

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,

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

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

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

mapping theorem.[14] Standard errors for the partial effects can be obtained via the Delta method or bootstrap.

To estimate the partial effects, it is necessary to specify a value of z. In principle, any value of z is valid for estimating the partial effects. However, a significant proportion (or all) of the terms in z are functions of x. For this reason, it is advisable to ensure that the values of x and z used to calculate the partial effects are consistent with one another to avoid issues similar to those of extreme counterfactuals (King and Zeng 2006). 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.[15]

3.3 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 I develop in this chapter.[16]

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.[17,18]

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

[15] 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.

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

[17] 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.

[18] The asymptitic bias is of order $O_p(T^{-1})$, meaning that it disappears as T tends to infinity.

In light of the inconsistency of the FE estimator, bias correction procedures have been proposed.[19] These corrections reduce the bias; however, they do not eliminate it.[20] 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.[21] 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.[22]

The CMLE approach has two main shortcomings. First, it does not provide estimates of the partial effects.[23] 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 of information, and consequently larger standard errors

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.

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

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

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

[22] 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) and Abrevaya (2000).

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

in the estimates.[24,25]

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.[26]

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 leads to a flexible specification of the unobserved heterogeneity. This flexible specification can capture a wide range of correlation forms between the unobserved heterogeneity and the observed covariates in the model.

If the exchangeability assumption holds, the PF-CRE estimator has several advantages relative to the FE and CMLE approaches. Unlike the FE approach, it does not suffer from the incidental parameters problem. It also allows for the estimation of probabilities and partial effects, which cannot be done with CMLE. Finally, PF-CRE also provides more efficient estimates of the model parameters than FE and

[24]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 chapter is that despite the presence of unobserved heterogeneity, individuals' behavioral rules are the same. That is, they all have the same β.

[25]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).

[26]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.

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).

3.4 Specification Test

The method outlined in section 3.2 relies on the assumption that the unobserved heterogeneity in the data can be appropriately captured through the flexible correlation specification represented by the $z\gamma$ terms. This assumption 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 proposed in this chapter 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.[27] 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:

$$\delta = d'V(d)^{-1}d, \quad with \quad d = \widehat{\beta}_{CMLE} - \widehat{\beta}_{PF-CRE}, \tag{3.12}$$

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

[27] 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.

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:[28]

$$V(d) = V(\widehat{\beta}_{CMLE}) - V(\widehat{\beta}_{PF-CRE}). \tag{3.13}$$

Hence, putting equations 3.12 and 3.13 together:

$$\delta \equiv \left(\widehat{\beta}_{CMLE} - \widehat{\beta}_{PF-CRE}\right)' \left(V(\widehat{\beta}_{CMLE}) - V(\widehat{\beta}_{PF-CRE})\right)^{-1} \left(\widehat{\beta}_{CMLE} - \widehat{\beta}_{PF-CRE}\right). \tag{3.14}$$

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.

3.5 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 Maximum Likelihood estimate that uses the exact specification of the data generating process. In the first set of simulations I analyze the of 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.[29] 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:

$$Prob(y_{it} = 1|x_{it}, c_i) = \Lambda(\alpha + x_{it}\beta + c_i), \ with\ x_{it} \in \mathbb{R}^5, \tag{3.15}$$

$$\beta = (0.7, 1.3, -0.4, 1.2, -0.2), \ \alpha = 0.2, \tag{3.16}$$

$$c_i|\mathbf{x}_i \sim \mathcal{N}(\mu_i, \sigma_c^2),$$

[28] 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.

[29] 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.

where $\Lambda(\cdot)$ is the logistic cumulative distribution, $x_{it} \sim \mathcal{N}(0, I_5)$. 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:*

  $\mu_i = 0$

- *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}. The three data generating processes for the conditional mean of the unobserved heterogeneity (μ_i) have several characteristics that make them interesting to study. The sparse specification represents one of the best case scenarios for PF-CRE. In this case, the model is relatively simple and should lead to significant efficiency gains relative to CMLE. Moreover, because the inclusion of an interaction term, the traditional CRE approach should be biased. Finally, the sparsity of the specification helps illustrate the gains from the penalization step. The random effect specification is included to show the performance of the PF-CRE estimator when the only heterogeneity present in the data is actually random heterogeneity. Finally, the complex specification is useful to study the performance of PF-CRE in a case in which the efficiency gains from the penalization 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.

Parameter Simulations

The DGP in equation 3.15 satisfies the assumptions of both the CMLE and PF-CRE estimators, and therefore the estimates of β from both of them are consistent. Table 3.1 shows the 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.[30] As expected,

[30]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 3.A1 in the appendix.

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 80% 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.

Table 3.1: $\widehat{\beta}$ 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 cross-sectional variation will have a smaller efficiency loss than in simpler specifications

Partial Effects Simulations

Here I compare the Partial Effects for the DGP in equation 3.15 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 3.2 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, in turn, produces estimates with RMSEs that can be more than 400% 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 35% 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 400% 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 3.A3 and 3.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 3.2: $\widehat{PE}$ 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	1.19	2.49	1.28	4.16
β_2	1.00	5.05	1.23	5.57	1.00	0.97	1.83	3.37	1.36	1.44	1.53	5.93
β_3	1.02	1.23	1.39	2.93	1.00	1.00	1.48	1.86	1.07	6.04	1.11	2.56
β_4	1.00	1.86	1.50	6.46	1.00	0.97	1.68	3.04	1.36	1.43	1.52	5.88
β_5	1.03	1.02	1.38	1.90	1.00	1.00	1.38	1.48	1.02	5.54	1.03	1.52

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

Specification Test Simulations

Using the same setting as for the previous simulations, I calculate the rejection rate of the model specification test in equation 3.14 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. Table 3.3 shows that the rejection rate of the (true) null hypothesis that the PF-CRE estimator is consistent and more efficient than the CMLE estimator is close to the theoretical 5% and 10% values for the Sparse Specification.

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 consistent. This implies that the test will provide a conservative recommendations

when the unobserved heterogeneity is complex. However, this over-rejection rate approaches theoretical levels with larger sample sizes.

Table 3.3: 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 3.A7 and 3.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.[31] 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 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 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.

[31] 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

3.6 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).[32]

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).[33]

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 me to reduce or eliminate the concerns outlined in the previous paragraph. In particular, the unobserved heterogeneity modeled by PF-CRE captures voters' overall characteristics and tendencies, which will reflect the fact that parties choose to contact some voters but not others.

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.[34] 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

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

[33] 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, Green, and Larimer (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 moral dilemmas.

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

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.[35] I defined 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 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.[36] 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.

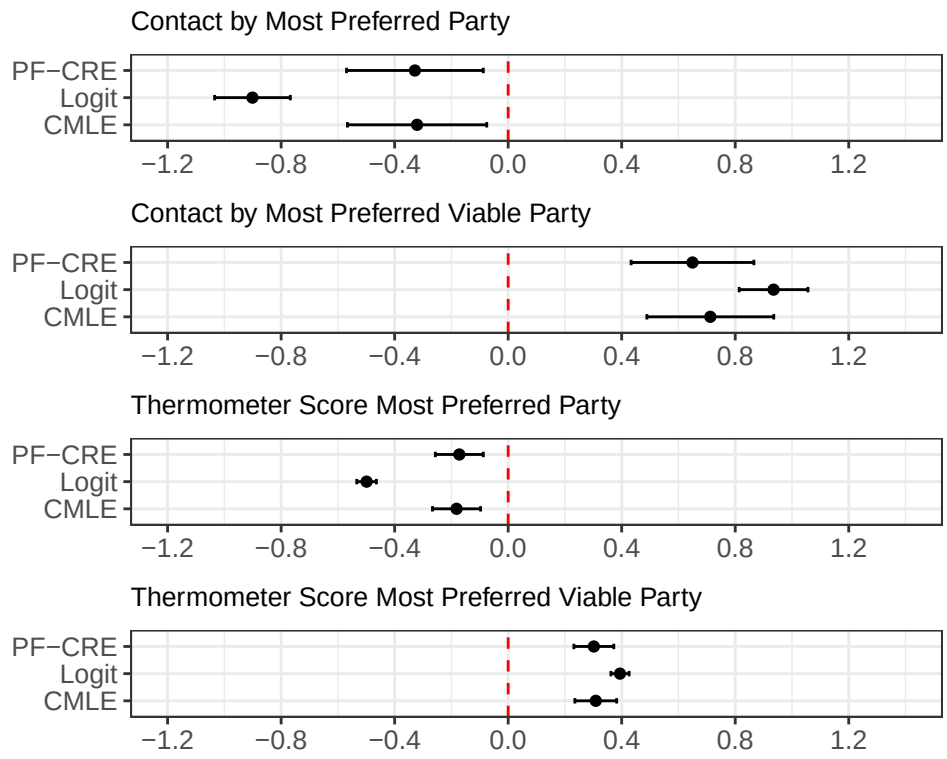
[35] In these cases, a tactical vote for these voters only occurs when none of their most preferred parties are viable and they cast a vote for the most liked viable party.

[36] I do not include a traditional CRE estimator here because the CRE estimator is nested in PF-CRE.

Results

Figure 3.1 shows that the coefficient estimates of PF-CRE and CMLE are very similar to one another.[37] 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 3.1 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.

Figure 3.1: Coefficient Estimates, Tactical Voting 2015 U.K. Election



The tuning parameter for 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

[37] See Table 3.B1 in the appendix for details with the estimates from the three models.

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 number 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.

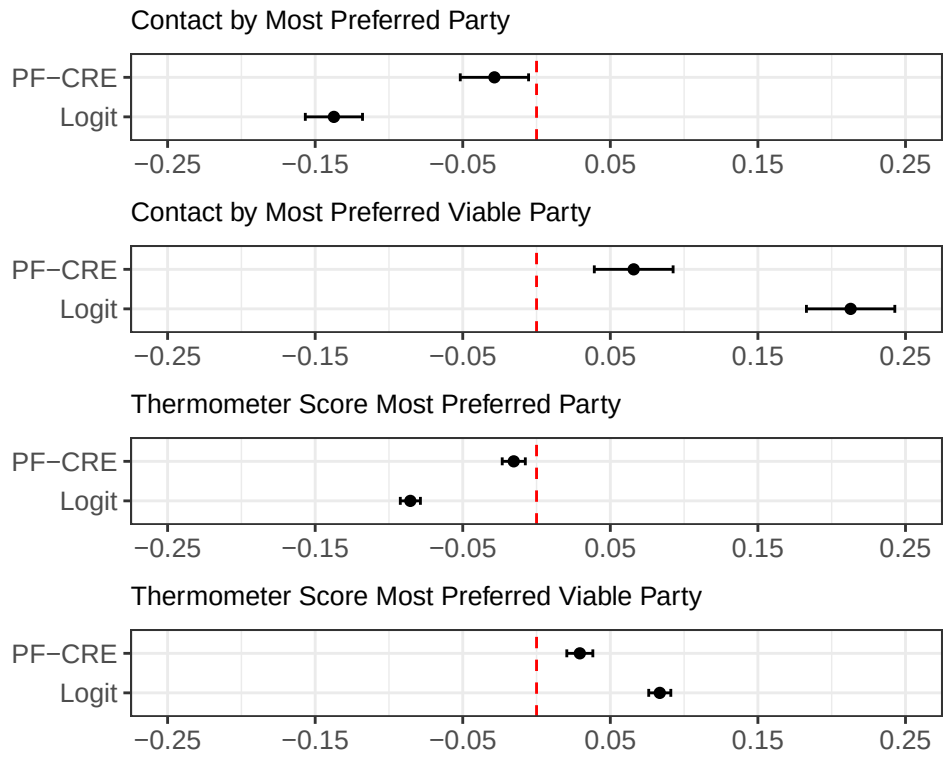
Accounting for the unobserved heterogeneity, as the PF-CRE estimator does, controls for the selection effect introduced by the way parties choose voters for contact. This reduces or eliminates the bias introduced by this selection effect, as it captures voters' overall characteristics, which are likely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

Figure 3.2 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

Figure 3.2: 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.

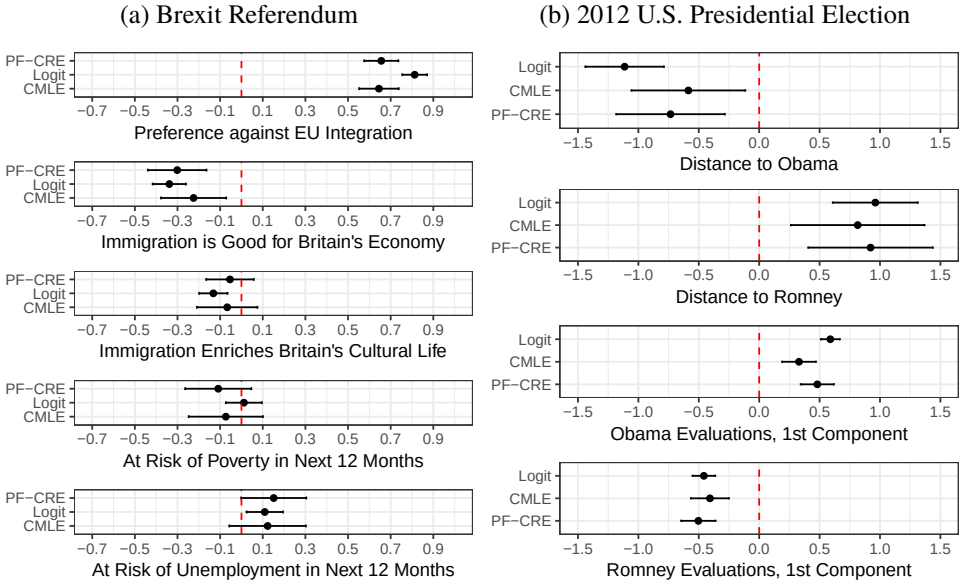
cannot, and the results show that parties' efforts to contact voters during the pre-election season have significant effects on the probability that voters cast a tactical vote. These results are important, because they show that parties can benefit from contacting voters as a way to encourage or discourage them from voting tactically.

3.7 Additional Applications

In this section I present a very brief discussion of two additional applications: 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 3.C and 3.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 1258 terms (of which 22 are selected by the penalization term). The specification test for PF-CRE returns a p-value of 0.074, which provides mixed statistical evidence for its validity. However, coefficient estimates from PF-CRE are similar to those of CMLE, and have smaller standard errors (see Figure 3.3a). Importantly, pooled logit overestimates some effects and provides excessively small confidence intervals for other variables.

Figure 3.3: Coefficient Estimates, Additional Applications



The tuning parameters for the 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 868 terms. The specification test supports the PF-CRE specification, with a p-value of 0.97. As Figure 3.3b 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).

3.8 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 requires imposing weak 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 an important contribution is to the study of comparative political institutions and international relations. In this type of environment, 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.

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 2010). 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, Muller, and Welsh 2017).

References

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

– . 1997. "The Equivalence of Two Estimators of the Fixed-Effects Logit Model". *Economic Letters* 55:41–43.

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

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

Alvarez, R. Michael, and John Brehm. 2002. *Hard Choices, Easy Answers: Values, Information, and American Public Opinion*. New Haven: Princeton University Press.

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

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

Ansolabehere, Stephen, Jonathan Rodden, and James M. Snyder. 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, Nathaniel, and Jonathan N. Katz. 2001. "Throwing out the Baby with the Bath Water: A Comment on Green, Kim, and Yoon". *International Organization* 55:487–495.

Becker, Sascha O., Thiemo Fetzer, and Dennis Novy. 2017. *Who Voted for Brexit? A Comprehensive District-Level Analysis*. CEP Discussion Paper No. 1480.

Bonhomme, Stephane, Thibaut Lamadon, and Elena Manresa. 2017. "Discretizing Unobserved Heterogeneity". Wokring paper.

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

Brinegar, Adam, Seth Jolly, and Herbert Kitschelt. 2004. "Varieties of Capitalism and Political Divides over European Integration". In *European Integration and Political Conflict*, ed. by Gary Marks and Marco Steenbergen, 62–89. Cambridge: Cambridge University Press.

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

Chamberlain, Gary. 1980. "Analysis of Covariance with Qualitative Data". *Review of Economic Studies* 47:225–238.

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

– . 1984. "Panel Data". Chap. 22 in *Handbook of Econometrics*, ed. by Zvi Griliches and Michale D. Intriligator, vol. II. Amsterdam: North-Holland.

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

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

Dhaene, Geert, and Koen Jochmans. 2015. "Split-Panel Jackknife Estimation of Fixed-Effect Models". *Review of Economic Studies* 82:991–1030.

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

Fan, Jianqing, and Runze Li. 2001. "Variable Selection Via Nonconcave Penalized Likelihood and its Oracle Properties". *Journal of the American Statistical Association* 96:1348–1360.

Feldman, Stanley. 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, Ivan. 2009. "Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models". *Journal of Econometrics* 150:71–85.

Fernandez-Val, Ivan, and Francis Vella. 2011. "Bias Corrections for Two-Step Fixed Effects Panel Data Estimators". *Journal of Econometrics* 163:144–162.

Garry, John, and James Tilley. 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, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment". *American Political Science Review* 102 (1): 33–48.

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

Greene, William H. 2015. "Panel Data Models for Discrete Choice". In *The Oxford Handbook of Panel Data*, ed. by Badi H Baltagi. Oxford University Press.

