Estimating group fixed effects in panel data with a binary dependent variable: how the LPM outperforms logistic regression in rare events data

Joan C. Timoneda

University of Maryland, College Park

Abstract

Estimating fixed effects models can be challenging with rare events data. Researchers often face difficult trade-offs when selecting between the Linear Probability Model (LPM), logistic regression with group intercepts and the conditional logit. In this paper, I survey these trade-offs and argue that, in fact, the LPM with fixed effects produces more accurate estimates and predicted probabilities than maximum likelihood specifications when the dependent variable has less than 25 percent of ones. I use Monte Carlo simulations and two examples from the literature to show when the LPM with fixed effects should be preferred. I perform these simulations on common panel data structures found in the literature as well as big data. This paper provides clarity around fixed effects models in panel data and a novel technique to identify which one to use as a function of the frequency of events in y.

Word Count: 9,942

1. Introduction

In this article I address common issues with fixed effects estimation in panel data with a binary outcome variable. The three most common techniques used in political science to estimate fixed effects are the conditional logit (CL), the logit with dummies (LD), and the linear probability model (LPM) with fixed effects (LPMFE). Currently, the choice of method varies by subfield: scholars in political economy generally favor the LPMFE (see Acemoglu et al. 2008) and democratization and conflict scholars tend to prefer the CL (Miller 2012). Yet confusion remains regarding the benefits and drawbacks of these techniques as well as *when* to apply them. In this paper, I suggest that the decision on which method produces the most accurate estimates depends on the number of events in the dependent variable. I show that both maximum likelihood (ML) methods are best suited for dependent variables whose number of events is between 25 and 75 percent, provided the average number of observations per group is at least over 30. The LPMFE, on the other hand, performs much better with fewer than 25 percent of occurrences.[1] I also show that researchers can use the LD to produce predicted probabilities if certain conditions are met.

Whence the confusion with fixed effects? In panel data with a binary outcome variable, such models often require certain sacrifices that researchers are unwilling to make. With rare events data and fixed effects, a substantial portion of the sample may be lost in ML models since those groups without variation in the dependent variable are dropped. Coefficients are also notoriously inaccurate when large numbers of covariates are added to the model. The deficiencies of the LPM, on the other hand, are equally well-known: heteroskedasticity, predicted probabilities outside the 0-1 range, and the imposed linearity assumption. Because of the glaring deficiencies of both approaches with regards to fixed effects specifications, many researchers cite simplicity of interpretation as the

[1] Or, trivially, more than 75 percent of ones. Henceforth I only make reference to rare events, but the logic applies to common events by trivial recoding.

reason for using the LPMFE or 'best practices' when using ML (Miller 2012, Harding and Stasavage 2014).

While the reasoning is justified, I argue that this is the wrong conclusion. Fixed effects estimation can be difficult with a majority of available datasets, but the right steps can be taken to produce reliable results. In this paper, I seek to provide some clarity around existing fixed effects specifications. I argue that doing so is theoretically and empirically important, either as the main model or to add robustness to the results. Second, and most importantly, I provide a new technique by which researchers can identify the fixed effects model that best suits their data according to the frequency of events in y. When observations per group are above 30, the ML and LPMFE models produce practically identical predicted probabilities *when the proportion of events in the sample is around 50 percent.* When the percentage of ones (or, trivially, zeros) is between 25 and 50, ML performs only slightly better than the LPMFE —primarily because nonlinear models are a bit more accurate at extreme values. Below 25 percent of events or non-events, however, the LPMFE produces predicted probabilities much closer to the observed probability for a majority of the distribution. Therefore, while the bias of ML with rare events is well known (King and Zeng 2001), bias occurs even when the number of events is relatively large (below 25 percent) when fixed effects are used. Indeed, the most important finding of this article is that, in cases of true rare events, i.e. when the number of ones is 1 percent or less, the LPMFE *is* the best method. Third, I show that logistic regression with dummies performs better than expected in big data analysis with a large number of both observations and groups. Despite this counterintuitive finding, researchers using big data should still use the LPM with fixed effects if the outcome variable is binary.

The structure of the paper is as follows. I first provide some theoretical reasons why using fixed effects is important and can help us identify false positives in our results. I follow this section with a discussion of the benefits and pitfalls of the three most common fixed effects models in the literature mentioned above —the LPMFE, the LD and the CL. I then use evidence from Monte Carlo experiments and empirical examples from the literature to establish that the LPM with fixed effects outperforms the LD and the CL in a majority of cases, especially in rare events data.

2. Fixed Effects: Theoretical Background and Notation

2a. Why Should We Use Fixed Effects?

Comparisons across countries based on panel data, which are pervasive in comparative politics, are particularly susceptible to omitted variable bias, as differences among countries are substantial and difficult to model. Fixed effects models provide a good, if blunt, solution to this problem. Researchers who run the fixed effects equivalent of their model and find that their main coefficient of interest is *smaller* should be particularly alert to omitted variable bias (Allison 2009). Second, comparative and IR theories are usually based on interactions among key actors of different groups –governments, opposition parties, and rebel groups, for instance. Surprisingly, however, within-group variance is not often the focus of the models used in these subfields, even if the theories largely describe domestic change. Fixed effects models help us reach conclusions using within-group variance rather than simply exploiting variation between groups.

Even if we agree on the importance of using fixed effects in cross-national research, questions remain about *how* and *when* to apply fixed effects. Should researchers use the CL, the LD, or the LPMFE? Theoretically speaking, since the dependent variable is limited, ML should be preferred over ordinary least squares. So one option is to add group dummy variables to a standard logit model. However, this solution has two problems. First, if the number of group variables or intercepts increases with the sample size, ML's asymptotic assumptions are violated and coefficients are biased. This is known as the 'incidental parameters problem' (Neyman and Scott 1948; see Section 3), which is particularly acute in dynamic models (Heckman 1981, Wooldridge 2002). Second, adding group dummies leads to systematic but small upward bias in the coefficients. This bias is the result of calculating many group intercepts – as is well known, logit models are sensitive to the addition of many parameters (Beck 2015). The bias does not stem from the violation of any core assumptions and, in this sense, is less problematic.

A solution to *both* of these problems is to use the CL (Chamberlain 1980; see Heckman 1981). The CL produces a single coefficient that represents the log likelihood of observing y conditional on observing x for those groups that at least have one occurrence of x and variation in y. This makes the calculation more efficient and eliminates the incidental parameters problem. However, while the CL, at least in theory, might be the most appropriate method to perform fixed effects analysis in ML, it is especially complicated to use in practice. Predicted probabilities and confidence intervals often are difficult to compute or produce unreasonable results given our theory. Predicted probabilities help us understand substantive significance, i.e. whether the effect of increases in economic equality on democratic transitions is politically meaningful (Achen 1982). The issues described above make this process difficult (see Woldridge 2002).

The LPM with group intercepts is an attractive alternative to ML models (Beck 2015). While unconventional with limited dependent variables, the LPMFE presents two key advantages. One is that it allows groups with no variation in y to be included in the estimation, mitigating the logit's issues with reduced sample size (see Section 5b for a practical discussion of this important modeling choice). Another benefit is ease of interpretation, both for the coefficients as well as the predictions. The main complication, however, is the linearity assumption, which can generate inaccurate probabilities at particular points of x even when the range of the predictions is accurate. Moreover, at extreme values, the LPMFE is likely to underperform. It may predict values over one or under zero or force linearity on a nonlinear relationship.