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

Hahn, Jinyong, and Whitney Newey. 2004. "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models". *Econometrica* 72:1295–1319.

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

Heckman, James J. 1981. "The Incidental Parameters Problem and the Problem of Initial Conditions in Discrete Time-Discrete Data Stochastic Process". In *Structural Analysis of Discrete Data with Econometric Applications*, ed. by Charles Manski and Daniel McFadden. Cambridge: MIT Press.

Hui, Francis K.C., Samuel Muller, and Alan H. Welsh. 2017. "Hierarchical Selection of Fixed and Random Effects in Generalized Linear Mixed Models". *Statistica Sinica* 27:501–518.

Ibrahim, Joseph G., et al. 2011. "Fixed and Random Effects Selection in Mixed Effects Models". *Biometrics* 67 (2): 495–503.

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

King, Gary, and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals". *Political Analysis* 14:131–159.

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

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

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

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

Neyman, Jerzy, and Elizabeth L. Scott. 1948. "Consistent Estimates Based on Partially Constistent Observations". *Econometrica* 16:1–32.

Poole, Keith, et al. 2013. `basicspace`*: A Package to Recover a Basic Space from Issue Scales*.

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

Shipman, Tim. 2016. *All Out War: The Full Story of How Brexit Sank Britain's Political Class*. London: William Collins.

Su, Liangjun, Zhentao Shi, and Peter C. B. Phillips. 2016. "Identifying Latent Structures in Panel Data". *Econometrica* 84 (6): 2215–2264.

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

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Second. Cambridge, MA: MIT Press.

3.A Additional Figures and Tables from Simulations

Table 3.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 3.A2: $\widehat{\beta}$ Standard Deviation relative to true $|\beta|$

	Sparse		RE		Complex	
	CMLE	PF-CRE	CMLE	PF-CRE	CMLE	PF-CRE
β_1	13.50%	10.50%	13.79%	7.98%	16.66%	11.95%
β_2	9.53%	6.51%	9.19%	5.50%	11.22%	7.46%
β_3	21.61%	12.81%	21.89%	12.53%	26.32%	19.72%
β_4	9.79%	5.31%	9.50%	5.45%	11.72%	7.85%
β_5	41.45%	23.72%	41.69%	22.99%	47.81%	37.12%

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

Table 3.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%	−1.51%	25.36%	−1.91%	−23.86%
β_2	−1.88%	−30.81%	−1.47%	−14.40%	−1.65%	−1.01%	−1.39%	−3.39%	−1.11%	7.96%	−1.44%	−23.71%
β_3	−2.29%	−9.84%	−2.01%	−14.84%	−1.05%	−0.39%	−1.37%	−3.47%	−1.73%	−108.07%	−2.34%	−24.28%
β_4	−1.45%	−8.66%	−0.90%	−13.84%	−1.72%	−1.09%	−1.12%	−3.23%	−1.35%	9.04%	−1.68%	−23.76%
β_5	−1.84%	−8.41%	−1.38%	−14.05%	−1.50%	−0.87%	−0.96%	−3.06%	0.14%	−189.76%	−0.31%	−22.04%

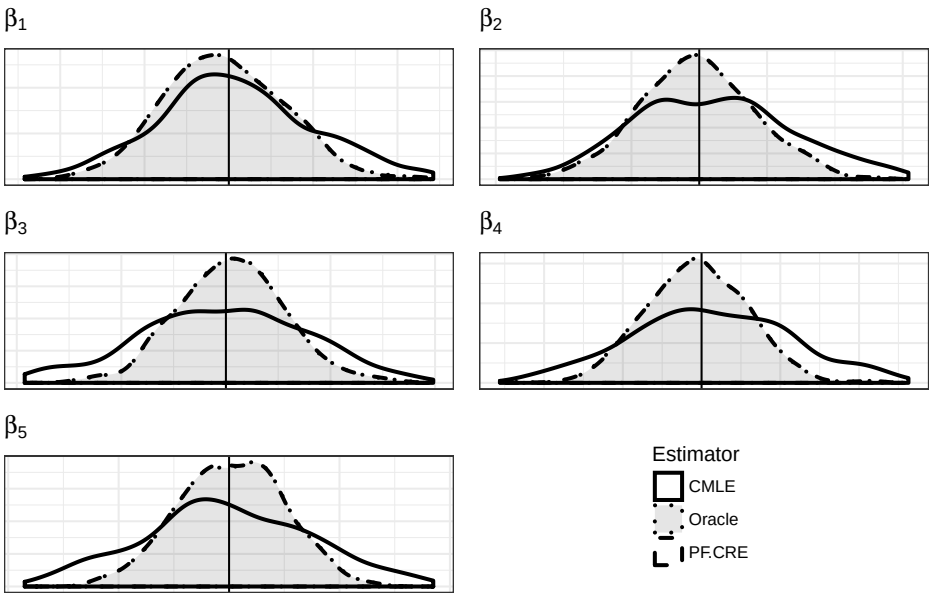
The quantities in this table are calculated as: $(\widehat{PE}/PE - 1) \times 100$.

Table 3.A4: $\widehat{PE}$ Standard Deviation 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	10.21%	7.41%	10.92%	31.91%	7.23%	7.30%	11.27%	17.21%	12.53%	7.78%	13.52%	37.32%
β_2	5.85%	4.31%	7.46%	31.20%	4.47%	4.48%	8.61%	15.68%	9.97%	7.07%	11.23%	36.82%
β_3	12.80%	12.24%	17.65%	34.33%	11.94%	12.03%	17.68%	21.99%	19.14%	11.65%	19.79%	39.23%
β_4	5.14%	4.81%	7.96%	31.52%	5.02%	5.03%	8.87%	15.82%	10.03%	5.64%	11.24%	36.84%
β_5	24.55%	22.87%	33.08%	43.33%	23.50%	23.63%	32.36%	34.71%	35.13%	24.79%	35.70%	47.70%

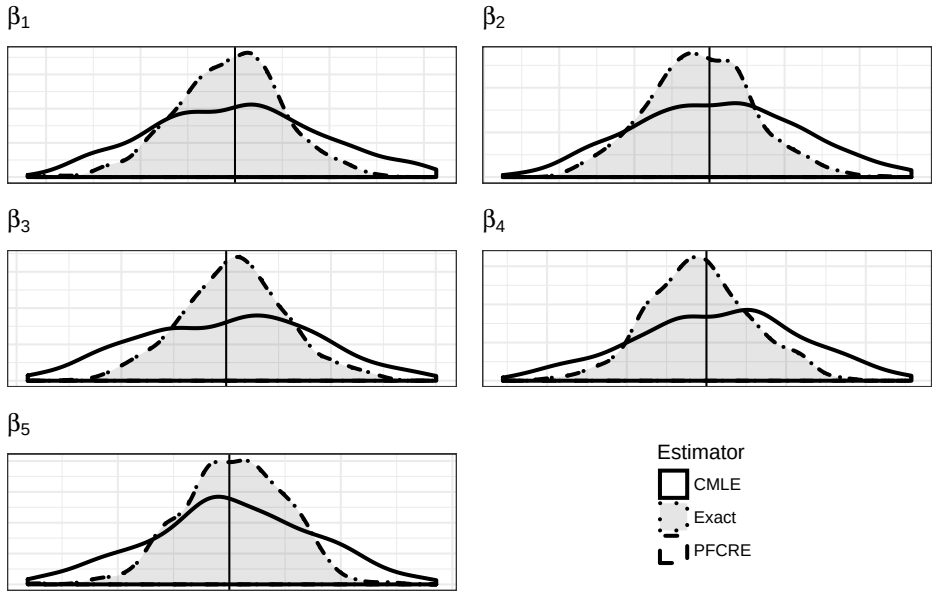
The quantities in this table are calculated as: $\sqrt{V(\widehat{PE})}/|PE| \times 100$.

Figure 3.A1: $\widehat{\beta}$ Distributions: PF-CRE v. CMLE, Sparse Specification



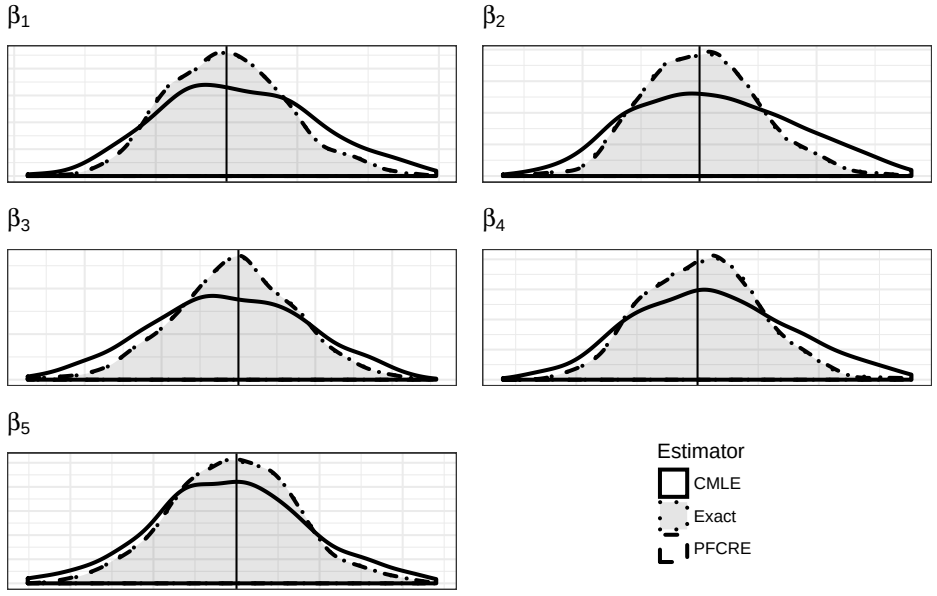
Vertical lines represent the true value of the parameters. The distributions correspond to the estimates for each parameter and estimator.

Figure 3.A2: $\widehat{\beta}$ Distributions: PF-CRE v. CMLE, Random Effect Specification



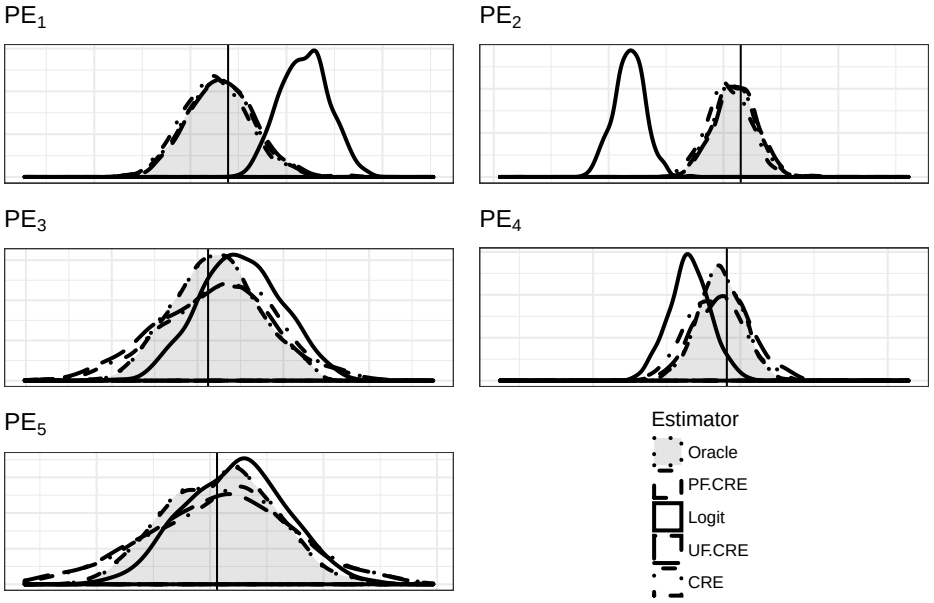
Vertical lines represent the true value of the parameters. The distributions correspond to the estimates for each parameter and estimator.

Figure 3.A3: $\widehat{\beta}$ Distributions: PF-CRE v. CMLE, Complex Specification



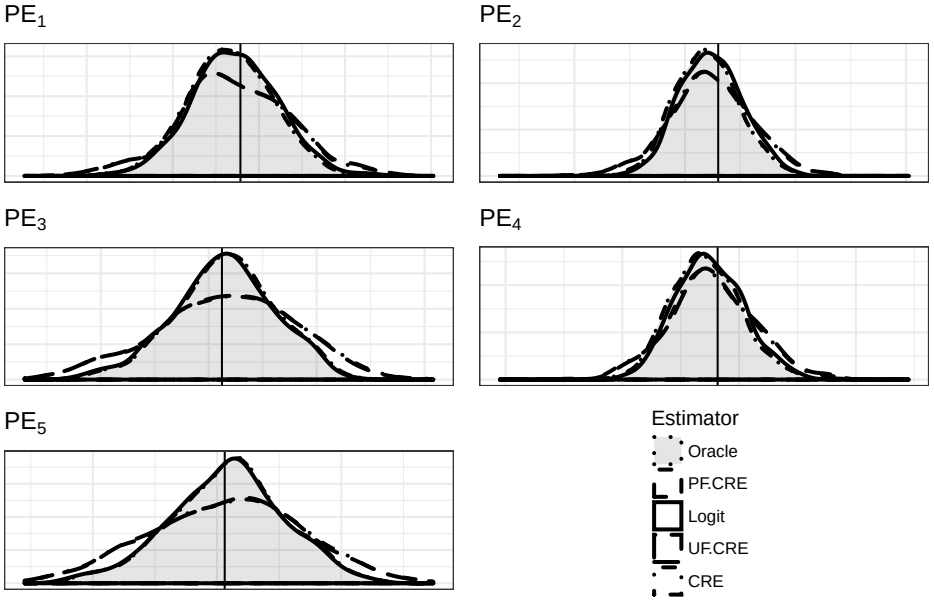
Vertical lines represent the true value of the parameters. The distributions correspond to the estimates for each parameter and estimator.

Figure 3.A4: $\widehat{PE}$ Distributions, Sparse Specification Specification



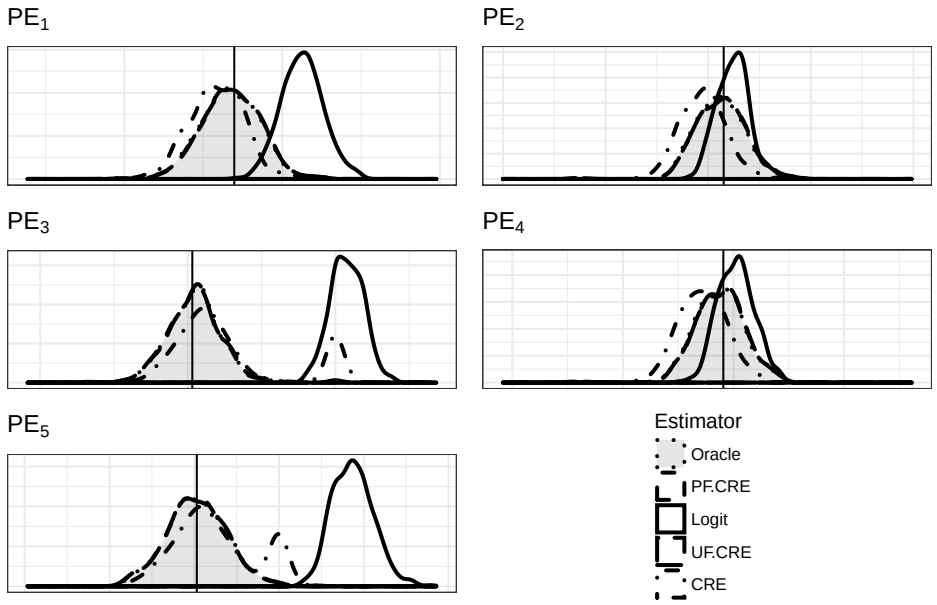
Vertical lines represent the true value of the parameters. The distributions correspond to the estimates for each parameter and estimator.

Figure 3.A5: $\widehat{PE}$ Distributions, Random Effect Specification



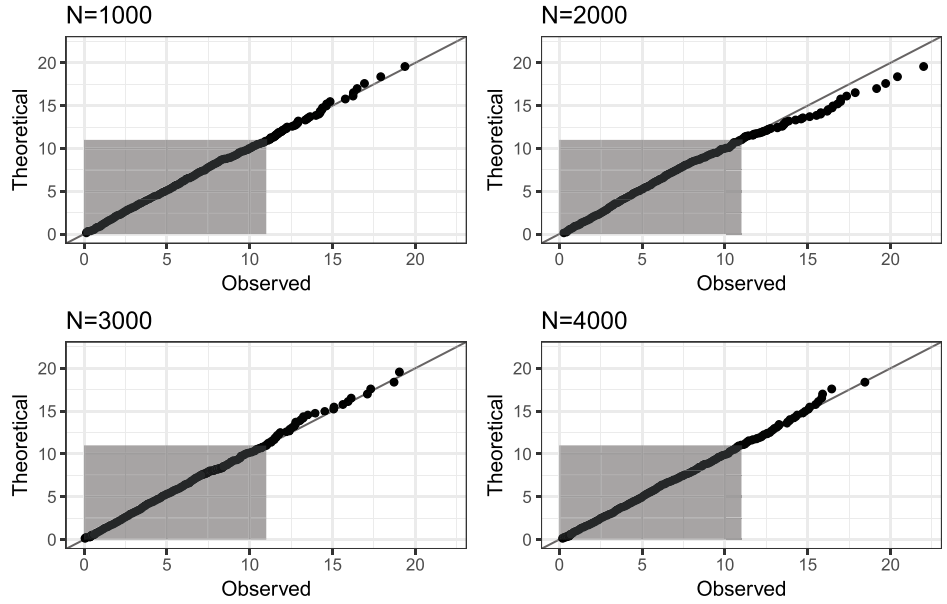
Vertical lines represent the true value of the parameters. The distributions correspond to the estimates for each parameter and estimator.

Figure 3.A6: $\widehat{PE}$ Distributions, Complex Specification



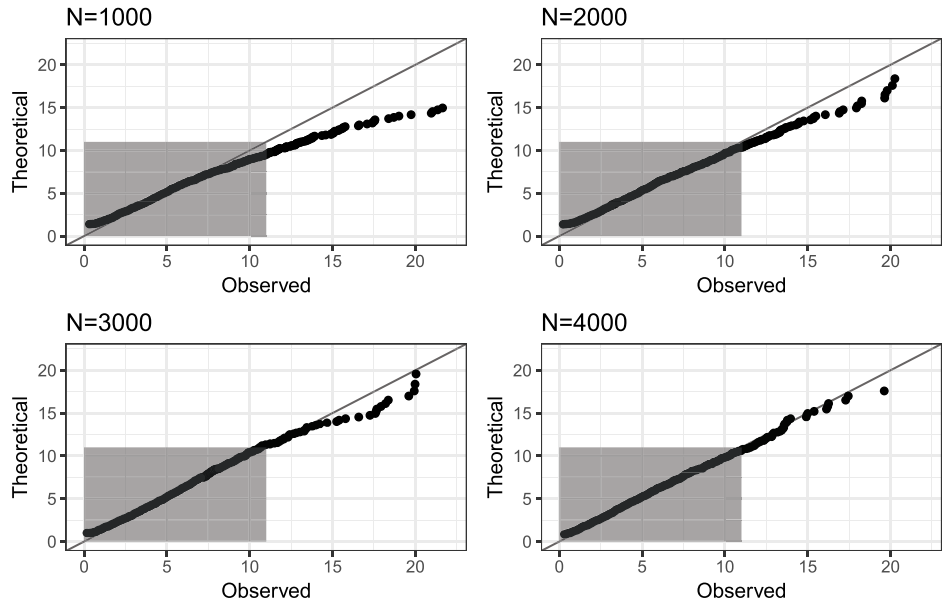
Vertical lines represent the true value of the parameters. The distributions correspond to the estimates for each parameter and estimator.

Figure 3.A7: Specification Test, Quantile-Quantile Plots, Sparse Specification



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

Figure 3.A8: Specification Test, Quantile-Quantile Plots, Complex Specification



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

3.B Additional Figures and Tables from Tactical Voting Application

Table 3.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. 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.

3.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, Fetzer, and Novy 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.[38] 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.[39]

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.

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 changes in individuals' economic fears and immigration concerns played out in their referendum vote decisions.

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).[40]

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

[39] There is a relatively large literature that focuses on an utilitarian approach to European integration. See, for example, Tucker, Pacek, and Berinsky (2002), Brinegar, Jolly, and Kitschelt (2004), and Garry and Tilley (2015).

[40] 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

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, gender, race, education level, employment status, household income, and indicators for whether the respondent has ever lived abroad, has friends from E.U. countries, and whether his/her parents were born in a foreign country.

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 three-way interactions among these terms, for a total of 1258 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.

Results

Table 3.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 0.074, which implies the validity of PF-CRE, albeit with weak statistical evidence. The pooled logit, which ignores the unobserved heterogeneity, overestimates some effects, particularly the coefficient on preferences against European 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.

slightly when using this version of the question. However, qualitative (and to a large extent) quantitative results remain the same.

Table 3.C1: Coefficient Estimates for Brexit Referendum

	PF-CRE			CMLE			Logit		
	β	Low	High	β	Low	High	β	Low	High
Against Integration	0.66	0.58	0.74	0.64	0.55	0.74	0.81	0.75	0.87
Immigration, Cultural	-0.05	-0.17	0.06	-0.07	-0.21	0.07	-0.13	-0.20	-0.07
Immigration, Economic	-0.30	-0.44	-0.16	-0.22	-0.38	-0.07	-0.34	-0.42	-0.026
Risk Poverty	-0.11	-0.26	0.05	-0.07	-0.25	0.10	0.01	-0.07	0.10
Risk Unemployment	0.15	0.00	0.30	0.12	-0.06	0.30	0.11	0.02	0.19
Controls	No			No			Yes		
N^o γ terms	1258			-			-		
Selected γs	22			-			-		
Observations	9,466			9,466			9,466		
Effective Obs	9,466			2,175			9,466		
$\chi^2_{(5)}$	10.04			-			27.50		
p-value	0.074			-			0.00		

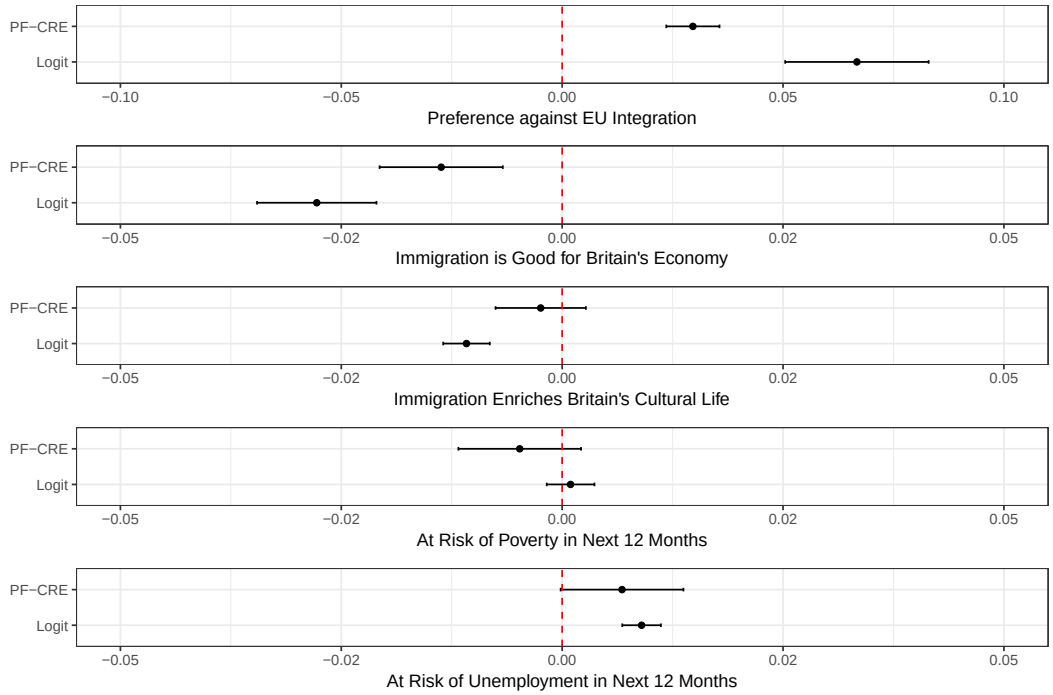
All confidence intervals are at the 95% level. Logit standard errors are clustered at the individual level. The effective number of observations refers to the number of actual observations used in estimation for CMLE. There is no χ^2 test reported for CMLE since this estimator is the basis for that test.

Figure 3.C9 presents the partial effects estimated for a baseline individual.[41] The PF-CRE estimates show that an increase in preferences against European integration are associated with a 2.95% increase in the probability of voting in favor of Brexit; logit overestimates this effect by 3.73 percentage points. In terms of respondents' views on immigration, the results show that those who find that immigration is good for the economy are 1.36 percentage points less likely to support Brexit, whereas there is no effect on the cultural side. Logit overestimates both these effects, by 1.42 and 0.82 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 0.7 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

[41] 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 3.C9: Partial Effects from Brexit Referendum



The tuning parameter for 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.

drivers of the referendum results, and that values related to Britain's culture did not play a significant role.

3.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.[42] 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

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

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).[43] 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 unobserved 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.

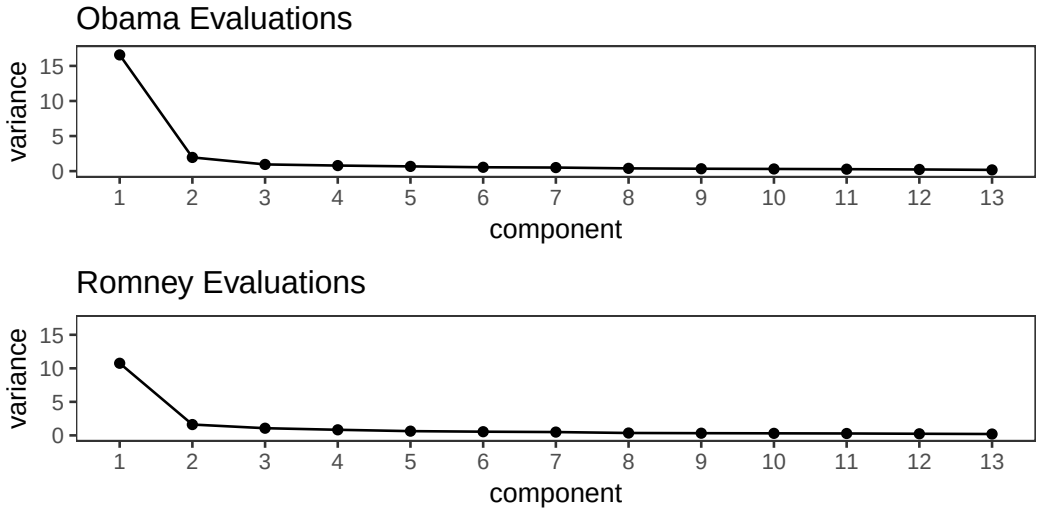
Data and Model Specification

To study the effect of ideological distance and candidate personality evaluations on vote choice, I use data from three waves (February, June, and October) 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

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

Figure 3.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).

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.[44] These evaluations are very highly correlated with each other, and using them all together in a model introduced more noise than explanatory power (Ansolabehere, Rodden, and Snyder (2008) make a similar argument for the case of issue positions). For this reason, I simplify the personality evaluations by replacing them with their first three principal components for each candidate, as additional dimensions do not contribute significantly to explaining the variance in each candidate evaluations (see Figure 3.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 868 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.[45]

[44]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.

[45]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.

I compare the estimates from PF-CRE to those of CMLE, a pooled logit estimator that includes the time-invariant characteristics as control variables, and a traditional CRE approach that only uses the time-means of the covariates to model the conditional distribution of the unobserved heterogeneity.

Results

Table 3.D1 shows the coefficient estimates for the main variables of interest in the model: ideological distance and the first three components of the candidate personality evaluations 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 0.97. The pooled logit model, which does not account for the unobserved heterogeneity, significantly overestimates the effect of ideological distance to Obama by a factor of two. Logit also overestimates the effect of the Obama personality evaluations on the probability of voting for Obama. These differences highlight the importance of controlling for unobserved heterogeneity in the estimation of vote choice.

Figure 3.D2 shows the partial effects estimated from PF-CRE and the pooled logit estimators. The baseline individual for these partial effects is a 40 years old white woman, with median income, some college education, and with ideological distances and personality evaluations at the average for an individual with these demographic characteristics.

The partial effects from PF-CRE show that, for the baseline individual, increasing the ideological distance to Obama is associated with a 12.5 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 15 percentage points. Pooled logit, which ignores the unobserved heterogeneity, overestimates these effects by about 75 and 50 percent, respectively. A similar picture arises from personality evaluations. PF-CRE estimates that a more positive evaluation of Obama is associated with a 7.75 percentage points increase in the probability of voting for him, whereas better personality evaluations of Romney are associated with a decrease of about 9 percentage point in the probability of voting for Obama. Pooled logit overestimates these effets by 75 and 25 percent, respectively.

Table 3.D1: Coefficient Estimates for 2012 U.S. Presidential Election

	PF-CRE			CMLE			Logit		
	β	Low	High	β	Low	High	β	Low	High
Distance BO	-0.73	-1.18	-0.28	-0.59	-1.06	-0.12	-1.11	-1.44	-0.79
Distance MR	0.92	0.41	1.44	0.82	0.26	1.37	0.96	0.61	1.31
BO Eval, 1st	0.48	0.34	0.62	0.33	0.19	0.47	0.59	0.51	0.67
MR Eval, 1st	-0.50	-0.65	-0.36	-0.41	-0.57	-0.25	-0.46	-0.55	-0.36
BO Eval, 2nd	-0.25	-0.45	-0.04	-0.23	-0.51	0.05	-0.30	-0.51	-0.08
MR Eval, 2nd	0.19	-0.01	0.39	0.04	-0.22	0.31	0.014	-0.03	0.32
BO Eval, 3rd	0.10	-0.13	0.33	0.18	-0.16	0.53	0.07	-0.12	0.27
MR Eval, 3rd	0.06	-0.17	0.28	0.10	-0.19	0.40	0.00	-0.20	0.20
Controls	No			No			Yes		
N^o γ terms	868			-			-		
Selected γs	19			-			-		
Observations	3,825			3,825			3,825		
Effective Obs	3,825			2,175			3,825		
$\chi^2_{(8}$	2.21			-			-143.69		
p-value	0.97			-			NA		

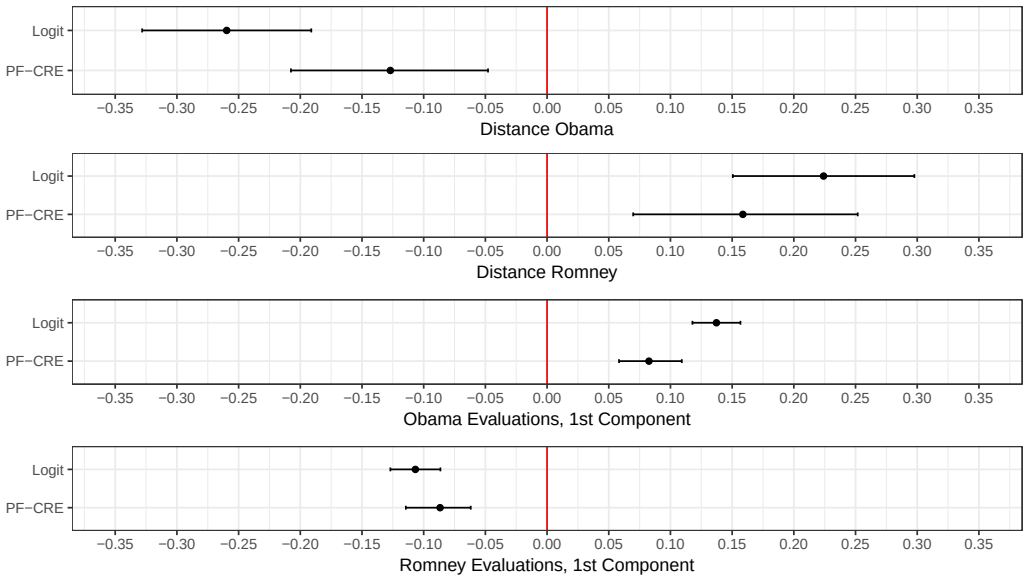
All confidence intervals are at the 95% level. Logit standard errors are clustered at the individual level. The effective number of observations refer to the number of actual observations used in estimation for CMLE. There is no χ^2 test reported for CMLE since this estimator is the basis for that test.

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

Figure 3.D2: Partial Effects from 2012 U.S. Presidential Election



The tuning parameter 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.

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.

Chapter 4

ENCOURAGING LOYALTY AND DEFECTION: THE EFFECT OF PARTY CAMPAIGNS ON TACTICAL VOTING IN BRITAIN

4.1 Introduction

In winner-take-all electoral systems, voters are often faced with the situation that their most preferred candidate is unlikely to carry the seat. In elections with more than two candidates voters facing this situation often decide to cast a tactical vote for a less preferred candidate with chances of winning, if only to prevent their most disliked candidate from being elected (Duverger 1954).[1]

There is a large empirical literature on tactical voting that has generally focused on measuring the extent to which it occurs. Evidence from a variety of countries and electoral systems shows that around 15 to 40 percent of voters who are in a position to cast a tactical vote actually decide to do so (see Alvarez, Kiewiet, and Núñez (2018) for a review on this subject).[2] In British elections this figure is typically estimated at around 30 to 40 percent (Alvarez, Boehmke, and Nagler 2006; Kiewiet 2013). These levels of tactical voting can have important consequences for overall electoral outcomes. For example, Kiewiet (2013) finds that as many as one in five Labour seats in Westminster are won thanks to tactical votes by Liberal Democrat supporters.

Perhaps because of the challenges in measuring the extent of tactical voting, the literature has focused significantly less on why some voters behave tactically while others do not. Although many important correlates of tactical voting have been found (see Section 4.2 below), they typically relate to demographic characteristics of the voters or to electoral circumstances. However, these correlates are generally non-actionable; that is, they are not variables that an electoral participant can modify (at least within the time frame of a campaign) so as to encourage or discourage tactical voting behavior. Therefore, our understanding of tactical voting remains limited.