2b. Three Models To Estimate Fixed Effects: The LPMFE, the Conditional Logit, and the Logit *with Group Specific Intercepts*[2]

In logit specifications with fixed effects, groups without variation in the dependent variable do not enter the likelihood function. For instance, if a country never experienced a transition in the period under study, and thus the dependent variable is all zeroes, the coefficient for the group cannot be estimated due to lack of variation. Another way to see this is that the group dummy is perfectly correlated with the dependent variable, and the group is therefore dropped. The LPM specification with fixed effects does incorporate all groups into the estimation, as variation in y is not a requirement of OLS. The most interesting aspect of this specification is that the $\boldsymbol{\beta}$ is a weighted average of two sets of groups, those countries with variation in the DV and those which are all zeros (Beck 2015). For those groups where $y_{g,i} = 0$, the estimate $\boldsymbol{\beta}$ is zero. For all others, it is the coefficient produced for all countries where there is variation in the DV. To see this process, let us take first the general equation for fixed effects and plug in a linear function such that

$$\begin{aligned} y_{g,i} &= x_{g,i}\beta + f_{g'} I_{g'=g} + \epsilon_{g,i} \\ P(y_{g,i} = 1) &= E(x_{g,i}\beta + f_{g'} I_{g'=g} + \epsilon_{g,i}) = x_{g,i}\hat{\beta} + \hat{f}_i I_{g=k}. \end{aligned} \tag{1}$$

I is an indicator function for group membership. OLS can calculate these probabilities, which may extend beyond the zero-one range, under the assumption that the error term $\epsilon_{g,i}$ satisfies the Gauss-Markov assumptions. Now, to see how OLS averages out the coefficients for groups with no variation (hereafter $all0$) in the DV and those with variation ($allV$), I show the equations for the OLS estimate for $\hat{\beta}_{all}$ and $\hat{\beta}_{allV}$ –for the entire sample and for only those that vary, respectively. I use the scalar case x, with all variables group mean centered, such that

[2] This discussion is based fundamentally on Angrist and Pischke (2008), Beck (2015) and Hosmer, Lemeshow and Sturdivant (2013).

$$\hat{\beta}_{all} = \frac{\sum (x_{allV,i})(y_{allV,i})}{\sum x_{all,i}^2} \tag{2}$$

$$\hat{\beta}_{allV} = \frac{\sum (x_{allV,i})(y_{allV,i})}{\sum x_{allV,i}^2}. \tag{3}$$

The only difference in the two equations is that the denominator in (2) includes all values of x per group, while (3) includes only values for those groups in which there is variation in y. Given that the denominator includes values of the independent variable for groups $all0$, but the numerator does not, the coefficient $\hat{\beta}_{all}$ is a weighted average of the coefficient of both groups —it is as if the numerator had a second summation that returned zero. It is intuitive now that coefficients for OLS incorporate cases without variation in the DV, and will be smaller than if those cases were dropped. The standard errors for the two equations are given by

$$SE_{\beta_{all}} = \sqrt{\frac{\widehat{\sigma_{all}^2}}{\sum x_{all,i}^2}} \quad ; \quad SE_{\beta_{allV}} = \sqrt{\frac{\widehat{\sigma_{allV}^2}}{\sum x_{allV,i}^2}}. \tag{4}$$

These equations show that, while the coefficients will be smaller because $all0$ groups are taken into account, the standard errors will be smaller, as well. This is because the numerator in (4) left includes all observations, while the one in (4) right includes only those from the $allV$ groups. Therefore, statistical significance will be affected to a lesser extent with the inclusion of $all0$ groups. Substantively, however, the differences are large, as the predicted probabilities will certainly be smaller. This is, in my view, a strong suit of the LPMFE, as it does not statistically select on those countries that have variation in the DV. Some argue, however, that the LPMFE underestimates the coefficients by adding *all*0 groups (Beck 2015). But I contend that this is akin to selecting on the dependent variable and the LPMFE allows us to avoid statistical selection bias. If China has not transitioned to democracy as inequality has increased because, let us imagine, repression has been particularly effective, why should we not include China in an analysis of the effects of repression on democratic transitions? This logic is even starker if we take the argument to its logical extreme: if only one country had transitioned to democracy in the period under study, we could not accept an argument that based its evidence on only one case. 'Dropping' cases, therefore, is logically difficult to justify, mostly when the ones that remain do so because our outcome of interest is observed.[3] This is a more complex problem when events are rare and, therefore, the average is weighed down heavily by those groups without variation. I will show how to best deal with that in the next section.

For logistic regression with group dummies, the coefficients are given by

$$P(y_{g,i} = 1) = \frac{1}{1 + e^{-(x_{g,i}\beta + f_{g'} I_{g'=g})}} \,, \tag{5}$$

where the probability is estimated per group first, and therefore it cannot be computed for those groups that do not vary in y. The average coefficient for all groups will therefore include only *allV* groups. In rare events data, this estimation process complicates the reliability of the results, as a majority of groups drops from the sample, leading to overestimated coefficients, small samples and, consequently, larger standard errors.

The last method is another ML technique, the conditional logit, which uses the true conditional likelihood of observing an outcome given a set of parameters provided at least one event is observed (Chamberlain 1980). Rather than providing a group specific intercept for all groups in the regression,

[3] This is, indeed, a philosophical difference that does not affect the results or the application of the results of this paper. It is relevant, however, because in rare events *many* groups are usually dropped for lack of variation in the DV. I will develop this argument further in my first case study of the literature.

the conditional logit assumes the fixed effect to be zero, and computes the conditional probability of observing an event. No intercept is therefore provided in the model. Let g denote groups, T_g the total number of observations for each group g, and y_{gt} is the value of the dependent variable, which takes on values of 0 and 1. Then,

$$k_{1g} = \sum_{t=1}^{T_g} y_{gt} \ , \tag{6}$$

where k_{1g} is the number of observed events in group g. What the conditional logit then calculates is the probability of y_g conditional on $\sum_{t=1}^{T_g} y_{gt} = k_{1g}$ such that

$$P(y_g | \sum_{t=1}^{T_g} y_{gt} = k_{1g}) = \frac{exp(\sum_{t=1}^{T_g} y_{gt} x_{it} \beta)}{\sum_{d_g \in S_g} exp(\sum_{t=1}^{T_g} d_{it} x_{it} \beta)} \ , \tag{7}$$

where d_{gt} is 0 or 1 with $\sum_{t=1}^{T_g} d_{it} = k_{1i}$, and the term S_g represents all possible combinations of ones and zeros per group (Hosmer, Lemeshow and Sturdivant 2013, eq. 7.4). Without going into more estimation details, what can be seen from this equation is that the conditional logit does not estimate a constant or different group specific intercepts, making it more robust to incidental parameters such as country dummies. In doing so, however, it also complicates the estimation of predicted probabilities, upon which political scientists rely for interpreting the substantive significance of statistical results.