In this chapter, I estimate the effect that being contacted by political parties has on voters' propensity to cast a tactical vote. This is an actionable (potential) correlate,

[1] What I call tactical voting in this paper has also been referred to as strategic voting.

[2] Voters in a position to cast a tactical vote are those voters whose most preferred party is considered to be out of the race.

because parties can decide whether to contact voters during the electoral campaign. The empirical challenge lies in correctly identifying the effect of party contact itself, independent of the effect of unobserved confounders. The worry is that parties 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 a researcher's point of view, this is unobserved heterogeneity in voters' behavior that is also correlated with the observed covariates (in this case, being contacted by a party).[3]

To address this estimation challenge, I use a panel data survey collected prior to the 2015 and 2017 United Kingdom General Elections. The use of panel data with multiple measurements of vote intention and indicators of contact by the different political parties allows me to significantly reduce or eliminate the concerns outlined in the previous paragraph. In particular, panel data estimators can control for unobserved heterogeneity to the extent that it is constant in time (at least within the time-frame of the study). In that sense, this unobserved heterogeneity will capture voters' overall characteristics and tendencies, which are likely closely related to the information that parties use to decide which voters they want to contact. While there are several methods that can be used to estimate panel data models with unobserved heterogeneity, in this chapter I use the *Penalized Flexible Correlated Random Effects* (PF-CRE) developed in Chapter 3. This estimator allows me to control for unobserved heterogeneity in this context, while at the same time yields consistent estimates of partial effects and predicted probabilities.

My findings show that in 2017 voters who are contacted by their most preferred party are 7.02 percentage points less likely to cast a tactical vote in 2017, indicating that party contacts can enforce loyalty and sincerity at the polling booth. However, being contacted by the most preferred viable party increases the probability of a tactical vote by 13.41 percentage points. These quantities stand at 2.75 and 7.03 percentage points in 2015, respectively. I also show, through counterfactual exercises, the effect that party contacts had on Parliamentary seats through tactical voting. Among other things, I find that if non-viable parties in each constituency gave up contacting their supporters, the Conservative party would have obtained a narrow majority following the 2017 General Election; instead, the Conservative

[3]Ideally, disentangling the effects of party contacts from that of parties choosing whom to contact could be done by relying on field experiments, in the spirit of Gerber, Green, and Larimer (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 moral dilemmas.

party formed minority government following a confidence and supply agreement with the Democratic Unionist Party from Northern Ireland.

The rest of this chapter is organized as follows. In Section 4.2 I discuss related literature; in Section 4.3 I describe the 2015 and 2017 election data and the methodology used for estimation; in Section 4.4 I present the main estimation results; in Section 4.5 I present and develop counterfactual estimates of electoral outcomes under different contact behavior by the parties; and Section 4.6 concludes

4.2 Related Literature

As mentioned in the introduction to this chapter, despite the attention that measuring tactical voting has received by empirical researchers, there is less understanding on why and when some voters cast tactical votes when given the opportunity to do so. In particular, the determinants identified by the literature do not provide with actionable recommendations. However, it is worthwhile to review these determinants. The variables that the literature has found to be associated with tactical voting can be grouped in two categories: those that relate to the individual voter, and those that relate to the electoral environment.

Among the individual voter characteristics the literature has found that voters that have strong partisan or ideological attachments are significantly less likely than others to cast a tactical vote (Blais 2002; Lanoue and Bowler 1992; Karp et al. 2002).[4] There is also evidence that voters' political sophistication and political knowledge (sometimes proxied by education levels) are positively associated with tactical voting (Alvarez, Boehmke, and Nagler 2006; Gschwend and van der Kolk 2006; Karp et al. 2002). There is also evidence that when voters are experienced with the electoral system they are more likely to exploit it and vote tactically (Spenkuch 2017; Duch and Palmer 2002), and that voters who believe the media influences the voting decisions of others are more likely to behave tactically (Cohen and Tsfati 2009).

Among the electoral environment variables, theoretical models put particular emphasis on the closeness of the election (see, for example Cox 1997). That is, when the race between the top-two contenders is considered to be close, it is expected that third party supporters will be more likely to vote tactically, as a defection from their most preferred party is more likely to be pivotal. Empirical results tend to support this theoretical expectation, albeit weakly (Lanoue and Bowler 1992; Fisher 2000;

[4] Similar effects have been found in the study of split ticket voting in the U.S. (Burden and Kimball 1998; Beck et al. 1992).

Kiewiet 2013; Elff 2014; Núñez 2016). The empirical literature has also found that the presence of a viable close ideological substitute to a non-viable preferred party encourages tactical voting (Karp et al. 2002), and that the presence of an incumbent politicians interferes with the decision to cast a tactical vote (Moser and Scheiner 2005).[5]

This paper is also related to the literature that studies campaign effectiveness. Experimental evidence from the United States shows that citizens are responsive to efforts aimed at getting them out to vote (see, for example, Gerber, Green, and Larimer 2008; Arceneaux and Nickerson 2009). Studies based on observational data have also found this positive effect (see, for example, Geys 2006; Karp, Banducci, and Bowler 2008). There is a substantial literature on British elections that studies the effects of local campaigning (like canvassing and other methods). While the 'received wisdom' prior to the early 1990s was that that constituency campaigns were made irrelevant by the advent of mass national media, more recent research has found that parties benefit electorally from more organized and intense local campaigning, both in terms of mobilization (Clarke et al. 2004; 2009; Cutts 2014; Fieldhouse et al. 2011; Whiteley and Seyd 1994, and references therein) and in terms of their vote share (Fieldhouse et al. 2011; Fieldhouse et al. 2015; Pattie and Johnston 2003; Johnston et al. 2013, and references therein).

A significant portion of the literature on campaign effectiveness in the United Kingdom focuses on aggregate data at the constituency level, e.g., the effect of constituency campaign spending (or other measures of campaign intensity) on turnout and parties' vote shares. A smaller portion of the literature focuses on analyses at the individual voter level, and is closer to the study in this chapter. For example, Pattie and Johnston (2003) use data from the 1997 British Election Study (BES) and find that door-step canvassing has an impact on respondents' vote choices, but that telephone contacts do not. Denver, Hands, and MacAllister (2004) use data from the 1992, 1997, and 2001 BES on respondents' turnout and find that campaign intensity at the constituency level increases turnout by those voters, and that individuals that were contacted by the parties were more likely to turn out to vote. However, these studies suffer from the problem that parties may be more likely to contact those voters who are already more likely to turn out to vote and vote for their party. Fieldhouse et al. (2011) partially resolve this issue using data from the 2010 BES. They use the

[5]Of course the most important electoral environment consideration is the electoral system. I avoid discussion of it here because of the focus on the First-Past-The-Post electoral system used in Britain.

pre-election wave to identify those respondents who initially declared themselves to be undecided. With this subsample, they study whether voters contacted by the different campaigns were more or less likely to support the Conservatives, Labour, or the Liberal Democrats. They find strong campaign contact effects in all cases.

A notable exception that deals with the problem of parties selecting which voters to contact is Whiteley and Seyd (2003). They study respondents' intention to turn out to vote and vote choice. Importantly, they are able to control for respondents' self-reported willingness to turn out and vote derived from an earlier panel wave. This way, their estimates of party campaign efforts are teased out from parties' mobilization efforts. Their findings show substantially smaller effects than those found in the rest of the literature. In particular, they find that canvassing face to face for the Labour party increased the probability of voting for Labour by 6%, whereas canvassing by phone increased it by 5%. These effects are in line with the ones found in this paper on the effect of party contacts on tactical voting for the 2015 and 2017 General Elections.

4.3 Data and Methodology

Data

To study the effect of party contacts on the probability of casting a tactical vote, I use data from six waves of the British Election Study (BES) Online Panel. The first three waves apply to the 2015 General Election, while the others cover the run-up to the 2017 General Election.[6] I restrict the sample to respondents that reported vote intention and party preferences in at least two of the waves for each election.

The analysis focuses on those voters whose most preferred party is not viable (i.e., those voters in a position to cast a tactical vote). I define a party as viable if it finished among the top-two contenders in a given district in the corresponding election.[7] To define the outcome of interest, it is first necessary to define voters' preferences. To

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

[7]This leaves a total of 3,824 individuals in the 2015 sample, and 4,744 individuals in the 2017 sample. Note that there are alternative ways to define party viability. For example, the BES data includes questions that asks respondents to gauge the probability that a given party wins in their constituency. However, using this data involves several challenges. First, probabilities do not sum up to one (in many cases they add up to more than 4). Second, this question is not answered by all respondents, limiting the sample in potentially biased ways. Finally, this question is only asked in two waves, thus limiting the sample further. Another alternative is to use the results of the prior election, as research has shown that voters tend to follow an election history heuristic (Lago 2008). However, given the strong wins by the SNP in Scotland in 2015 and the strong performance of UKIP in England, prior election results may not be good measures of viability in this study.

do this, I define a voter's most preferred party in the following way: (1) the party with the highest feeling thermometer score; (2) if the are ties, these are broken by the thermometer score of the party leaders; and (3) remaining ties are kept and all tied parties are considered as the most preferred parties for those voters. I also define voters' most preferred *viable* party as the most preferred party from among the viable ones (ties are dealt with in the same manner as for the most preferred party).

With these variables in hand, tactical voting is defined as occurring if voters cast a vote for the most preferred viable party, and as not occurring otherwise.[8] The main covariates of interest are two indicators for whether a voter's most preferred party or most preferred viable party contacted the voter during the four weeks prior to each election wave. I also include as independent 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: respondents' labor force status (5 categories), education level (5 categories), ethnicity (5 categories), gender (2 categories), and home ownership (3 categories).

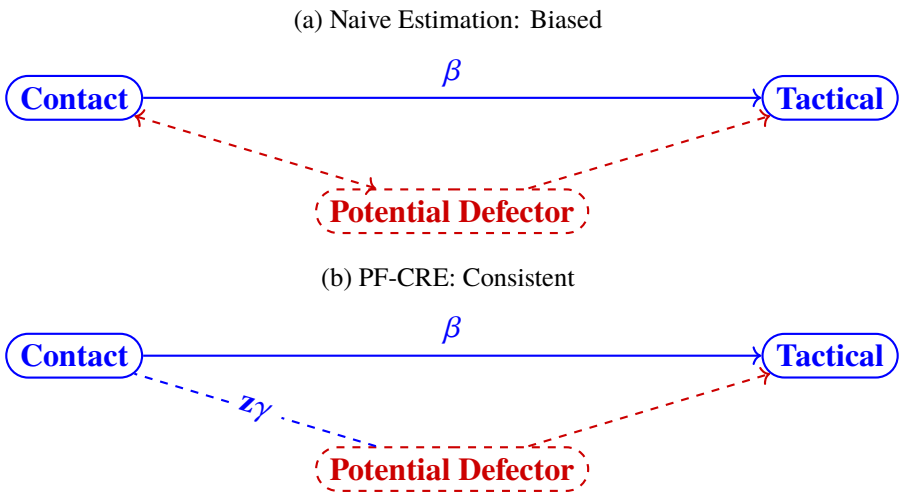
Methodology

The empirical challenge in estimating the effect of party contacts on the probability that a voter casts a tactical vote lies in the fact that parties will tend to contact voters they think will be more likely to be swayed. In terms of tactical voting, parties have strong incentives to try to contact those voters who might defect their preferred party when it is considered to be out of the race. In that case, a simple model that estimates the probability of a tactical vote as a function of contact indicators will be upward biased: parties do not contact voters at random, but are instead more likely to seek out potential defectors. Figure 4.1a represents the source of this bias graphically.

Without knowledge of how parties choose which voters to contact, it is not possible to directly control for this confounder. However, the panel structure of the BES data allows me to account for individual level heterogeneity. To the extent that who is a potential defector does not vary significantly across the survey waves *and* parties' contact strategies remain relatively stable, controlling for unobserved heterogeneity removes the bias introduced by the way parties decide to contact voters. While the

[8] Votes cast for neither the most preferred nor the most preferred viable party are considered to be non-tactical.

Figure 4.1: Estimation Challenge



A tactical vote is affected by contact from the parties (the effect of interest β), but also by who are potential defectors, which is unobserved; the correlation between potential defectors and contact by parties introduces the bias in estimating β. zγ accounts for this correlation and eliminates this bias in PF-CRE.

first condition is not testable, it is unlikely that in the short term voters will change their tendencies strongly.[9] The second condition, that parties' contact strategies do not vary significantly over the period of analysis, is testable. In Appendix 4.B, I show that the types of voters that parties contact throughout the waves considered in this study have almost exactly the same observable characteristics, suggesting that there is no change in parties' contact strategies.[10]

The model I estimate, therefore, considers the probability of casting a tactical vote as a function of the covariates of interest and the unobserved heterogeneity:

$$P(y_{it} = 1) = \Lambda(\alpha + \beta_1 CMP_{it} + \beta_2 CMPV_{it} + Controls_{it} + c_i), \qquad (4.1)$$

where y_{it} indicates whether respondent i intends to cast a tactical vote at wave t; CMP_{it} indicates whether i's most preferred party contacted her in the four weeks prior to wave t; $CMPV_{it}$ indicates the same thing for i's most preferred viable party; $Controls_{it}$ are i's thermometer ratings for the most preferred and most preferred

[9] A voter who is extremely averse to casting a tactical vote, will probably remain very averse throughout the entire period covered by the survey.

[10] Voters contacted and not contacted could differ in their unobservable characteristics. However, it is unlikely that these unobservables are fully independent of the observables. Therefore, finding no significant differences in voters' observed characteristics provides reasonable evidence that they do not differ in the unobservable ones.

viable parties; c_i represents the unobserved heterogeneity; and $\Lambda(\cdot)$ is the cumulative logistic distribution.

As is well known in the literature (see Chapter 3), estimating binary outcome models with unobserved heterogeneity either requires restricting the form of the unobserved heterogeneity or forgoing the estimation of probabilities and partial effects. In this chapter, I apply the *Penalized Flexible Correlated Random Effects* (PF-CRE) developed in Chapter 3, which imposes mild restrictions on the form of the unobserved heterogeneity while at the same time allowing for the estimation of probabilities and partial effects. Estimating these probabilities is key for Section 4.5, where I conduct several counterfactual analyses.

Using PF-CRE, I account for the correlation between the unobserved heterogeneity and the covariates of interest by explicitly modeling the unobserved heterogeneity with a flexible polynomial form, denoted by $z\gamma$ in Figure 4.1b, where z represents the polynomial terms and γ their coefficients. By accounting for this correlation, the polynomial allows for the consistent estimation of β, as it accounts for the fact that parties may target voters who are already potential defectors. This polynomial is composed of the moments of the covariates for each individual[11] as well as the observed time-invariant characteristics. Given that this polynomial is composed of a very large number of terms, PF-CRE estimates the model parameters using a penalized maximum likelihood approach that reduces its dimensionality in order to achieve efficiency (see Chapter 3). That is, PF-CRE estimates β by penalized maximum likelihood on:

$$P(y_{it} = 1) = \Lambda(\alpha + \beta_1 CMP_{it} + \beta_2 CMPV_{it} + Controls_{it} + z_i\gamma). \qquad (4.2)$$

With estimates of the parameters in equation 4.2, partial effects can be readily obtained.

4.4 Estimation Results

Naive Estimation

Before presenting the estimates that control for unobserved heterogeneity, it is useful to discuss naive estimates. The first naive estimates are based on simple differences from cross-tabulations. Table 4.1 shows that in 2017 those voters contacted by their most preferred party were 31.55 percentage points less likely to cast a tactical vote than those who were not. In turn, those voters who were contacted by the most

[11] For example, in how many waves of the panel was individual i contacted by her most preferred party.

preferred viable party were 16.53 percentage points more likely than those who were not contacted, according to this simple measure. For 2015, these figures stand at -13.29 and 12.41 percentage points.

Table 4.1: Naive Estimation

Variable	Technique	2015	2017
Contact MP	Difference	-13.29%	-31.55%
	Logit	-17.61%	-33.54%
Contact MPV	Difference	12.41%	16.53%
	Logit	19.52%	22.58%

Difference refers to the difference in tactical voting among those contacted and not contacted by the corresponding party. Logit refers to the average partial effects for the corresponding party contact. For details in the logit estimation see Table 4.A1 in the Appendix.

The second naive estimate is based on a pooled logit model of tactical voting as a function of contact indicators and demographic controls.[12] The logit model generally estimates slightly larger effects of contact by the parties than simple differences. For example, for 2015 it estimates that contact by the most preferred party diminishes the chances of tactical voting by 17.61 percentage points, whereas contact by the most preferred viable party increases these chances by 19.52 percentage points.

It is noteworthy that these estimated effects are quite large, substantively. However, as mentioned in the previous section, all these estimates potentially suffer from upward bias. This is because parties are likely to contact those voters whom they believe are more likely to sway their intended way. The next subsection presents estimates that account for this selection effect.

Accounting for Unobserved Heterogeneity

Before proceeding to the presentation of the results proper, it is necessary to determine whether the PF-CRE approach can be validly applied to the data at hand. Table 4.2 shows the result of the specification test for PF-CRE, for both the 2015 and 2017 samples.[13] In both cases, the PF-CRE specification is not rejected, and therefore PF-CRE provides consistent and efficient estimates of the parameters and partial effects for both years.