Why, then, can we not just use the LD? The answer is that most applied researchers, in fact, are able to use the LD because the asymptotics of a majority of models are not in the group. Others who work with individual-level survey data, however, may be violating the incidental parameters problem.

3. The Incidental Parameters Problem and the Conditional Logit

The incidental parameters problem, identified by Neyman and Scott (1948), is a central issue to ML estimation with fixed effects: when the number of parameters to be added grows with the number of observations, the estimates can no longer converge to the population parameter with shrinking standard errors as the sample size increases. Therefore, the asymptotic assumption at the core of maximum likelihood can no longer be maintained. Andersen (1973) and Chamberlain (1980) showed that using a conditional ML estimator solved the issue by assuming the fixed effect to be zero and, therefore, the number of added parameters does not affect the estimates.

The key question to ask, then, is whether the types of data that this paper relates to are susceptible to the incidental parameters problem. That is, whether the number of observations grows with the number of parameters as we approach infinity. The answer is no. Most datasets have their asymptotics at the group level, and the most common groups are countries, districts or regions in comparative politics or congresses, states and districts in American politics. Such groups in panel data cannot grow too large to generate a problem, and it is rather more likely that the estimation is 'fixed' in both time and group. The complication emerges with individual-level data, such as survey data, and we estimate a model using fixed effects at the *individual* level. If the asymptotics are in the individual, then we do have an incidental parameters problem and are forced into conditional estimation in ML.

Therefore, if the asymptotics are at the group level, and the group is fixed, researchers can use the LD for estimation, and one can make a strong case that there is no incidental parameters problem. However, special attention needs to paid to the fact that the logit with dummies produces estimates that are slightly overstated. This bias stems from the fact that the logit with dummies actually estimates all the fixed effects in the model, that is, a coefficient is produced for each of the 'incidental' group parameters. Since, as is well known, ML is susceptible to large numbers of independent variables, adding a large number of groups introduces bias, but this bias is generally small and in a clear direction, as compared to the conditional logit.

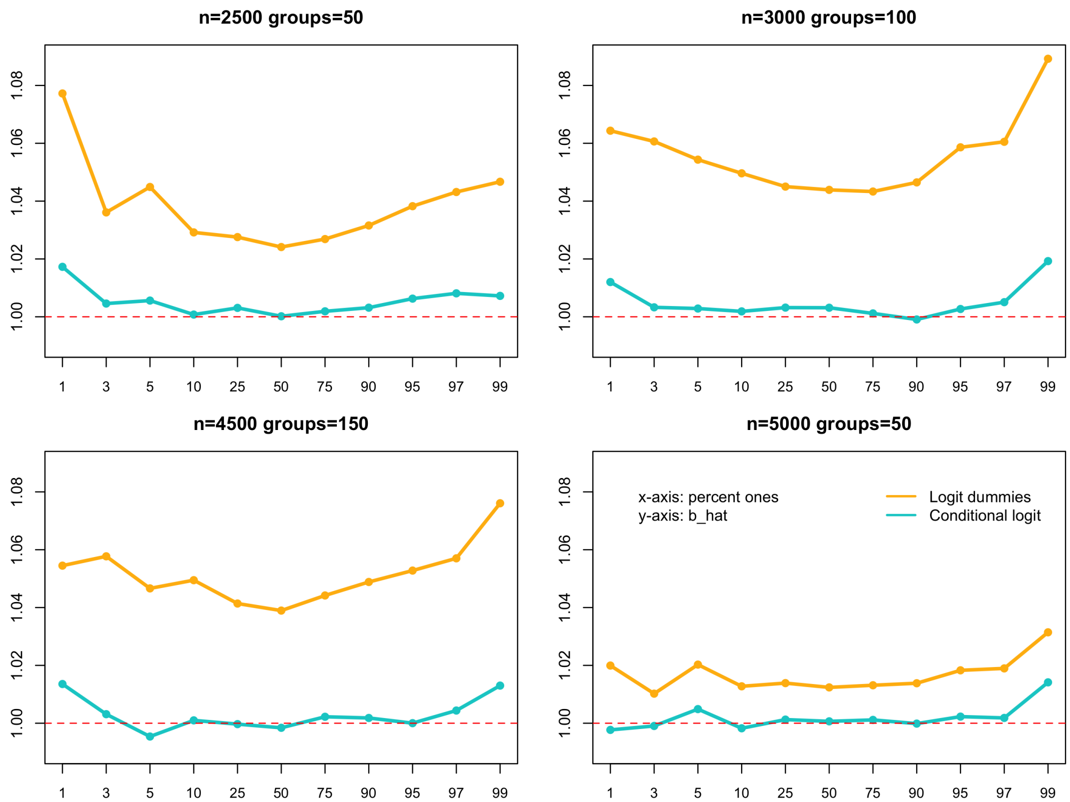


Figure 1: Conditional Logit Coefficients versus Logit with Dummies

Figure 1 shows the bias of the logit with dummies and the accuracy of the conditional logit. It draws from evidence obtained through four of the Monte Carlo simulations performed for the analysis section of this paper. The title of each plot shows the dataset structure used in each simulation —sample size and the number of groups in each dataset. Eleven simulations are performed by dataset, each one increasing in the proportion of events observed in y in the data. They range from rare events (1 percent) to highly common events (99 percent), and constitute the x-axis. In each of these simulations, I obtained average coefficients for the LD and the CL (blue and gold lines). 1 is the value of the β_1 coefficient in the true model in all the simulations. We can see how even with a large number of observations per group (100; bottom right), the logit with dummies produces consistently biased estimates around 1.02 when the true value of the parameter is 1. When the data is rare events (the points furthest to the left and right of the lines in each plot) the coefficients become more biased, as expected. The CL, on the other hand, produces unbiased coefficients throughout, except when the data is true rare events, even though the bias is always smaller than in the LD. Plot (b) shows that the tendency is exacerbated when the number of observations per group decreases to around 30. Here the logit with dummies produces biased coefficients around the 1.05 mark when events are not rare, and between 1.06 and 1.08 when they are. The conditional logit, on the other hand, is much more reliable, with the exception of true rare events, where coefficients tend to also be biased upward.

The problem, however, with the conditional logit lies in calculating predicted probabilities. To do so, one must *assume* that the fixed effect is zero. As will be shown later, doing this can lead to unrealistic probabilities in certain applications. Also, since it does not estimate a baseline, the model's prediction when x is at its lowest point is often overstated in rare events. Thus, even if conditional logit coefficients are better estimates, they are a hindrance for substantive significance (see Achen 1982). I contend that the LPMFE, when events in y are infrequent, solves many of these issues and is the appropriate tool for applied researchers.

4. The Problems With Fixed Effects Regression Using ML and the LPMFE

Multiple issues arise when we estimate fixed effects with ML. First is the loss of data that originates in ML estimation for those groups that have no variation in the DV. As shown in Section 2, no ML estimate can be produced for those groups with no variation in the DV. These groups are then dropped from the final sample. Conversely, the LPMFE produces 0 for such groups and the observations are kept. In rare events, the loss of data in ML can represent a problem, as a large portion of the data is often lost. Relatively large samples with over 30 observations per group can be turned into small samples of under one thousand observations, where we know ML estimates are biased.

A related issue, in the author's view, is the implicit statistical selection bias in choosing only those cases that have variation in the dependent variable. If, let us say, a country with very high levels of inequality *never* transitioned to democracy, it would be removed from the estimation in ML models. Yet the substantive importance of this one country cannot be questioned. I refer to this process as 'statistical' selection bias. Knowingly eliminating groups that *are at risk of experiencing an event* but which did not actually experience it in the data is conceptually problematic. As King and Zeng (2001) show, for extreme cases of rare events, one may justify this statistical selection on the DV on certain solid theoretical grounds. But, I argue, doing so simply for the reason that a group did not experience an event is weak. Again, this is not an issue with the LPMFE, which constructs an average for those countries with no variation in the DV, yielding zero, and a separate estimate for those that do vary.

Moreover, the accuracy of the coefficients decreases in ML when a large number of covariates is added to the model. Even with relatively large ratios of observations per group (which usually need to be at least above 30), ML coefficients become increasingly biased as we add covariates to the model (Beck 2015). Another issue, which we saw in Section 3, is the difficulty in calculating predicted probabilities and, more importantly, marginal effects with Chamberlain's conditional logit. Assessing substantive significance can be complex for researchers that do not possess ideal datasets –very large samples with a lot of observations per group and strong variation in the DV.

The LPMFE is not without its own issues. First, as it fits a linear model on a probability space, it may produce nonsensical predictions over 1 and under 0. Second, it explicitly violates the heteroskedasticity assumption, since errors are by definition not randomly distributed across observations. Third, it issues predictions linearly, which means that predicted probabilities at different points may be understated or overstated if the relationship is indeed nonlinear –and we may never know. There are some interesting solutions to get around these problems: we can simply assume that nonsensical probabilities are simply close to 0 or 1, and we can use a Huber-White (1980) correction for heteroskedasticity. The linearity assumption is certainly the hardest to 'correct', but I suggest that we can confidently apply the LPMFE and its linearity assumption *as a function of the occurrence of events in the DV*. If events are rare or highly common, or if the number of ones is around 50 percent, the LPM will map on linearly to parts of the logistic CDF and produce reliable probabilities. I now proceed to the analysis section, where this argument should become clearer.[4]

5. Analysis

I use Monte Carlo simulations to compare the performance of the LPMFE and the LD in calculating predicted probabilities. I do not use the CL here, for two reasons. First, because it does not estimate the fixed effect or a constant term and its predicted probabilities are generally inaccurate, a problem which is exacerbated with rare events data (see Section 5b for an example of this from the literature). Therefore, it is difficult to interpret the substantive significance of our results using the CL as our main model, but it serves as a useful robustness check given its accurate coefficients. Second, because in substantive comparative research the incidental parameters problem is not usually an issue, the LD can be used without incurring in this violation – even though, as I showed earlier, the

[4]Two other 'philosophical' objections to fixed effects need to be addressed. First is the claim that fixed effects is an overly blunt solution to control for unobservable heterogeneity. The second is that it is not sufficiently grounded in theory. There is some truth to the first objection, but in a field where omitted variable bias is a salient problem, fixed effects models provide a better alternative to model overspecification and the inevitability of type-I errors. The second objection is not completely true. Rather, fixed effects exploits *within* variation, and I argue that this is often a better test of our theories, which usually revolve around rational choice explanations of domestic phenomena. Cross-national differences are exploited, but most theories use a logic that applies to changes within countries, not differences between them. I expand more on this in the case study of Houle (2016).

coefficients will be increasingly biased as events in y become less frequent. I will show that, provided a large number of observations per group (around 100), the ML and the LPMFE provide similar results when the number of observed events in the data (the proportion of ones) is between 25 and 75 percent. Below or above this threshold, the LPMFE approximates the observed probability with increasing accuracy. At rare events (1% of ones or less) or highly frequent events (99% of ones or more), the argument will be stronger, that is, that the LPMFE *ought to* be used as it yields accurate predictions. Lastly, while the performance of the LD improves with big data, which is counterintuitive, the LPMFE is still preferred with big data when events are rare.

5a. Logit versus the LPMFE: Monte Carlo Results

In this section, I introduce evidence from Monte Carlo simulations to ascertain the accuracy of the LPMFE and logit in calculating predicted probabilities using group fixed effects. The simulations test these two modeling techniques, LD and LPMFE, along three dimensions: (1) the distribution of the outcome variable, i.e. how frequent observed events are, (2) the number of groups or incidental parameters added to the model, and (3) the number of total observations and observations per group.

Simulation Design

Variables Y^* and Y are generated from a true model

$$Y^* = 1.5 + 1 \times X_i + \mu_i,$$

$$Y = 1 \text{ if } Y^* > q_j$$

$$Y = 0 \text{ if } Y^* < q_j,$$

		Sample Size								
		1050	**2000**	**2500**	**3000**	**3750**	**4500**	**5000**	**7500**	**100k**
Groups	**20**		***100***							
	50			*50*		*75*		*100*		
	75	*14*								
	100		***20***		***30***			*50*	*75*	
	150						*30*			
	1,000									***100***
	4,000									***25***

**Datasets in bold are the focus of the analysis.*

Table 1: Dataset Structures Used in Simulations

where X_i is a randomly generated variable with mean 0 and a standard deviation of 1. q_j is a vector of j quartiles, each of which produces exactly the same percentage of ones in the data as the percentile in separate simulations. The vector used establishes breakpoints at six different percentiles: the 1st, 3rd, 5th, 10th, 25th, and 50th. This creates exactly one, three, five, ten, twenty-five and fifty percent of events in each dataset, respectively. For the big data tests, I run the analysis only at true rare events (the 1st percentile, with only 1 percent of observed events in the data). Thus, a total of six sets of simulations are run for each dataset structure presented below in Table 1 – except for datasets with 100,000 observations, with one set of simulations at rare events. Each of these produces results at different levels of event occurrence. The dataset structures chosen in this paper are intended to (1) approximate those usually found in comparative politics research and (2) reflect what would be ideal datasets in terms of total observations and observations per group. I report all the dataset structures used in the simulations in Table 1. They vary between 1,050 and 100,000 total observations and between 20 and 4,000 groups, yielding balanced panels with varying numbers of observations per group. From a brief survey of the literature, I consider that most time-series cross-section datasets in comparative politics vary between around 1,000 and 7,500 observations. Some datasets may be larger than 7,500 observations, but these are rare. Similarly, datasets below 1,000 observations exist, but logistic models are known to be biased below this threshold –and this bias will only increase if fixed effects are used. I also include two datasets, both of which have 100,000 observations, which are intended to produce results for researchers who use big data with binary response models and who may be uncertain about using fixed effects specifications. One of these datasets has 1,000 groups and 100 observations per group, while the other has 4,000 groups and just 25 observations per group. Again, in the case of the LD, the results hold only when the asymptotics are in N, not in the group. Again, this is not an issue with the LPMFE.