[12] See Table 4.A1 in the Appendix for more information.

[13] See Chapter 3 for more details on the test.

Table 4.2: PF-CRE Specification Test

	χ^2	p-value
2015	0.87	0.93
2017	1.64	0.80

Figure 4.2 presents the parameter estimates obtained through PF-CRE for the 2015 and 2017 U.K. General Elections.[14] The results show that voters who are contacted by their most preferred viable are more likely to defect their preferred party and vote tactically. On the other hand, voters contacted by their most preferred party are encouraged to remain loyal, thus reducing the chance that they will cast a tactical vote. Importantly, the degree of tactical voting depends on the extent to which voters like their most preferred and most preferred viable party. The more a voter prefers its most preferred party, the less likely she is to cast a tactical vote; the more she prefers the viable party, the more likely she is to vote tactically.[15]

Figure 4.3 shows the partial effects for 2015 and 2017 calculated for a baseline individual.[16] The results show that a 1 point increase in the feeling thermometer for the most preferred viable party increases the probability that this voter casts a tactical vote by 2.99 and 1.80 percentage points in 2015 and 2017, respectively. In turn, a 1 point increase in the feeling thermometer for the most preferred party is associated with a decrease in the probability of a tactical vote of 1.59 and 4.09 percentage points for 2015 and 2017, respectively. In terms of the main variables of interest, when the baseline voter is contacted by his most preferred viable party he is 7.03 and 13.41 percentage points more likely to cast a tactical vote in 2015 and 2017, respectively. On the other hand, being contacted by his most preferred party reduces the probability that he casts a tactical vote by 2.75 and 7.02 percentage points in 2015 and 2017, respectively.[17]

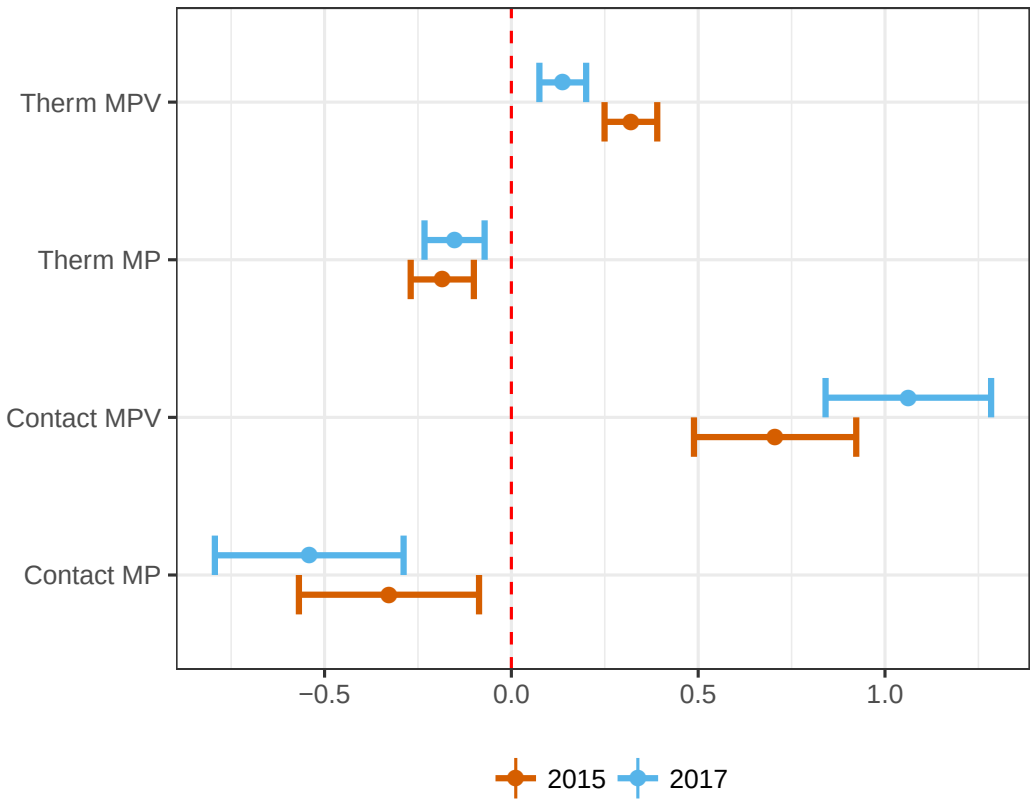
[14] See Table 4.A2 in the Appendix for additional information.

[15] Note that voters who are more favorable to their most preferred viable party still have higher ratings for the most preferred party.

[16] The baseline individual is a man between 40 and 50 years of age who works full time, owns his home outright, and has secondary education. The values of all other variables are set at the median for an individual with these characteristics. I construct the baseline this way, as it helps ensure that the baseline individual is realistic and, therefore, extrapolation from the model is reduced. Figure 4.A1 presents estimates of the average partial effects, which are in line with the partial effects presented here.

[17] Note that these differences between the most preferred and most preferred viable parties are not an artifact of the baseline individual used for estimation. Figure 4.A1 in the appendix presents the average partial effects, with an almost identical pattern.

Figure 4.2: Coefficient Estimates

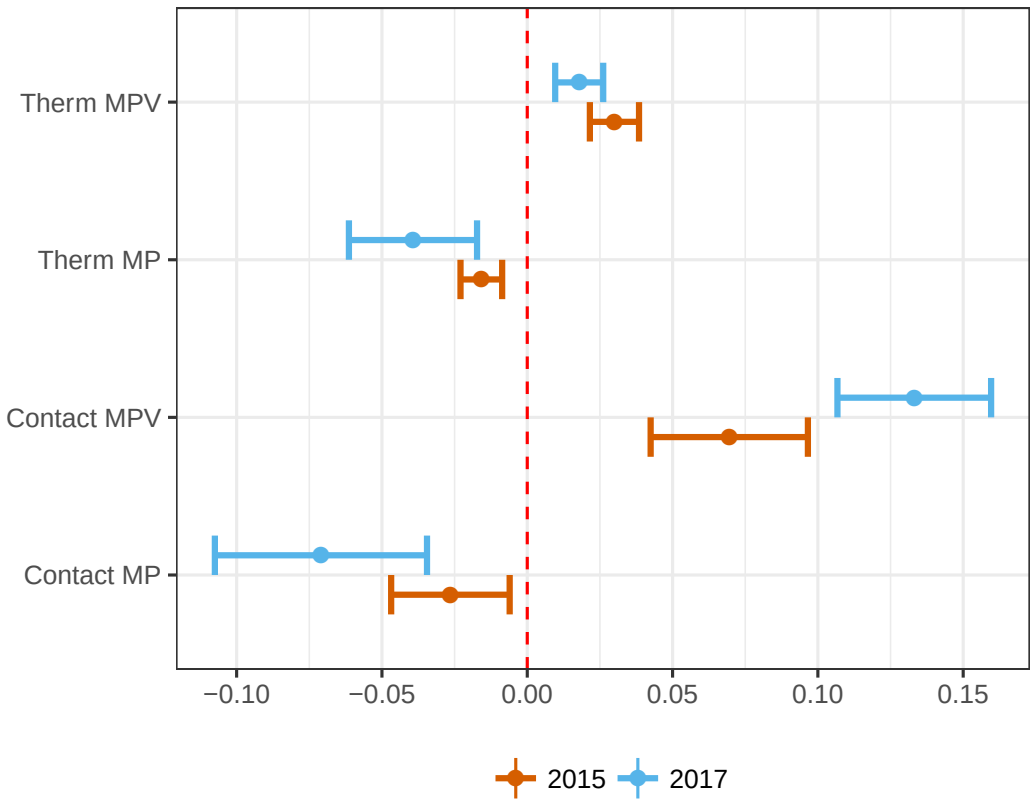


The tuning parameter was obtained through 10-fold cross validation using the Akaike information criterion. MP refers to the most preferred party, and MPV the most preferred viable party.

These results point to two interesting differences. The first is that contact by the most preferred and most preferred parties have asymmetric effects. That is, contact by the most preferred party (which in this analysis is always out of the race) has a smaller effect on voter behavior than contact by the most preferred viable party. This result is quite intuitive: it is harder to convince a voter to waste his or her vote, than it is to vote for a less preferred alternative that might actually alter the outcome of the election. The second difference is between the 2015 and 2017 estimates. In particular, in 2017 parties were more effective at changing voters' behavior when contacting them than they were in 2015. Further research can explore the reasons behind these different magnitudes between the 2015 and 2017 estimates.

The PF-CRE estimates of the effect of party contacts on tactical voting presented here are less than half the size of those obtained with naive estimation. These large

Figure 4.3: Partial Effects Estimates



The tuning parameter was obtained through 10-fold cross validation using the Akaike information criterion. The baseline individual is a male between 40 and 50 years of age who works full time, owns his home outright and has secondary education. The values of all other variables are set at the median for an individual with these characteristics.

differences are strongly suggestive of the magnitude of the selection effect. Parties in the U.K. are generally contacting those voters who will give them the highest rewards. That is, parties that are out of the race are generally contacting those voters they might persuade not to defect, whereas parties who are viable manage to contact potential defectors to sway them their way. The naive estimates ignore these facts, instead suggesting that party contacts are extraordinarily effective at influencing voting decisions. In reality, however, party contacts provide moderate influences to voter behavior.

4.5 Counterfactuals

To gain better insights into the magnitude of the effects of party contacts on tactical voting, I conduct two counterfactual studies:

> **Counterfactual A** Viable parties in each constituency contact none of the voters whose most preferred party is out of the race
>
> **Counterfactual B** Non-viable parties in each constituency contact none of the voters whose most preferred party is out of the race

Note that these counterfactuals are only partial counterfactuals: for example, the change in behavior of the non-viable parties in Counterfactual B is assumed not to generate an optimal change in behavior from the viable parties in response. That is, these counterfactuals are not the result of the equilibrium of a game. However, they are intended to measure the magnitude and direction of the incentives that parties face in such a strategic interaction.

For both counterfactuals, the baseline is the predicted voter behavior using parties' contacts as observed in the data. For Counterfactual A, I set the contact indicator for viable parties to zero for all voters; then, I use the model to predict voter behavior as with the baseline. For Counterfactual B, I set the contact indicator from the non-viable parties in each constituency to zero for all voters and then use the model to predict voter behavior. All other variables remain at their observed quantities for every respondent.

Table 4.3: Counterfactuals

	Counterfactual A			Counterfactual B		
Year	Effect	L	U	Effect	L	U
2015	-2.70	-3.51	-2.27	1.04	0.11	1.53
2017	-3.88	-4.47	-3.16	1.45	0.75	2.04

95% confidence intervals are calculated by bootstrap from the distribution of the PF-CRE estimates for the corresponding year.

Table 4.3 presents the aggregate outcome of these counterfactuals. For 2015, if viable parties gave up on contacting those voters whose most preferred party was out of the race, tactical voting would have decreased by 2.70 percentage points. On the other hand, if non-viable parties had given up contact, tactical voting would have increased by about 1.04 percentage points. For 2017, the effects are somewhat

larger, with tactical voting decreasing by 3.88 percentage points if the viable parties gave up contacting these voters, and an increase in tactical voting of 1.45 percentage points if the non-viable parties did.

Changes in Number of Parliamentary Seats

To have a better picture of the effect that party contacts has on tactical voting, and in turn, on overall election results, I study the distribution of seats in Westminster that would occur under both counterfactuals (relative to the observed voter behavior).

To obtain the distribution of seats in Westminster based on the survey data I use Multilevel Regression and Poststratification, or MRP (Park, Gelman, and Bafumi 2004). MRP models individual survey vote intentions as a function of demographic and constituency predictors, partially pooling respondents across constituencies to an extent determined by the data. Then, predicted outcomes from the model for each demographic-constituency group are poststratified by the percentages of each type in the actual constituency populations. This provides an estimate of each party's support in each constituency, from where a predicted winner can be derived. More details about the specific MRP procedure I use can be found in Appendix 4.C.

To estimate the Multilevel Regression model, I use the full sample from the final waves for the 2015 and 2017 elections; that is, the sample that includes all voters, not only the voters who are in a position to cast a tactical vote. The baseline seats per party are obtained through MRP from the actual survey responses.[18] The seats per party in each counterfactual are obtained by estimating the MRP model on survey responses that were modified in accordance with each counterfactual. That is, for Counterfactual A, I simulate individuals' vote intention if the viable parties contact none of the voters whose most preferred party is out of the race; and similarly for Counterfactual B.

Table 4.4a shows the baseline and Counterfactual A predicted number of seats in 2015. When viable parties contact none of the voters whose most preferred party is out of the race, a total of 11 seats in Parliament change hands: the Conservative party loses 3 seats to Labour, but in turn gains 6 seats from Labour and 1 from UKIP, for a net gain of 4 seats; and the Liberal Democrats gain one seat in Scotland

[18]Please note that the baseline seats per party obtained through MRP do not match exactly the seats that each party won in the election. In particular, for 2015 the MRP model predicts one additional seat for UKIP and one less seat for the Labour party than the actual election results. For 2017, MRP predicts 4 additional seats for the SNP and 4 less seats for the Conservative party than the actual election results, all in Scotland.

from the SNP.

The results for Counterfactual B for 2015 are presented in Table 4.4b. In this case, the Conservative party loses 3 seats to Labour, but gains 4 from Labour and 1 from UKIP, for a net gain of 2; while the Liberal Democrats get 1 seat from the SNP.

Table 4.5a presents the results for Counterfactual A in 2017. The results show that when viable parties give up contacting voters whose most preferred party is out of the race, the Conservative party would have lost 4 seats to Labour, but won 4 seats from Labour and 3 from th Liberal Democrats, for a net gain of 3 seats. The Labour party, in addition to gaining and losing 4 seats from the Conservatives, would have won 2 seats from the SNP but lost 1 other to the SNP as well, for a net gain of 1 seat. The SNP instead, would have lost 2 seats to Labour, but won 2 from the Conservatives, 1 from Labour, and 2 from the Liberal Democrats, for a net gain of 3 seats.

Finally, Table 4.5b shows what would have happened if the non viable parties gave up contacting their supporters in constituencies where they are out of the race in 2017. In that case, the Conservative party would have gained 6 seats from Labour and 5 from the Liberal Democrats, only losing a seat to the SNP, for a net gain of 10 seats. The Scottish Nationalist Party instead, would have lost 1 seat to Labour, but won 1 seat from the Conservatives and another one from Labour, for a net gain of 1 seat.

Overall, the results show that a number of seats change hands because parties' campaign efforts influence tactical voting behavior. As noted before, for 2015 11 seats change hands under Counterfactual A and 9 under Counterfactual B. For 2017, 18 seats change hands in Counterfactual A and 14 in Counterfactual B. These changes can be substantively important. For example, under Counterfactual B in 2017, the Conservative party would have had a net gain of 10 Westminster seats, which would have given the Conservative party a narrow majority in Parliament (of 1 seat), transforming the second May ministry from a minority to a majority government, without the need to rely on confidence and supply votes from the Democratic Unionist Party from Northern Ireland.[19]

[19]In practice, because of Sinn Fein winning 7 Parliamentary seats from Northern Ireland and maintaining a policy of abstensionism, the Conservative majority would have been slightly larger, at 4 seats.

Table 4.4: Counterfactual Seat Allocation, 2015

(a) Counterfactual A

	Counterfactual						
	Con	Lab	LD	SNP	PC	UKIP	Grn
Con	328	3	-	-	-	-	-
	[316,339]	*[0,5]*	*[0,1]*	*[0,0]*	*[0,0]*	*[0,1]*	*[0,0]*
Lab	6	225	-	-	-	-	-
	[1,10]	*[217,241]*	*[0,0]*	*[0,0]*	*[0,0]*	*[0,2]*	*[0,1]*
LD	-	-	7	1	-	-	-
	[0,0]	*[0,0]*	*[7,9]*	*[0,1]*	*[0,0]*	*[0,0]*	*[0,0]*
SNP	-	-	-	56	-	-	-
	[0,0]	*[0,1]*	*[0,0]*	*[55,47]*	*[0,0]*	*[0,0]*	*[0,0]*
PC	-	-	-	-	3	-	-
	[0,0]	*[0,1]*	*[0,0]*	*[0,0]*	*[2,3]*	*[0,0]*	*[0,0]*
UKIP	1	-	-	-	-	1	-
	[0,1]	*[0,0]*	*[0,0]*	*[0,0]*	*[0,0]*	*[0,3]*	*[0,0]*
Grn	-	-	-	-	-	-	1
	[0,0]	*[0,1]*	*[0,0]*	*[0,0]*	*[0,0]*	*[0,0]*	*[0,1]*

(b) Counterfactual B

	Counterfactual						
	Con	Lab	LD	SNP	PC	UKIP	Grn
Con	328	3	-	-	-	-	-
	[315,339]	*[0,7]*	*[0,0]*	*[0,1]*	*[0,0]*	*[0,1]*	*[0,0]*
Lab	4	227	-	-	-	-	-
	[0,7]	*[218,242]*	*[0,0]*	*[0,0]*	*[0,1]*	*[0,1]*	*[0,0]*
LD	-	-	7	1	-	-	-
	[0,1]	*[0,0]*	*[6,9]*	*[0,1]*	*[0,0]*	*[0,0]*	*[0,0]*
SNP	-	-	-	56	-	-	-
	[0,0]	*[0,0]*	*[0,0]*	*[56,57]*	*[0,0]*	*[0,0]*	*[0,0]*
PC	-	-	-	-	3	-	-
	[0,0]	*[0,0]*	*[0,0]*	*[0,0]*	*[2,3]*	*[0,0]*	*[0,0]*
UKIP	1	-	-	-	-	1	-
	[0,1]	*[0,1]*	*[0,0]*	*[0,0]*	*[0,0]*	*[0,2]*	*[0,0]*
Grn	-	-	-	-	-	-	1
	[0,0]	*[0,1]*	*[0,0]*	*[0,0]*	*[0,0]*	*[0,0]*	*[0,1]*

Rows represent the number of seats for each party in the baseline (actual) election. Columns represent the counterfactuals. Numbers in square brackets are 95% confidence intervals. Seats in the diagonals are those that are held by the same party in the baseline and the counterfactual.