For each of these panels, the first simulation performs the estimation with rare events data (1 percent of ones) and the last one with full variation in y (50 percent of ones). Thus, a total of 68 simulations are run —11 simulations for each of the 10 'common' dataset structures and one each for the big data structures. The one where the median is used provides the maximum amount of variation in the dependent variable and, therefore, is the one where the logit model should most closely approximate the observed probability. Comparing the coefficients and the predicted probabilities from these simulations sets the basis for the contrast between the LPMFE and the logit models in rare events data versus more common types of events.

Each simulation of 1,000 iterations produces, first, coefficients for the LD, CL and LPMFE.[5] With the LD and LPMFE estimates, I then calculate predicted probabilities using the observed value approach (Hanmer and Kalkan 2013) and store them for each simulation. Since applied researchers rely on predicted probabilities to assess substantive significance, these predicted probabilities will be the focus of the comparison –rather than the coefficients themselves. I then plot these predictions for the LD and the LPMFE at different levels of event occurrence in the DV. These comparisons will

[5]The CL coefficients were already shown in Section 3.

be the central source of evidence in this paper. To gauge which model provides a better estimate in each simulation, I calculate the observed probability of y at different levels of x. [6]

Results From Monte Carlo Experiments

In Figures 2, 4 and 5, I report a representative set of results from the simulations described above. These results are from simulations using three dataset structures: 2,000 observations and 20 groups, 3,000 observations and 100 groups, and 2,000 observations and 100 groups. These are common dataset structures in comparative politics and cover three different important scenarios. The first (2,000/20) has 100 observations per group and adds a smaller amount of incidental parameters, 20. The second data structure has a bigger overall sample of 3,000 observations and a borderline number of observations per group of 30. The last one (2,000/100) yields 20 observations per group, and any average group size below 30 is known to introduce bias in ML coefficients (Beck 2015). Therefore, the first model should have the least amount of bias in the estimation, since there are 100 observations per group.

Each of the graphs shows the results for the LD and LPMFE models as well as the observed or 'true' probability of y at each level of x. The x-axes in all graphs represent values for an x variable used as the main IV in the simulations, generated randomly with mean 0 and a standard deviation of 1. The y-axes are predicted probabilities. The blue line represents LPMFE predictions and the red line, LD probabilities. The black line plots the observed or 'true' probability of each event occurring at different levels of x, calculated as the average of y across values of x in the simulations.

Figure 2 reports the results for a dataset with 100 observations per group. The subplots (a)

[6] In Appendix A1, I provide the following robustness checks by modifying the data generating process in two important ways. First, I change the true value of the parameter of interest in the simulations. In the main results section of the paper, I use a true value of 1. For robustness, I use true values of 0.5 and 2 using two dataset structures, one known to be biased (2000 observations and 100 groups) and one unbiased (2000 observations and 20 groups) – see Figures A1 through A4. The results obtain if we change the true value of the parameter. Second, I add another variable into the regression to see if the results hold in multivariate analysis, which they do (Figures A5 and A6).

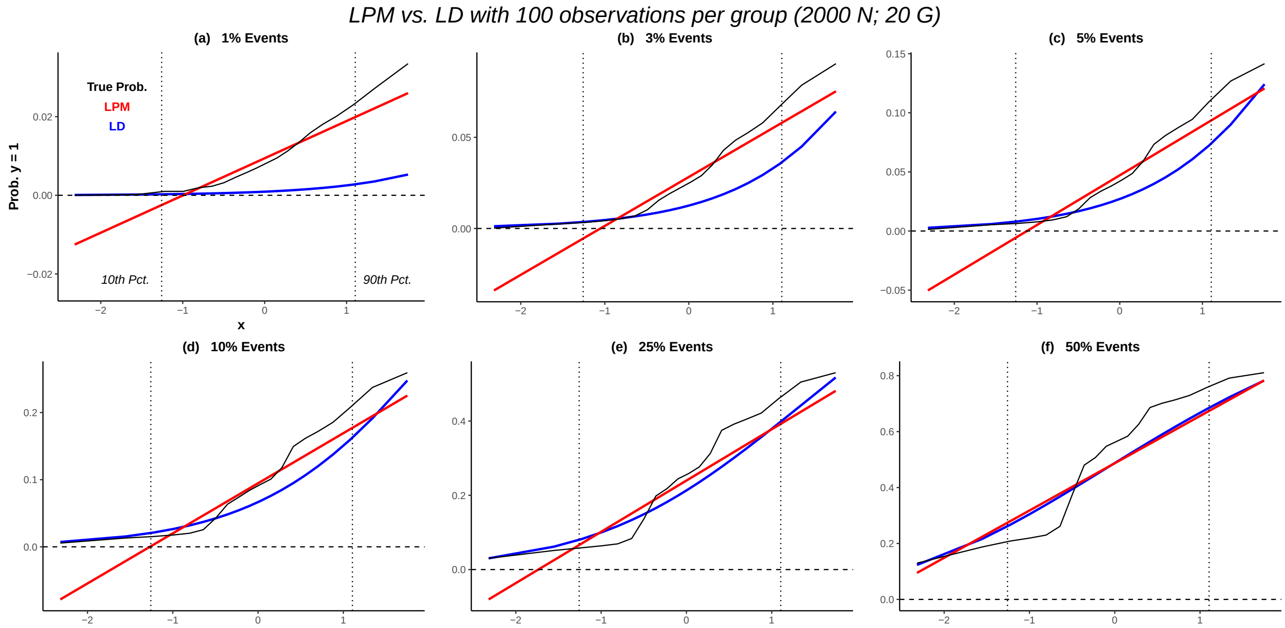


Figure 2: LPMFE, LD and observed probabilities at different levels of event occurrence in the DV.

through (f) display the predicted probabilities produced by the LPMFE and the LD by the level of occurrence of the event in y, ranging from rare events in (a), where only 1% of the observations are ones, to full variation in plot (f), where there are 50% of ones and 50% of zeros. Presented this way, the results clearly show how the predictions of the LPMFE and the LD evolve as we have more variation in y. The black line, or the 'true' probability, gives us a sense for which model produces more accurate estimates. It must be pointed out, however, that this line is not the result of any model but just the *average* of y in the simulations at different levels of x, and therefore it is not the 'correct' or exact answer that the right model should produce. What this 'observed' probability gives us is an approximation to where the model's prediction should lie, since no model ought to produce probabilities that are far off the actual level at which y is observed in the data as a function of x.

Three results are particularly important from Figure 2, and these will apply to Figures 4 and 5 as well. First, notice that the LPMFE and the ML models produce practically the same predicted probabilities when there is sufficient variation that no groups are dropped in the ML model – see plots (f). The two predictions may differ slightly at the extremes, but they map on to each other perfectly below the 90th percentile and above the 10th, a range more commonly used to interpret substantive significance. Indeed, this is consistent with what we know about the LPMFE and logit models without fixed effects, that is, that the LPMFE tends to underestimate or overestimate the extremes, but that probabilities in the center of the distribution are similar to those produced by logit models.