Table 4.5: Counterfactual Seat Allocation, 2017

(a) Counterfactual A

	Counterfactual						
	Con	Lab	LD	SNP	PC	UKIP	Grn
Con	308	4	-	2	-	-	-
	[294,321]	*[0,10]*	*[0,1]*	*[0,3]*	*[0,0]*	*[0,1]*	*[0,0]*
Lab	4	257	-	1	-	-	-
	[3,15]	*[240,269]*	*[0,0]*	*[0,5]*	*[0,0]*	*[0,0]*	*[0,0]*
LD	3	-	7	2	-	-	-
	[2,8]	*[0,0]*	*[4,8]*	*[0,2]*	*[0,0]*	*[0,0]*	*[0,0]*
SNP	-	2	-	37	-	-	-
	[0,3]	*[0,1]*	*[0,0]*	*[33,40]*	*[0,0]*	*[0,0]*	*[0,0]*
PC	-	-	-	-	4	-	-
	[0,0]	*[0,0]*	*[0,0]*	*[0,0]*	*[3,4]*	*[0,0]*	*[0,0]*
UKIP	-	-	-	-	-	-	-
	[0,0]	*[0,0]*	*[0,0]*	*[0,0]*	*[0,0]*	*[0,0]*	*[0,0]*
Grn	-	-	-	-	-	-	1
	[0,0]	*[0,0]*	*[0,0]*	*[0,0]*	*[0,0]*	*[0,0]*	*[1,1]*

(b) Counterfactual B

	Counterfactual						
	Con	Lab	LD	SNP	PC	UKIP	Grn
Con	313	-	-	1	-	-	-
	[296,324]	*[0,8]*	*[0,2]*	*[0,2]*	*[0,1]*	*[0,0]*	*[0,0]*
Lab	6	255	-	1	-	-	-
	[5,20]	*[236,264]*	*[0,0]*	*[0,5]*	*[0,0]*	*[0,0]*	*[0,0]*
LD	5	-	7	-	-	-	-
	[2,8]	*[0,0]*	*[4,9]*	*[0,1]*	*[0,0]*	*[0,0]*	*[0,0]*
SNP	-	1	-	38	-	-	-
	[0,3]	*[0,2]*	*[0,1]*	*[32,39]*	*[0,0]*	*[0,0]*	*[0,0]*
PC	-	-	-	-	4	-	-
	[0,0]	*[0,0]*	*[0,1]*	*[0,0]*	*[2,4]*	*[0,0]*	*[0,0]*
UKIP	-	-	-	-	-	-	-
	[0,0]	*[0,0]*	*[0,0]*	*[0,0]*	*[0,0]*	*[0,0]*	*[0,0]*
Grn	-	-	-	-	-	-	1
	[0,0]	*[0,0]*	*[0,0]*	*[0,0]*	*[0,0]*	*[0,0]*	*[1,1]*

Rows represent the number of seats for each party in the baseline (actual) election. Columns represent the counterfactuals. Numbers in square brackets are 95% confidence intervals. Seats in the diagonals are those that are held by the same party in the baseline and the counterfactual.

4.6 Conclusion

Voters' preferred party is often unlikely to win in their district. In that situation, some voters choose to cast a tactical vote for another party with a greater chance of carrying the seat. The literature that studies tactical voting has typically focused on measuring its extent. However, less is known about the motivations behind the decision to cast a tactical vote, beyond voters' demographics and the electoral environment.

In this chapter, I show that parties' campaigning in constituencies, measured by whether they contacted voters during the campaign, affects voters decisions. In particular, I show that voters whose most preferred party is out of the race are between 2.75% and 7.02% more likely to remain loyal if their party contacted them. On the other hand, contact by the most preferred viable party encourages voter defections, increasing the probability of a tactical vote between 7.03 and 13.41 percentage points. Estimating counterfactuals combined with Multilevel Regression and Poststratification, I show how the changes in tactical voting behavior due to parties' campaigning affected the distribution of seats in the U.K. Parliament for both 2015 and 2017. Interestingly, I show that if parties had given up contacting their supporters in constituencies where they were out of the race, the Conservative party would have gained enough seats for an outright majority in Parliament in 2017, instead of forming a minority government.

The findings in this chapter increase our understanding of tactical voting behavior in mass elections by considering the effects of party campaigns. However, a significant proportion of tactical votes (non-tactical votes) from among those voters in a position to cast them remains unexplained. This is particularly true when focusing on variables or incentives that are under the control of an election participant, be it the government, political parties, the media, or voters themselves. Future research on tactical voting should expand its attention to considerations like party campaigns, which are under the control of election participants, as well as to psychological underpinnings of tactical vote decisions which might help us better understand why some voters choose to forfeit the opportunity to affect election outcomes, instead maintaining their support for a party or candidate who is unlikely to win.

References

Alvarez, R. Michael, Frederick J. Boehmke, and Jonathan Nagler. 2006. "Strategic Voting in British Elections". *Electoral Studies* 25:1–19.

Alvarez, R. Michael, D. Roderick Kiewiet, and Lucas Núñez. 2018. "Preferences, Constraints, and Choices". In *Routledge Handbook of Elections, Voting Behavior and Public Opinion*, ed. by Justin Fisher et al. Elsevier.

Arceneaux, Kevin, and David Nickerson. 2009. "Who is Mobilized to Vote? A Reanalisis of Eleven Randomized Field Experiments". *American Journal of Political Science* 53 (1): 1–16.

Beck, Paul A., et al. 1992. "Patterns and Sources of Ticket Splitting in Subpresidential Voting". *American Political Science Review* 86:916–928.

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

Burden, Barry C., and David C. Kimball. 1998. "A New Approach To the Study of Ticket Splitting". *American Political Science Review* 45:533–544.

Clarke, Harold, et al. 2009. *Performance Politics and the British Voter*. Basingstoke: Palgrave Macmillan.

– . 2004. *Political Choice in Britain*. Basingstoke: Palgrave Macmillan.

Cohen, Jonathan, and Yariv Tsfati. 2009. "The Influence of Presumed Media Influence on Strategic Voting". *Communication Research* 36:359–378.

Cox, Gary W. 1997. *Making Votes Count*. Cambridge, MA: Cambridge University Press.

Cutts, David. 2014. "Local Elections as a 'Stepping Stone': Does Winning Council Seats Boost the Liberal Democrats' Performance in General Elections?" *Political Studies* 62 (2): 361–380.

Denver, David, Gordon Hands, and Iain MacAllister. 2004. "The Electoral Impact of Constituency Campaigning in Britain, 1992-2001". *Political Studies* 52:289–306.

Duch, Raymond M., and Harvey D. Palmer. 2002. "Strategic Voting in Post-Communist Democracy?" *British Journal of Political Science* 32:63–91.

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

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

Fieldhouse, Edward, et al. 2015. "Is All Campaigning Equally Positive? The Impact of District Level Campaigning on Voter Turnout at the 2010 British General Election". *Party Politics* 22 (2): 215–226.

Fieldhouse, Edward, et al. 2011. "The Electoral Effectiveness of Constituency Campaining in the 2010 British General Election: The 'Triumph' of Labour?" *Electoral Studies* 30 (4): 816–828.

Fisher. 2000. “Intuition versus Formal Theory: Tactical Voting in England: 1987-1997”. Paper presented at the Annual Meeting of the American Political Science Association, Washington, DC.

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

Geys, Benny. 2006. “Explaining Voter Turnout: A Review of Aggregate-level Research”. *Electoral Studies* 25:637–663.

Gschwend, Thomas, and Henk van der Kolk. 2006. “Split Ticket Voting in Mixed Member Proportional Systems: The Hypothetical Case of The Netherlands”. *Acta Politica* 41:163–179.

Hanretty, Chris, Benjamin E. Lauderdale, and Nick Vivyan. 2016. “Comparing Strategies for Estimating Constituency Opinion from National Survey Samples”. *Political Science Research and Methods*: 1–21.

Johnston, Ron, et al. 2013. “The Long and the Short of it: Local Campaigning at the British 2010 General Election”. *Political Studies* 61 (S1): 114–137.

Karp, Jeffrey A., Susan A. Banducci, and Shaun Bowler. 2008. “Getting Out the Vote: Party Mobilization in a Comparative Perspective”. *British Journal of Political Science* 38:91–112.

Karp, Jeffrey A., et al. 2002. “Strategic Voting, Party Activity, and Candidate Effects: Testing Explanations for Split Voting in New Zealand’s New Mixed System”. *Electoral Studies* 21:1–22.

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

Lago, Ignacio. 2008. “Rational Expectations or Heuristics? Strategic Voting in Proportional Representation Systems”. *Party Politics* 14 (1): 31–49.

Lanoue, David J., and Shaun Bowler. 1992. “The Sources of Tactical Voting in British Parliamentary Elections, 1983-1987”. *Political Behavior* 14 (2): 141–157.

Lax, Jeffrey R., and Justin H. Phillips. 2009. “How Should We Estimate Public Opinion in the States?” *American Journal of Political Science* 53:107–21.

Leemann, Lucas, and Fabo Wasserfallen. 2017. “Extending the Use and Prediction of Subnational Public Opinion Estimation”. *American Journal of Political Science* 61 (4): 1003–1022.

Moser, Robert G., and Ethan Scheiner. 2005. “Strategic Ticket Splitting and the Personal Vote in Mixed-Member Electoral Systems”. *Legislative Studies Quarterly* 30 (2): 259–276.

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

Park, David K., Andrew Gelman, and Joseph Bafumi. 2004. “Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls”. *Political Analysis* 12:375–385.

Pattie, Charles, and Ron Johnston. 2003. “Local Battles in a National Landslide: Constituency Campaigning at the 2001 British General Election”. *Political Geography* 22:381–414.

Spenkuch, Jörg L. 2017. “Expressive vs. Pivotal Voters: An Empirical Assessment”. Wokring paper.

Warshaw, Chris, and Jonathan Rodden. 2012. “How Should We Measure District-Level Public Opinion on Individual Issues?” *Journal of Politics* 74 (1): 203–219.

Whiteley, Paul, and Patrick Seyd. 1994. “Local Party Campaigning and Voting Behavior in Britain”. *Journal of Politics* 56 (1): 242–251.

– . 2003. “Party Election Campaigning in Britain: The Labour Party”. *Party Politics* 9 (5): 637–652.

4.A Additional Tables and Figures

Table 4.A1: Logit Average Partial Effects on Tactical Voting

	2015			2017		
	APE	Lower	Upper	APE	Lower	Upper
Contact MP	-0.18	-0.20	-0.15	-0.36	-0.38	-0.33
Contact MPV	0.20	0.17	0.22	0.23	0.21	0.24
Therm. MP	-0.10	-0.11	-0.09	-0.06	-0.06	-0.05
Therm. MPV	0.08	0.07	0.09	0.06	0.05	0.06
Employed Full	-0.03	-0.07	-0.00	-0.02	-0.04	0.01
Employed Part	-0.00	-0.04	0.04	0.02	-0.01	0.04
Student	-0.06	-0.11	-0.02	-0.01	-0.06	0.04
Retired	-0.01	-0.07	0.05	-0.02	-0.08	0.04
Unemployed	-0.02	-0.05	0.02	-0.01	-0.03	0.01
Education 1	-0.08	-0.11	-0.05	0.01	-0.02	0.03
Education 2	-0.06	-0.09	-0.04	-0.02	-0.04	0.00
Education 3	-0.02	-0.04	0.01	-0.01	-0.04	0.01
Mixed Race	0.10	0.01	0.19	0.05	-0.04	0.13
Asian	0.06	-0.02	0.14	0.06	-0.01	0.12
Black	0.21	0.10	0.33	0.05	-0.09	0.18
Female	0.03	0.01	0.04	0.03	0.02	0.05
Age	-0.00	-0.00	0.00	-0.00	-0.00	-0.00
Owns Home	0.02	-0.01	0.04	-0.00	-0.02	0.02
Owns Mortgage	0.05	0.02	0.07	-0.02	-0.04	0.00
Observations		10,378			12,539	
McFadden's R^2		0.11			0.10	

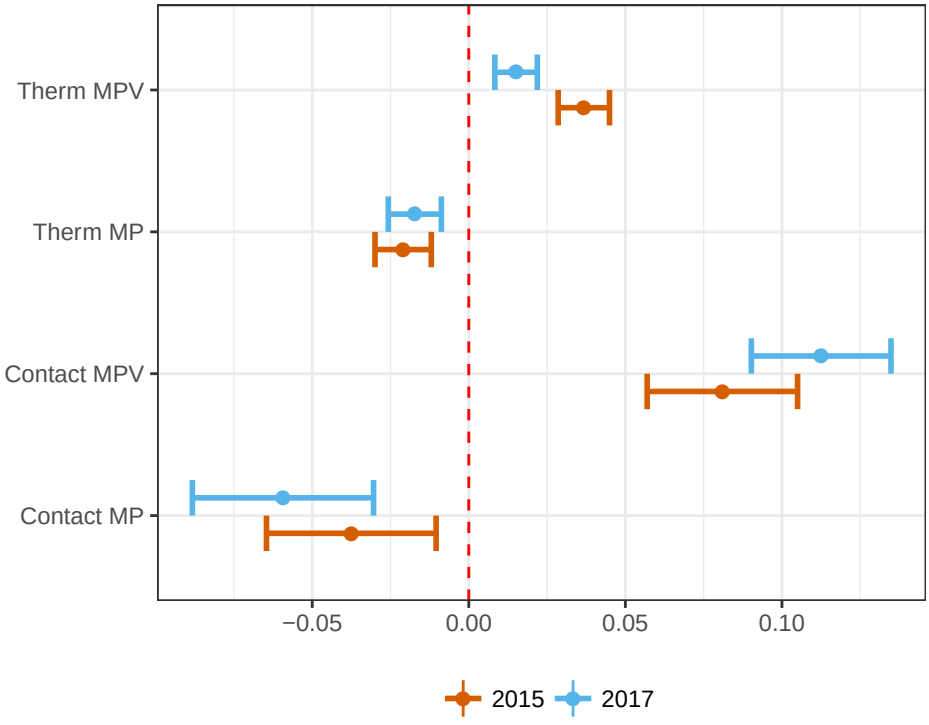
Confidence intervals are at the 95% level. Standard errors are clustered at the individual level.