Second, between 25 and 75 percent of ones (I only show 25 percent), the ML model performs rather well. The fit of the line is closer to the observed probability, even though the LPMFE model is not far behind. Third, and similarly, what is *much* more striking is the loss of accuracy of the LD predicted probabilities when the number of ones in the data is below 25 percent. Subplots (a) through (d) show that the LPMFE produces probabilities that are much closer the observed probability curve, and virtually all LPMFE lines cross the observed probability curve at the mean. The LD, on the other hand, becomes progressively more inaccurate as events in y become rarer. Thus, these results confirm, on the one hand, that the ML model is highly inaccurate at rare events, which we could reasonably expect given its performance in rare events even without fixed effects. But, on the other hand, they take this common knowledge a step further: inaccuracies in LD predicted probabilities with fixed effects occur even when the data is not rare events per se, but even when y is slightly more common – 3, 5, or 10 percent of events. As plot (d) shows, at 10 percent, even with 100 observations per group, the LD model tends to slightly underestimate the observed probability across the distribution. The LPMFE, on the other hand, maps on nicely to the observed probability when the predictions of the model are over 0. Here, we know that, on average, $y = 1$ is observed around 10 percent on the time in the data. Since the relationship is designed to be positive and statistically significant, we also know that the predictions may be much higher at higher values of x, and be closer to 0 at lower values of x. This is precisely what the observed probability shows and what the LPMFE does a good job of reflecting.

A third important feature of these results is that, at true rare events (that is, around 1 percent of ones or less), the LPM is highly accurate in relation to the observed probability. Intuitively, this makes sense: if we imagine the CDF of the logistic distribution, there are three levels at which we can expect the LPM to practically map onto the ML probabilities *a priori*: at the two extremes and in the middle. This reasoning is illustrated in Figure 3. The dashed red line represents a linear prediction at rare events, high frequency events, and at the center of the distribution. These lines correspond exactly to their respective parts of the logistic CDF, with the exception being at extreme values of x. Therefore, the linearity assumption is only an important issue when events are not rare, highly common, or evenly split between ones and zeros. When they are, however, we should be confident that the LPM estimates the right predicted probabilities for the substantively meaningful range of x.

This is, in fact, consistent with the results in Figure 2. If we look at plots (a) through (f), the black line flattens in comparison to the other plots. The LPMFE models produce probabilities that

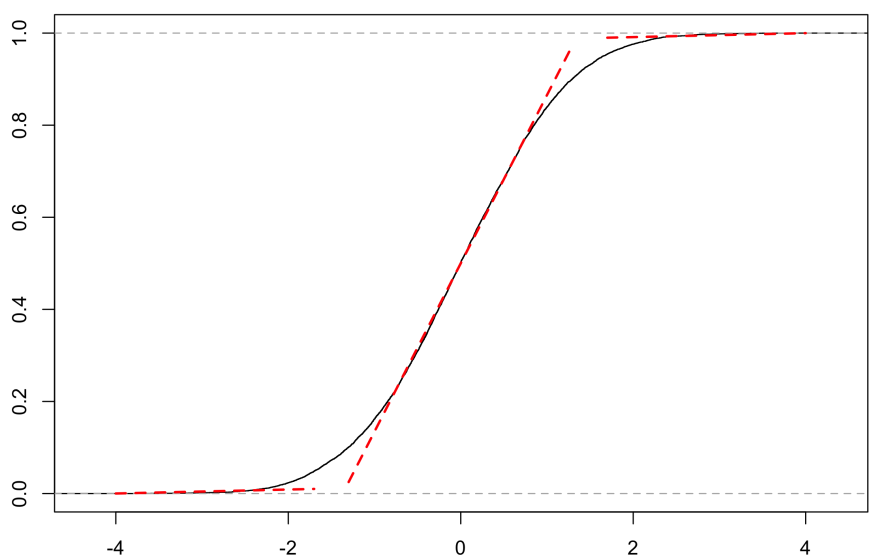


Figure 3: Logistic CDF with theoretical LPM probabilities overlaid

are much closer to the observed probability of y throughout the range of x. Since the range of predicted probabilities is much smaller in true rare events, say between near-zero probability and 3 or 4 percent, the ML assumption that the effect of x is greater in the middle of the distribution that at the extremes will not produce clear nonlinear patterns, as it may when the range of probabilities is wider. If we could ever fit an LD model and keep all the groups, the probabilities for the LD and the LPMFE would map onto each other the same way they do in plot (m) –at 50 percent of groups with 100 observations per group. Therefore, for fixed effects models at true rare events (below 1 percent) or highly uncommon events (below 5 percent), the LPMFE should be the model of choice. Its positive predicted probabilities will be meaningful and accurate. Lastly, it is important to note that, as expected, the LPMFE produces below-zero predictions in a majority of the plots, but those only occur when both the LD and the observed probabilities are essentially zero.

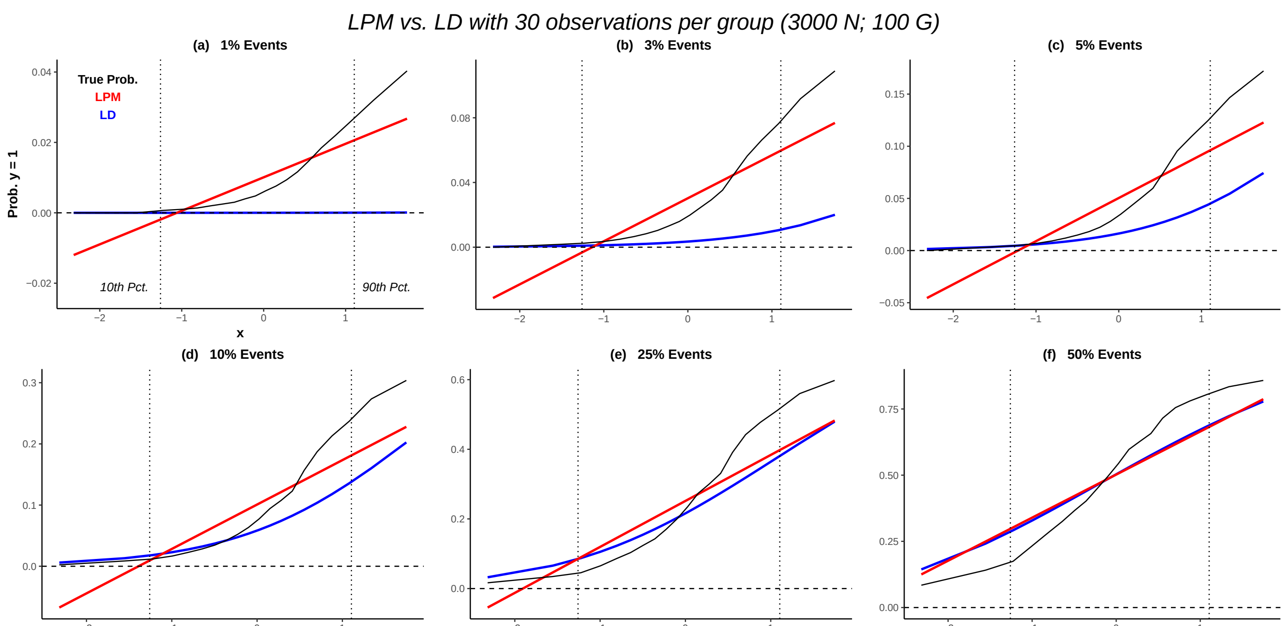


Figure 4: LPMFE, LD and observed probabilities at different levels of event occurrence in the DV.