Table 4.A2: Coefficient Estimates, PF-CRE

	2015			2017		
	β	Lower	Upper	β	Lower	Upper
Contact MP	-0.33	-0.57	-0.09	-0.54	-0.79	-0.29
Contact MPV	0.71	0.49	0.92	1.06	0.84	1.28
Therm. MP	-0.18	-0.27	-0.10	-0.15	-0.23	-0.07
Therm. MPV	0.32	0.25	0.39	0.14	0.08	0.20
γ terms	230			230		
γ selected	41			42		
n	3,824			4,744		
Observations	10,378			12,539		
$\chi^2_{(4)}$	0.87			1.64		
p-value	0.93			0.80		

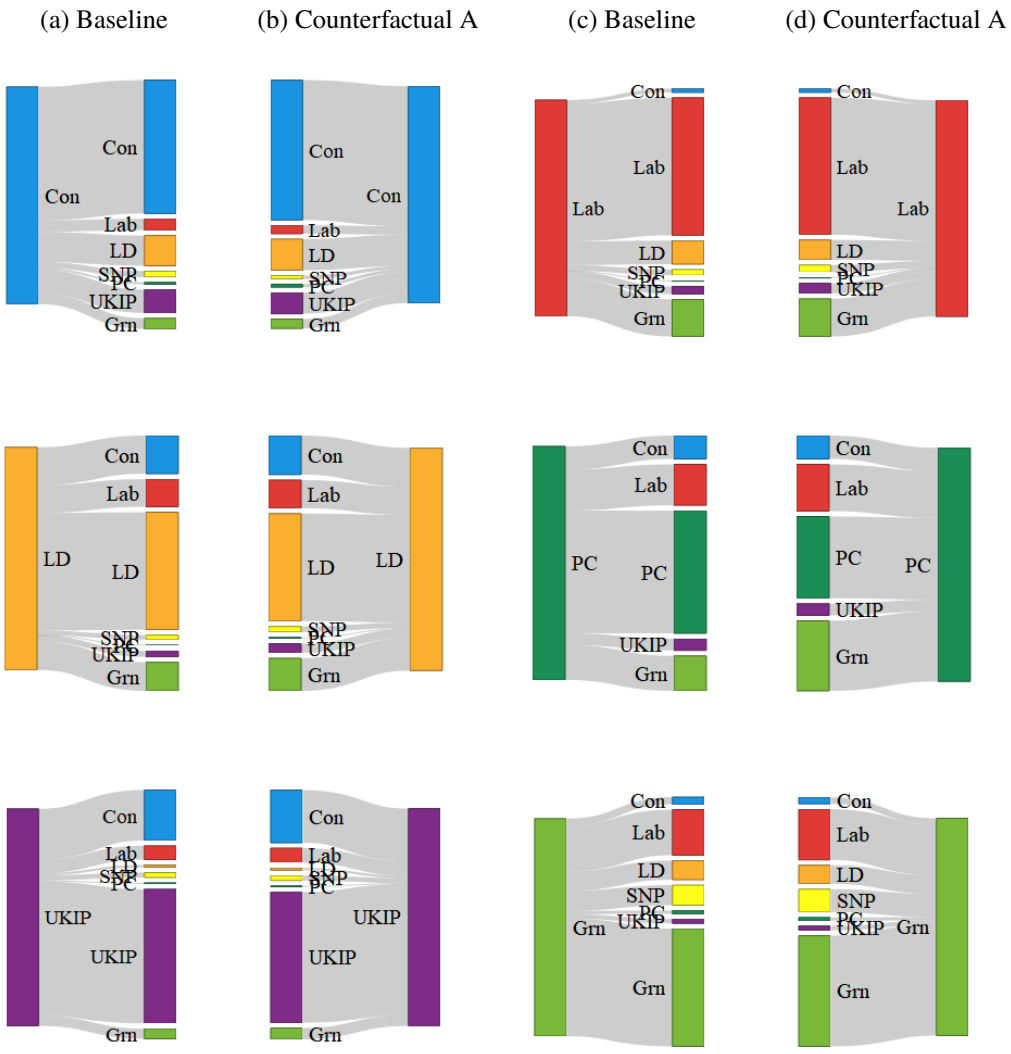
The tuning parameter λ for each year was obtained through 10-fold cross-validation. Confidence intervals are at the 95% level.

Figure 4.A1: Average Partial Effects Estimates



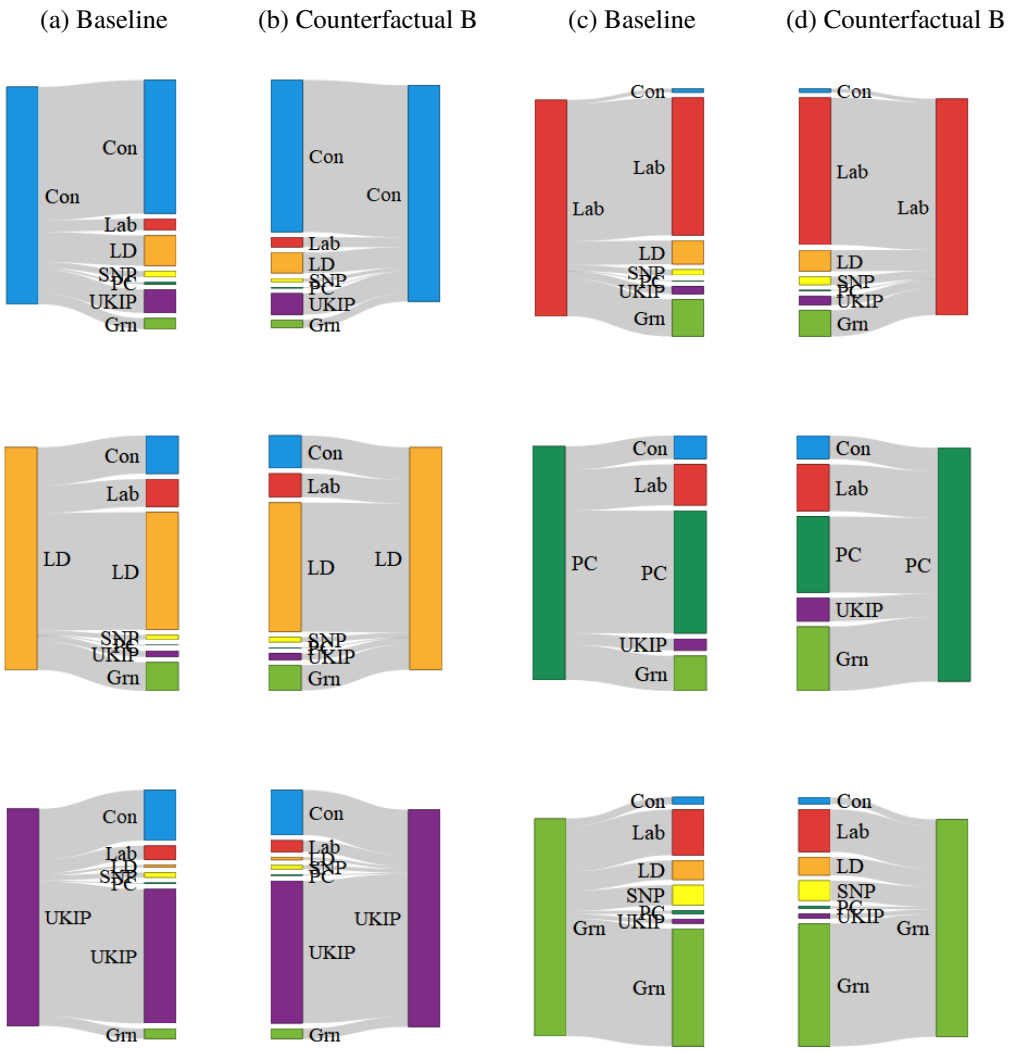
The tuning parameter was obtained through 10-fold cross validation using the Akaike information criterion. Confidence intervals are at the 95% level.

Figure 4.A2: Transitions 2015, Counterfactual A



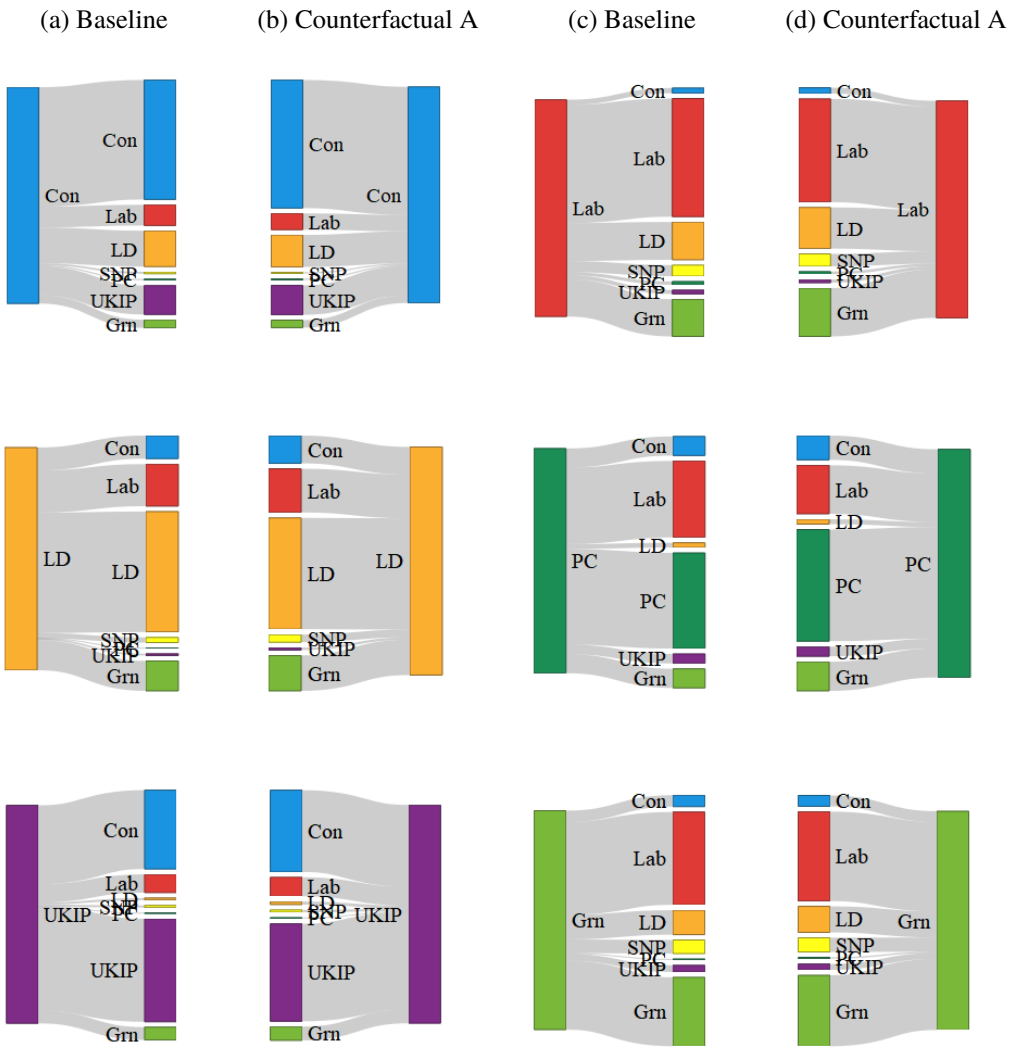
Plots for the SNP are not included as the SNP was viable in all Scottish constituencies

Figure 4.A3: Transitions 2015, Counterfactual B



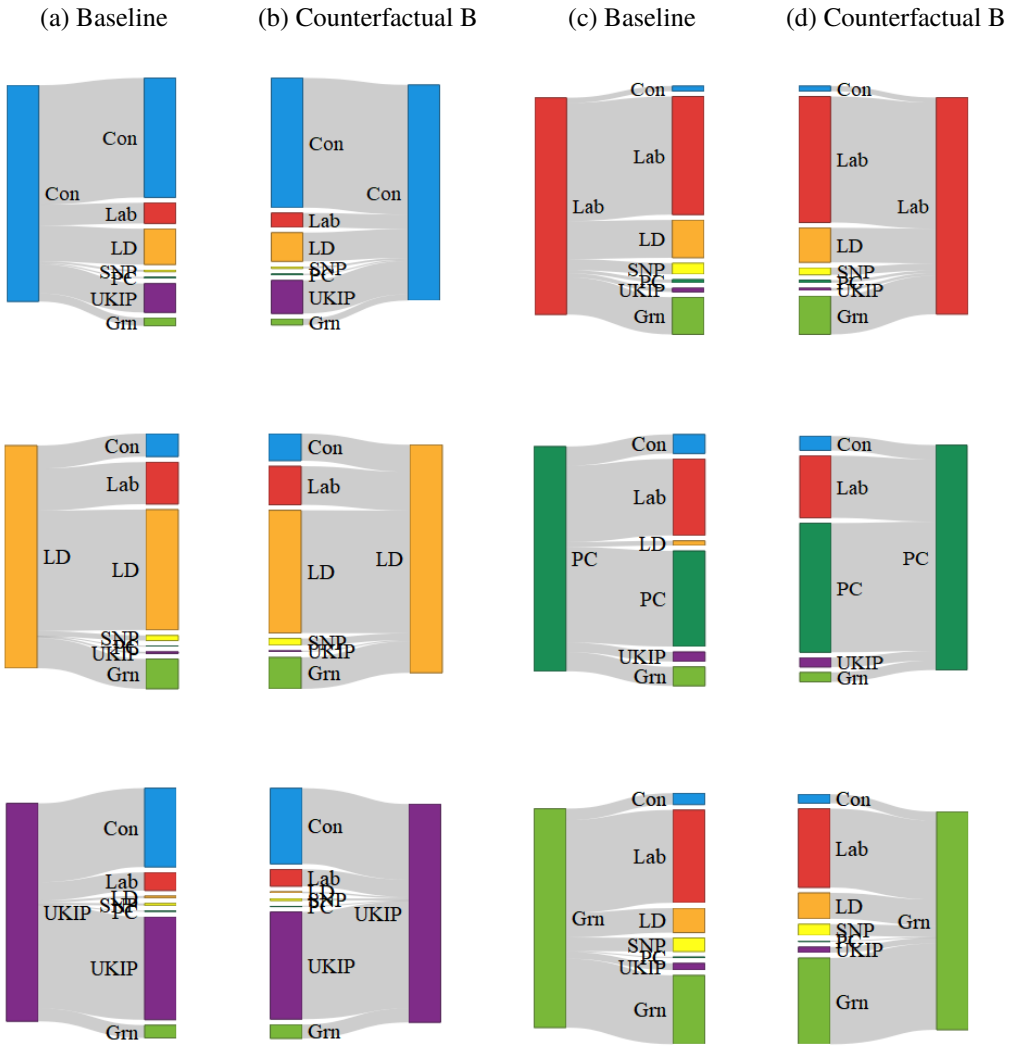
Plots for the SNP are not included as the SNP was viable in all Scottish constituencies

Figure 4.A4: Transitions 2017, Counterfactual A



Plots for the SNP are not included as the SNP was viable in all Scottish constituencies

Figure 4.A5: Transitions 2017, Counterfactual B



Plots for the SNP are not included as the SNP was viable in all Scottish constituencies

4.B Parties' Contact Strategies

For PF-CRE to be valid, it is necessary that the unobserved heterogeneity be time-invariant.[20] A key requirement for this to hold is that parties' contact strategies be time-invariant. That is, it requires that across survey waves the individuals that parties decide to contact have similar observed and unobserved characteristics. While it is not possible to test whether individuals contacted by the parties across the different survey waves have the same unobserved characteristics, it is possible to analyze whether their observed characteristics are actually the same.

To determine whether parties target the *same kinds* of voters across survey waves, I first compare the average characteristics of the voters contacted by their most preferred party in each of the ways, using a t-test for the comparison of means. As the results in Tables 4.B1 and 4.B2 show, there is no evidence that these means differ across waves, suggesting that parties tend to contact the same kinds of voters.[21]

In Table 4.B3 I provide another way to compare parties' contact strategies across survey waves. For each wave I split the sample into training (E_i, for *e*stimation, $i = 1, 2, 3$) and test (T_i for *t*est, $i = 1, 2, 3$) sets. Then, for each training set, I estimate a support vector machine where the outcome is being contacted by the most preferred party and the input variables are all those described in Tables 4.B1 and 4.B2. Then, for each of the support vector machines, I predict the outcomes for each of the test sets. Finally, for each test set, I compare whether the three support vector machines predict similarly (*a* for *a*greement). Formally:

$$f_i(\cdot) = SVM(E_i),\ i = 1, 2, 3 \tag{4.3}$$

$$p_i^r = f_i(T_r),\ i = 1, 2, 3,\ r = 1, 2, 3 \tag{4.4}$$

$$a_{i,j}^r = 1 - \text{mean}(|f_i(T_r) - f_j(T_r)|),\ i \neq j,\ r = 1, 2, 3. \tag{4.5}$$

Thus $a_{i,j}^r$ denotes the percentage of observations in test set r for which the support vector machines estimated with training sets i and j predict the same outcome.

The results in Table 4.B3 show that for both 2015 and 2017, the prediction rules derived from the support vector machines estimated with different training set waves produce similar predictions for each training set wave. For example, for 2015 the support vector machines estimated with the training sets from waves 1 and 2 produce

[20] This is also required for other methods like Fixed Effects and Conditional Maximum Likelihood estimation.

[21] Note, however, that whereas parties tend to contact the same kinds of voters, they do contact significantly more voters in the final survey wave, right before the election.

Table 4.B1: Comparison of Means, Contact by Most Preferred Party 2015

	μ_1	μ_2	μ_3	$t_{1,2}$	$t_{1,3}$	$t_{2,2}$
Age	53.54	54.73	54.25	-0.08	-0.05	0.03
Household Income	7.39	7.16	7.08	0.06	0.09	0.02
Income	348.75	300.32	392.08	0.03	-0.02	-0.05
Household Size	2.41	2.45	2.41	-0.04	-0.00	0.04
Children	0.31	0.33	0.30	-0.03	0.01	0.04
Agreeableness	5.93	6.03	5.95	-0.06	-0.01	0.05
Concientiousness	6.68	6.80	6.77	-0.06	-0.05	0.02
Extroversion	4.15	4.19	4.14	-0.02	0.00	0.02
Neuroticism	3.59	3.68	3.66	-0.04	-0.03	0.01
Openess	5.96	5.84	5.69	0.06	0.15	0.09
Risk Taking	2.65	2.58	2.58	0.11	0.11	0.00
Middle Class	0.40	0.43	0.41	-0.05	-0.01	0.04
Working Class	0.35	0.36	0.36	-0.02	-0.02	-0.00
Employed Full	0.40	0.37	0.35	0.07	0.10	0.04
Employed Part	0.13	0.16	0.15	-0.08	-0.06	0.02
Unemployed	0.03	0.02	0.02	0.03	0.03	-0.01
Student	0.07	0.04	0.05	0.13	0.06	-0.07
Retired	0.28	0.29	0.31	-0.03	-0.08	-0.04
Married	0.63	0.69	0.66	-0.11	-0.06	0.05
Separated	0.09	0.09	0.09	-0.01	0.02	0.03
Widowed	0.03	0.02	0.03	0.04	-0.03	-0.07
Never Married	0.25	0.20	0.22	0.12	0.07	-0.05
Education 1	0.10	0.12	0.11	-0.09	-0.05	0.04
Education 2	0.15	0.15	0.18	-0.00	-0.07	-0.07
Education 3	0.17	0.22	0.19	-0.12	-0.05	0.07
Education 4	0.58	0.51	0.52	0.15	0.12	-0.03
Other Race	0.02	0.02	0.01	0.01	0.05	0.05
Mixed Race	0.00	0.01	0.00	-0.03	-0.00	0.03
Asian	0.01	0.01	0.01	-0.07	-0.05	0.02
Black	0.00	0.00	0.00	-0.02	0.01	0.03
Female	0.37	0.43	0.43	-0.12	-0.12	-0.01
Own Home	0.43	0.46	0.46	-0.06	-0.06	-0.00
Mortgage	0.36	0.39	0.35	-0.07	0.02	0.09
Therm. Most Preferred	8.46	8.51	8.47	-0.04	-0.01	0.03
Therm. Most Pref. Viable	5.51	6.00	6.11	-0.27	-0.34	-0.07

μ_i refers to the mean of the variables in wave i; $t_{i,j}$ is a t-statistic of comparison of means between waves i and j.

the same prediction 86% of the time for the test set from wave 1 in 2015; the vector machines estimated with waves 1 and 3 coincide 90% of the time in that sample;

Table 4.B2: Comparison of Means, Contact by Most Preferred Party 2017

	μ_1	μ_2	μ_3	$t_{1,2}$	$t_{1,3}$	$t_{2,2}$
Age	54.10	52.24	52.07	0.12	0.13	0.01
Household Income	6.90	7.22	6.88	-0.09	0.01	0.09
Income	241.11	284.16	258.84	-0.03	-0.01	0.02
Household Size	2.26	2.31	2.31	-0.05	-0.04	0.01
Children	0.26	0.24	0.26	0.02	-0.01	-0.03
Agreeableness	6.04	6.03	6.05	0.01	-0.00	-0.01
Concientiousness	6.82	6.69	6.69	0.07	0.07	-0.00
Extroversion	4.19	4.14	4.09	0.02	0.05	0.03
Neuroticism	3.62	3.71	3.70	-0.04	-0.03	0.01
Openess	5.68	5.60	5.73	0.05	-0.03	-0.08
Risk Taking	2.58	2.53	2.56	0.07	0.04	-0.03
Middle Class	0.45	0.47	0.42	-0.04	0.05	0.09
Working Class	0.31	0.28	0.32	0.05	-0.04	-0.09
Employed Full	0.33	0.35	0.35	-0.04	-0.04	-0.00
Employed Part	0.14	0.12	0.13	0.05	0.01	-0.04
Unemployed	0.01	0.01	0.02	0.02	-0.04	-0.05
Student	0.04	0.05	0.04	-0.05	-0.03	0.03
Retired	0.32	0.31	0.30	0.01	0.03	0.03
Married	0.61	0.61	0.60	-0.00	0.02	0.02
Separated	0.09	0.10	0.10	-0.02	-0.03	-0.01
Widowed	0.05	0.04	0.04	0.04	0.07	0.02
Never Married	0.25	0.25	0.26	-0.01	-0.03	-0.02
Education 1	0.13	0.11	0.11	0.07	0.08	0.02
Education 2	0.13	0.12	0.16	0.02	-0.08	-0.10
Education 3	0.17	0.17	0.15	-0.02	0.04	0.06
Education 4	0.57	0.59	0.58	-0.05	-0.03	0.02
Other Race	0.01	0.01	0.01	-0.05	-0.03	0.02
Mixed Race	0.00	0.01	0.01	-0.07	-0.08	-0.01
Asian	0.01	0.01	0.01	-0.05	0.00	0.05
Black	0.00	0.00	0.00	-0.00	0.07	0.07
Female	0.43	0.42	0.44	0.01	-0.03	-0.03
Own Home	0.50	0.50	0.49	-0.00	0.02	0.03
Mortgage	0.34	0.31	0.31	0.08	0.07	-0.01
Therm. Most Preferred	8.38	8.27	8.39	0.07	-0.01	-0.08
Therm. Most Pref. Viable	6.01	6.02	6.13	-0.01	-0.07	-0.06

μ_i refers to the mean of the variables in wave i; $t_{i,j}$ is a t-statistic of comparison of means between waves i and j.

and support vector machines estimated with waves 2 and 3 agree 87% of the time. Therefore, this constitutes additional evidence that parties' contact strategies do not

Table 4.B3: Agreement Between Party Strategies

	2015			2017		
	T_1	T_2	T_3	T_1	T_2	T_3
f_1 v. f_2	0.86	0.89	0.86	0.90	0.90	0.89
f_1 v. f_3	0.90	0.85	0.89	0.91	0.88	0.89
f_2 v. f_3	0.87	0.89	0.87	0.89	0.90	0.88

Rows indicate the estimates of which two support vector machines are being compared. Columns indicate the test sets used in each case. The cells indicate the percentage of respondents for which the two vector machines in the corresponding row agree in their contact prediction.

vary significantly across the sample period, for both 2015 and 2017.

Tables 4.B4, 4.B5, and 4.B6 show that the same is true for contact by the most preferred viable party.

Table 4.B4: Comparison of Means, Contact by Most Preferred Viable Party 2015

	μ_1	μ_2	μ_3	$t_{1,2}$	$t_{1,3}$	$t_{2,2}$
Age	54.06	54.56	53.41	-0.03	0.04	0.07
Household Income	6.93	7.41	6.97	-0.14	-0.01	0.13
Income	294.53	230.89	321.14	0.04	-0.02	-0.05
Household Size	2.41	2.41	2.46	0.00	-0.04	-0.04
Children	0.27	0.37	0.33	-0.14	-0.09	0.05
Agreeableness	5.95	6.00	5.88	-0.03	0.04	0.07
Concientiousness	6.76	6.63	6.83	0.07	-0.04	-0.11
Extroversion	4.13	4.17	4.04	-0.02	0.04	0.06
Neuroticism	3.62	3.58	3.56	0.02	0.03	0.01
Openess	5.85	5.75	5.67	0.06	0.11	0.05
Risk Taking	2.58	2.58	2.53	0.00	0.07	0.06
Middle Class	0.41	0.42	0.41	-0.02	-0.02	0.00
Working Class	0.36	0.37	0.37	-0.01	-0.02	-0.01
Employed Full	0.37	0.41	0.39	-0.08	-0.04	0.04
Employed Part	0.15	0.17	0.15	-0.04	0.00	0.04
Unemployed	0.03	0.02	0.02	0.07	0.03	-0.04
Student	0.06	0.04	0.06	0.08	-0.01	-0.09
Retired	0.31	0.26	0.28	0.11	0.05	-0.06
Married	0.64	0.68	0.67	-0.07	-0.05	0.02
Separated	0.10	0.09	0.08	0.03	0.09	0.06
Widowed	0.03	0.03	0.03	0.03	0.01	-0.02
Never Married	0.22	0.20	0.23	0.05	-0.01	-0.06
Education 1	0.11	0.12	0.10	-0.02	0.05	0.07
Education 2	0.16	0.12	0.17	0.13	-0.02	-0.15
Education 3	0.15	0.22	0.19	-0.17	-0.09	0.08
Education 4	0.57	0.54	0.54	0.05	0.06	0.01
Other Race	0.01	0.01	0.01	-0.05	0.03	0.09
Mixed Race	0.01	0.01	0.01	0.03	0.09	0.06
Asian	0.01	0.01	0.01	0.06	-0.00	-0.06
Black	0.00	0.01	0.00	-0.10	-0.04	0.06
Female	0.43	0.43	0.45	-0.00	-0.04	-0.04
Own Home	0.45	0.46	0.46	-0.02	-0.03	-0.01
Mortgage	0.33	0.39	0.36	-0.14	-0.06	0.07
Therm. Most Preferred	8.08	8.28	8.32	-0.13	-0.16	-0.03
Therm. Most Pref. Viable	5.80	6.12	6.29	-0.17	-0.28	-0.10

μ_i refers to the mean of the variables in wave i; $t_{i,j}$ is a t-statistic of comparison of means between waves i and j.

Table 4.B5: Comparison of Means, Contact by Most Preferred Viable Party 2017

	μ_1	μ_2	μ_3	$t_{1,2}$	$t_{1,3}$	$t_{2,2}$
Age	52.32	51.25	51.66	0.07	0.04	-0.03
Household Income	6.94	6.72	6.79	0.06	0.04	-0.02
Income	276.20	339.97	265.13	-0.04	0.01	0.04
Household Size	2.28	2.35	2.29	-0.06	-0.00	0.06
Children	0.31	0.31	0.29	0.00	0.04	0.04
Agreeableness	6.13	6.19	6.02	-0.03	0.06	0.10
Concientiousness	6.70	6.75	6.74	-0.02	-0.02	0.00
Extroversion	3.90	3.94	4.01	-0.02	-0.05	-0.03
Neuroticism	3.81	3.72	3.83	0.04	-0.01	-0.05
Openess	5.78	5.78	5.74	0.00	0.02	0.02
Risk Taking	2.50	2.56	2.52	-0.08	-0.02	0.06
Middle Class	0.40	0.40	0.39	0.01	0.03	0.02
Working Class	0.36	0.38	0.35	-0.05	0.01	0.05
Employed Full	0.35	0.37	0.38	-0.03	-0.06	-0.03
Employed Part	0.14	0.14	0.12	-0.01	0.05	0.06
Unemployed	0.02	0.02	0.02	0.05	0.04	-0.01
Student	0.04	0.03	0.04	0.02	-0.00	-0.03
Retired	0.29	0.29	0.29	0.01	0.02	0.01
Married	0.60	0.61	0.60	-0.01	0.01	0.02
Separated	0.09	0.10	0.10	-0.04	-0.04	0.00
Widowed	0.04	0.03	0.03	0.11	0.07	-0.03
Never Married	0.26	0.27	0.27	-0.01	-0.01	-0.01
Education 1	0.12	0.10	0.12	0.04	-0.01	-0.05
Education 2	0.14	0.16	0.16	-0.04	-0.04	-0.00
Education 3	0.15	0.17	0.15	-0.03	0.02	0.05
Education 4	0.59	0.58	0.58	0.03	0.03	-0.00
Other Race	0.01	0.01	0.01	0.04	0.06	0.02
Mixed Race	0.01	0.01	0.01	-0.00	0.02	0.02
Asian	0.01	0.01	0.01	-0.04	-0.01	0.03
Black	0.01	0.01	0.00	0.00	0.11	0.10
Female	0.45	0.47	0.46	-0.05	-0.03	0.02
Own Home	0.48	0.47	0.48	0.02	-0.01	-0.04
Mortgage	0.33	0.33	0.32	0.02	0.02	0.01
Therm. Most Preferred	8.29	8.22	8.32	0.05	-0.02	-0.07
Therm. Most Pref. Viable	6.32	6.38	6.36	-0.04	-0.03	0.01

μ_i refers to the mean of the variables in wave i; $t_{i,j}$ is a t-statistic of comparison of means between waves i and j.

Table 4.B6: Agreement Between Party Strategies, Most Preferred Viable Party

	2015			2017		
	T_1	T_2	T_3	T_1	T_2	T_3
f_1 v. f_2	0.88	0.89	0.85	0.89	0.89	0.88
f_1 v. f_3	0.89	0.87	0.88	0.90	0.88	0.89
f_2 v. f_3	0.86	0.88	0.88	0.88	0.90	0.88

Rows indicate the estimates of which two support vector machines are being compared. Columns indicate the test sets used in each case. The cells indicate the percentage of respondents for which the two vector machines in the corresponding row agree in their contact prediction.

4.C Multi-level Regression and Poststratification

To estimate constituency level support for the different parties (and to derive an estimated winner in each constituency), I use multi-level regression and poststratification (MRP). MRP has been developed to generate good estimates of public opinion in subnational areas based on national survey samples.[22] In MRP, I first estimate a model of vote choice using survey individual demographics together with constituency-level information. I then use this model to predict the voting behavior for each demographic-constituency type. Finally, I use poststratification to true population characteristics to estimate the level of support for each party at the constituency level.

Multi-level Model

To implement the multi-level regression I use five individual demographic characteristics: age (9 categories), gender (2 categories), home ownership (3 categories), ethnicity (2 categories), and qualifications (6 categories). I also use the following constituency level information: an indicator for the party that actually won the constituency, the actual vote shares that the parties received in the constituency, the percentage of long-term unemployed, the percentage in industry manufacturing, population density, indicators for whether parties are standing in the constituency. With this information, I estimate the following model:

$$P(y_i^j = 1) = \Lambda(\beta_0 + \alpha_{a[i]}^{age} + \alpha_{g[i]}^{gender} + \alpha_{h[i]}^{home} + \alpha_{e[i]}^{ethnic} + \alpha_{q[i]}^{qual} + \alpha_{c[i]}^{const}), \qquad (4.6)$$

where y_i^j denotes where respondent i intends to vote for party j, $a[i]$ denotes respondent i's *a*ge group, $g[i]$ denotes respondent i's *g*ender, $h[i]$ denotes respondent i's *h*ome ownership status, $e[i]$ denotes respondent i's *e*thnicity, $q[i]$ denotes respondent i's highest *q*ualifications, and $c[i]$ denotes respondent i's constituency.

Each of the demographic effects is drawn from a normal distribution with mean zero

[22] See, for example, Park, Gelman, and Bafumi (2004), Hanretty, Lauderdale, and Vivyan (2016), Leemann and Wasserfallen (2017), Lax and Phillips (2009), and Warshaw and Rodden (2012).

and some (estimated) variance:

$$\begin{aligned}
\alpha_a^{age} &\sim \mathcal{N}(0, \sigma^2_{age}),\ a = 1, ..., 9 \\
\alpha_g^{gender} &\sim \mathcal{N}(0, \sigma^2_{gender}),\ g = 1, 2 \\
\alpha_h^{home} &\sim \mathcal{N}(0, \sigma^2_{home}),\ h = 1, 2, 3 \\
\alpha_e^{ethnic} &\sim \mathcal{N}(0, \sigma^2_{ethnic}),\ e = 1, 2 \\
\alpha_q^{qual} &\sim \mathcal{N}(0, \sigma^2_{qual}),\ q = 1, ..., 6.
\end{aligned} \tag{4.7}$$

The constituency effects are modeled as a function of the constituency characteristics:

$$\alpha_c^{const} \sim \mathcal{N}(\beta_0^{const} + \beta X, \sigma^2_{const}), \tag{4.8}$$

where X is a vector of constituency characteristics.

I estimate this model for England, Scotland, and Wales separately. Estimation for each country is performed using one-versus-the-rest. The left out party is the Conservative party in England, the SNP in Scotland, and Labour in Wales.

Once the model is estimated, I predict the voting probabilities of each demographic group in each constituency for each of the parties:

$$\widehat{p}_{gjc}, \tag{4.9}$$

where g denotes each of the 648 distinct demographic groups that arise from combining the demographic characteristics.

Poststratification

Performing MRP requires knowledge of the joint population distribution of all poststratifying variables for every constituency to generate the poststratification weights. However, this information is not available from the U.K. Census at the constituency level. For this reason, I combine two sources of information to impute the joint distribution for each constituency. First, I use data from the Annual Population Survey (APS), which provides data on over 300,000 respondents at the national level. Second, I use the constituency level marginal distributions of the demographic variables of interest obtained from Nomis Census Area Statistics.

To impute the joint distribution for each constituency, I rake the subsample of the APS corresponding to the region where the constituency lies.[23] The outcome of this raking is an imputation of the joint distribution of the demographics of interest for each constituency: π_{gc}, where c denotes the constituency and g denotes the demographic group. This procedure is related to those used by Hanretty, Lauderdale, and Vivyan (2016) and Leemann and Wasserfallen (2017).

To obtain the estimates of vote intention for each party in each constituency, I perform the following calculation:

$$\widehat{p}_{jc} = \frac{\sum_{g} \widehat{p}_{gjc}\pi_{gc}}{\sum_{g} \pi_{gc}}, \tag{4.10}$$

where $\widehat{p}_{jc}$ is the vote intention for party j in constituency c.

MRP Uncertainty

There are two sources of uncertainty in the estimation of the seat changes under the counterfactuals. The first is the estimation of the multilevel regression; this source applies both for the baseline and the counterfactuals. The second is from the individual counterfactuals that are then used to estimate the multilevel regression for the seat counterfactuals.

To incorporate the first source of uncertainty, I take draws from the posterior distribution of the corresponding multilevel regression model and re-poststratify each time to obtain predicted vote shares and seats. To obtain seat changes between the baseline and the counterfactual, I compare each draw from the baseline and the corresponding counterfactual. Some seat changes have a skewed distribution, which means that 95% confidence intervals will not be centered around the point estimate.

The second source of uncertainty is more challenging to incorporate. One way to do it is to take draws from the causal model in equation 4.2; then, for each draw, run a new MRP. However, this process is highly computationally demanding. Therefore, I do not incorporate this source of uncertainty to the estimates of seat counterfactuals reported in the paper.

[23] These regions are Scotland, Wales, and the 9 Government Office Regions of England: East Midlands, East of England, London, North East England, North West England, South East England, South West England, West Midlands, and Yorkshire and the Humber.