Notice also in Figures 4 and 5 that the LPMFE models produce practically identical probabilities *regardless* of the number of observations or observations per group. Figure 4 reports the Monte Carlo results using 3000 observations and 100 groups, and Figure 5 for a dataset with 2000 observations and 100 groups. (The structure of the plots and subplots is the same as in Figure 2). Some bias may exist in plots between rare events and common events due to the linearity assumption, but the mean prediction produced is usually at the mean of the observed probability of y. On the other hand, the LD probabilities become increasingly inaccurate as events in y become less common. The robustness of OLS to these problems in model specification are well known, but these results seem to tilt the debate between LPMFE and ML in fixed effects toward the LPMFE. The results are practically identical in both Figures.

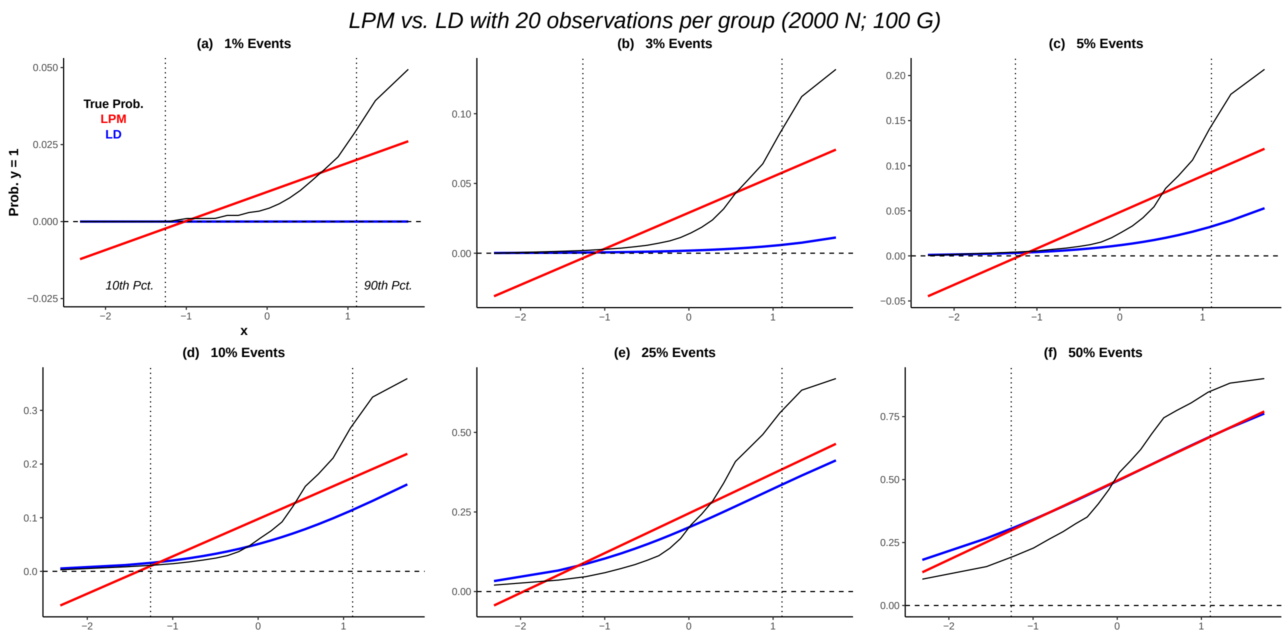
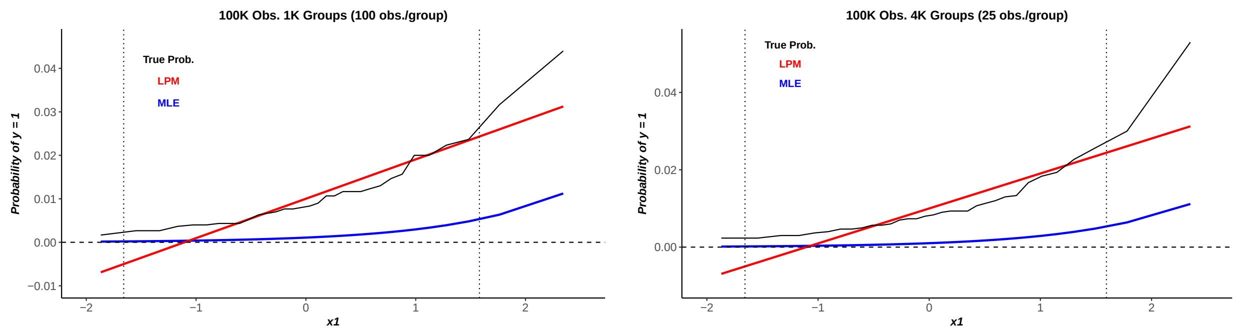


Figure 5: LPMFE, LD and observed probabilities at different levels of event occurrence in the DV.

Plots (f) in Figures 4 and 5 confirm the results from Figure 2 that, with over 30 observations, both models produce practically the same probabilities if the level of occurrence of y is around 50 percent. Even at 25 observations per group, however, the LD quickly becomes biased, as plots (e) show. At rare events, the LPMFE fits a model that maps on very closely to the observed probability (see plots (a)), again confirming what we saw in Figure 2. What is important to note is that these datasets are relatively common in comparative politics. It is rather *uncommon* to have much more than 30 observations per group, which is the threshold needed for the LD to produce generally



(a) (b)

Figure 6: LPMFE, LD and observed probabilities at different levels of event occurrence in the DV, with big data.

reliable probabilities. In these cases, the ML with fixed effects will be biased even if we manage to obtain data for a large number of countries, say 150. While the n of this dataset would be 3000 or more, the LPMFE should be preferred regardless of y's composition because the *average* number of observations per group is below 30.

Lastly, Figure 6 displays the predicted probabilities from both the LPMFE and the LD with big data with rare events – the data in both simulations includes only 1 percent of ones in the dependent variable. The big data results are consistent with the rest of the data structures (Figures 2, 4 and 5). Here, the most important result is that the LPMFE continues to provide accurate predicted probabilities within the relevant range of the distribution of x. It only produces wildly wrong probabilities when the observed probability is near-zero or above the 95th percentile of x. Again, this is expected of the LPMFE: it fits an accurate line through the center of the distribution but fails in the tails. However, this issue is substantively minor. On the one hand, negative probabilities can be effectively assumed to be zero, and probabilities at high values of x should not generally be used to assess substantively significant effects. The LD, on the other hand, does *better* in big data than in regular comparative dataset structures, in that it produces increasingly positive probabilities that, while still wrong, lie closer to the observed probability than they did in Figures 2, 4 and 5.

Three conclusions can be derived from this discussion. (1) The LD produces its most accurate predictions between 25 and 75 percent of ones in the dependent variable, and does best when variation in y is highest (50 percent of ones). The LD then becomes increasingly biased as variation in y decreases. (2) The LPMFE fits accurate predictions in true rare events or highly uncommon events data for the part of the distribution that is substantively relevant. (3) Results using big data with rare events remain substantively the same. I now proceed to place these findings into context by analyzing a recently published paper. I also refer back to the theoretical discussion to address some of the issues I raised in that section as to why fixed effects matters and why the LPMFE is often a good resource.

6. Fixed Effects In Comparative and IR Research

There are two reasons why fixed effects estimation is important in comparative political research, one empirical and one theoretical. The empirical reason is that comparativists often leverage variation among countries in their explanatory variables to explain variation in the dependent variable. But when within country variation is small and between country variation is large, the potential for omitted variable bias increases. One often sees research in which, once fixed effects are added to the equation, the standard errors remain similar but the coefficients drop and lose statistical significance. In a majority of cases, this is due to unobservables in the error term inflating our coefficients of interest (Allison, 2009). Admittedly, controlling for all possible confounders is difficult when researchers have to deal with a highly heterogeneous sample of cases —think of how different India, Sweden, Argentina and Burkina Faso are to each other. Omitted variable bias is particularly damaging because it does not decrease as sample size increases (Keele, 2015) and may lead to false positives.

The second reason for the importance of fixed effects is theoretical. Comparative research, as opposed to an important share of International Relations research, is concerned with how countries differ in their *domestic* politics. By comparing processes in different countries, we get better leverage to explain broad, generalizable political processes and phenomena. An influential theory, for instance, states that income inequality and democratization are negatively related because at high levels of inequality, elites find repression more cost-effective than redistribution (Boix 2003). This theory, as many others, is based on domestic mechanisms within different countries, while the empirics are pooled and explore differences *between* countries. Fixed effects, which explicitly uses within-country variation and then compares countries to each other, is a much better match for comparative theory than pooled models.

A parallel motivation is to raise awareness about which model is preferable and why. An example of this is Harding and Stasavage's (2014) excellent paper, published in *The Journal of Politics*, on the effects of democracy on basic social services. The authors correctly use a linear probability model with household-specific group intercepts, but fail to develop on the reasons for their choice. In a footnote, they explain that

> "There is some debate whether the most desirable way to estimate an equation with an endogenous dummy variable (as is the case with the election variable in Equation 3) is with a nonlinear model (such as probit and ivprobit) that is constrained to produce estimated probabilities between 0 and 1 or, alternatively, whether a linear probability model (estimated via either OLS or 2SLS) is preferable because it is not dependent on as restrictive a set of assumptions (see Angrist 2001). In practice, we obtained quite similar results using both approaches. As a result, in this section we will report the linear probability model estimates, which are more straightforward to interpret" (Harding and Stasavage 2014, 240).

Yet, if the asymptotics are in g, the group, a ML technique outside the conditional logit would incur in the incidental parameters problem (Chamberlain 1980). For some of their tests, given that the sample size is under 1,000 observations, such technique would not have been feasible. They may have attempted it with more confidence with their first test, which has a sample size of over half a million observations.

7. Conclusion

Panel data is a common dataset structure in all subfields of political science. Repeated observations across time (monthly, quarterly, or yearly data) are pooled for many different groups, such as individuals, districts, states, or countries. Two issues generally emerge with this type of data structure. First, unit-specific effects can lead to biased results if the model does not account for unobservable heterogeneity. Second, theories are often built on domestic processes that should leverage within-group variation, but researchers use between-group variation instead. Fixed effects addresses both of these issues, reducing type-I errors resulting from biased coefficients and using within-group variation in the model. In panel data, however, fixed effects estimation can be burdensome when the DV is dichotomous. The dichotomous nature of the DV calls for ML estimation, but a host of issues emerge when using the LD and the CL, as identified in this paper. Interpreting the substantive significance of CL coefficients is often challenging. The LPMFE is the best choice when dealing with rare events data – or even if events are just 'infrequent' – but the linearity assumption it imposes may not be always justified. Applied researchers are thus placed between a rock and a hard place when trying to use fixed effects models in their research.

In this paper, I have attempted to offer a way out of this dilemma. I have shown that the structure of the dataset is not sufficient to determine which fixed effects model to use. Most analysis of bias in fixed effects used the total number of observations and the number of observations per group to determine the usefulness of ML techniques over the LPMFE. This paper has demonstrated the importance of the *frequency* of events in y to determine model choice. ML increasingly overstates predicted probabilities as events become more rare. The LPMFE, on the other hand, is very accurate at both rare events and highly common events. The same results apply to big data.

References

Achen, Christopher H. 1982. *Interpreting and using regression.* Vol. 29. Sage Publications.

Allison, Paul D. 2009. *Fixed effects regression models.* Vol. 160. Sage Publications.

Andersen, Erling B. 1973. *Conditional inference and models for measuring.* Vol. 5 Mentalhygiejnisk forlag.

Angrist, Joshua D. 2001. "Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice." *Journal of business & economic statistics* 19(1):2–28.

Angrist, Joshua D and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

Beck, Nathaniel. 2015. Estimating grouped data models with a binary dependent variable and fixed effects: What are the issues? In *Annual meeting of the Society for Political Methodology.* pp. 22–25.
URL: *https://dataverse.harvard.edu/file.xhtml?fileId=2694346&version=RELEASED&version=.0*

Boix, Carles. 2003. *Democracy and redistribution.* Cambridge University Press.

Chamberlain, Gary. 1980. "Analysis of Covariance with Qualitative Data." *Review of Economic Studies* 47(1).

Chamberlain, Gary. 1984. "Panel data." *Handbook of econometrics* 2:1247–1318.

Greene, William H. 2003. *Econometric analysis.* Pearson Education.

Hanmer, Michael J and Kerem Ozan Kalkan. 2013. "Behind the curve: Clarifying the best approach to calculating predicted probabilities and marginal effects from limited dependent variable models." *American Journal of Political Science* 57(1):263–277.

Harding, Robin and David Stasavage. 2013. "What democracy does (and doesn't do) for basic services: School fees, school inputs, and African elections." *The Journal of Politics* 76(1):229–245.

Heckman, James J. 1981. *The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process.* MIT Press.

Hosmer Jr, David W, Stanley Lemeshow and Rodney X Sturdivant. 2013. *Applied logistic regression.* Vol. 398 John Wiley & Sons.

Houle, Christian. 2016. "Why class inequality breeds coups but not civil wars." *Journal of Peace Research* 53.

Huber, Peter J. 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1 pp. 221–233.

Katz, Ethan. 2001. "Bias in conditional and unconditional fixed effects logit estimation." *Political Analysis* pp. 379–384.

Keele, Luke. 2015. "The statistics of causal inference: A view from political methodology." *Political Analysis* 23(3):313–335.

Keele, Luke and Nathan J Kelly. 2006. "Dynamic models for dynamic theories: The ins and outs of lagged dependent variables." *Political analysis* 14(2):186–205.

King, Gary and Langche Zeng. 2001. "Logistic regression in rare events data." *Political analysis* pp. 137–163.

Miller, Michael K. 2012. "Economic development, violent leader removal, and democratization." *American Journal of Political Science* 56(4):1002–1020.

Neyman, Jerzy and Elizabeth L Scott. 1948. "Consistent estimates based on partially consistent observations." *Econometrica: Journal of the Econometric Society* pp. 1–32.

White, Halbert. 1980. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica: Journal of the Econometric Society* pp. 817–838.

Wooldridge, Jeffrey M. 2002. *Econometric analysis of cross section and panel data*. The MIT Press.