PA

Estimating Grouped Data Models with a Binary-Dependent Variable and Fixed Effects via a Logit versus a Linear Probability Model: The Impact of Dropped Units

Nathaniel Beck

Department of Politics, New York University, New York, NY 10003, USA. Email: nathaniel.beck@nyu.edu

Abstract

This letter deals with a very simple question: if we have grouped data with a binary-dependent variable and want to include fixed effects in the specification, can we meaningfully compare results using a linear model to those estimated with a logit? The reason to doubt such a comparison is that the linear specification *appears* to keep all observations, whereas the logit drops the groups where the dependent variable is either all zeros or all ones. This letter demonstrates that a linear specification averages the estimates for all the homogeneous outcome groups (which, by definition, all have slope coefficients of zero) with the slope coefficients for the groups with a mix of zeros and ones. The correct comparison of the linear to logit form is to only look at groups with some variation in the dependent variable. Researchers using the linear specification are urged to report results for all groups *and* for the subset of groups where the dependent variable varies. The interpretation of the difference between these two results depends upon assumptions which cannot be empirically assessed.

Keywords: binary logit, clustered data, marginal effects, fixed effects, panel data

1 Introduction

Many applied researchers include "fixed effects" (unit specific intercepts) to account for unmodeled heterogeneity in grouped data analyses. The inclusion of fixed effects, however, can lead to issues interpreting the results of the estimation. Researchers often use a linear probability model with unit specific intercepts ("LpmFE") which is sometimes compared to a logit model with the same unit specific intercepts ("LogitFE"). One reason researchers might choose the linear specification is that it *appears* to use all the data in estimation, whereas the logit specification drops all groups that show no variation in the dependent variable ("homogeneous groups"). Thus, the logit form is estimated on a subset of the data used by the linear form; this subset may be dramatically smaller if, for example, outcomes are rare events.[1]

This letter demonstrates the consequences of subsetting the data to eliminate homogeneous groups—a decision that cannot be assessed empirically. This letter encourages researchers to think carefully about the substantive implications of restricting the data to heterogeneous groups and calls for a common standard of presenting both results in robustness checks. Furthermore,

Political Analysis (2020)
vol. 28:139–145
DOI: 10.1017/pan.2019.20

Published
19 September 2019

Corresponding author
Nathaniel Beck

Edited by
Jeff Gill

Author's note: Replication data may be found in Beck (2018b). An earlier and longer version of this letter was given at the 2015 Annual Meeting of the Society for Political Methodology, University of Rochester, Rochester, NY, July; this letter benefits from helpful comments at that session. Special thanks to Chris Blattman who initially posed this question to me. The three referees were extremely helpful in seeing the correct structure of this letter and made a huge difference in the final copy. The actual letter was written while I enjoyed the hospitality of the United States Studies Centre, University of Sydney, Sydney, Australia.

1 This is a common problem in international relations research where groups are countries observed over years, and many countries experience no rare events such as wars or coups.

researchers comparing the logit and linear estimates should restrict the latter to the observations from the heterogeneous group data; comparing the logit results on the heterogeneous subset to the linear results on the full data set is comparing results on what might be two very different data sets.

Researchers are aware of this issue, but they often seem to take an *ad hoc* approach to the choice of specification and data set limitation. This can be seen in two recent examples in very prominent journals. Wright, Frantz, and Geddes (2013) examines the link between oil wealth and autocratic regime survival using a country–year design with a binary-dependent variable denoting whether a regime survived a given year. The article notes that fixed effects are generally included in models similar to theirs in order to deal with the unobserved unit heterogeneity. Wright, Frantz, and Geddes (2013, 294) go on to say:

> "[T]his strategy, however, drops [between 26 and 64] countries from the analysis that do not experience [various types of] regime change. . . . Dropping countries that do not experience regime change may bias estimates downward by selecting only those where regime change has occurred in the sample period, particularly if those stable political systems have high oil wealth. Below, we investigate the possibility that this restriction on the sample induces selection bias [by dropping fixed effects from the logit model]."

If fixed effects are needed to aid causal identification, omitting them to change the countries studied is not an obvious solution. Are changes in the results from dropping fixed effects due to the different sample or due to the failure to control for unobserved unit heterogeneity?

In the second example, Besley and Reynal-Querol (2011) test the probability of a leader having a graduate degree across regimes to study whether democracies elect more educated leaders. As in the previous example, panel data are used with countries observed over many years and country fixed effects are used to account for unmodeled heterogeneity. The article estimates the LpmFE specification (without justification) and then estimates a conditional logit model as a robustness check. The authors' only mention of this issue is ". . . we estimate a conditional logit model to recognize the discrete nature of the left-hand side variable. The core finding of [the LpmFE model] remains" (Besley and Reynal-Querol 2011, p. 556). This is clearly correct if we only care about the sign and significance of a coefficient; but, as we shall see, the difference between the two estimations is not trivial, albeit *perhaps* not enormous.

1.1 Notation

To make the issues concrete, we need a bit of notation and nomenclature. Let y_{gi} be a binary-dependent variable with the exogenous covariates being $\mathbf{x}_{gi}$, where g indexes groups and i indexes particular units in a group. It simplifies notation to assume that all groups are of the same size, and dropping this one extra subscript has no consequence for the argument. Let this group size be N, with G being the number of groups. α_g refers to the fixed effect for group g, that is the group specific (fixed) intercept.

For the purposes of this article, I take N to be large enough (approximately over 20) so that the simple logit form provides essentially unbiased and efficient estimators. G can take any value, but in typical applications, it is reasonably large.

The LogitFE model is

$$P(y_{gi} = 1) = \frac{1}{1 + e^{-(\mathbf{x}_{gi}\boldsymbol{\beta} + \alpha_g)}} \tag{1}$$

and $\hat{P}(y_{gi} = 1)$ is estimated via plugging in parameter estimates. In this letter, this model is estimated via a standard logit analysis with unit intercepts (dummy variables) adjoined.[2] The LpmFE model is

$$y_{gi} = \mathbf{x}_{gi}\boldsymbol{\beta} + \alpha_g + \epsilon_{gi}. \tag{2}$$

This specification is easily estimated by ordinary least squares ("OLS") with $\hat{P}(y_{gi} = 1)$ estimated by $\hat{y}_{gi}$ in the obvious way. For both the linear and logit forms, we can then estimate $\frac{\partial P(y_{gi}=1)}{\partial \mathbf{x}_{gi}}$, the marginal effect of the covariates on $P(y_{gi} = 1)$.[3]

There may be some groups where every member of the group has $y = 0$ ("AllZero" groups).[4] This letter deals with the issue of the consequences of such groups, and it is shown that the consequences, not surprisingly, increase as the number of AllZero groups increases. Section 2 illustrates the consequence of including the AllZero groups in the LpmFE estimation and the difference between LpmFE and LogitFE as a function of dropping those groups. Section 3 reanalyzes one result from Besley and Reynal-Querol (2011) to show the practical importance of the analytic results. The conclusion discusses some interpretative issues in these differences and suggests that researchers using LpmFE report results on the entire data set as well as the subset of the data which drops homogeneous groups. Those who compare LpmFE and LogitFE should do this comparison on the same data, that is, the subset of the data which drops the homogeneous groups.

2 Differences between what is estimated with LpmFE and LogitFE

As noted, the LogitFE specification can be estimated via any standard logit program, with the group specific dummy variables adjoined to the list of covariates.[5] This is because, for the homogeneous groups of all failures (AllZero), the likelihood is maximized when all the estimated probabilities of success are as close to zero as possible. To achieve this, the maximum likelihood program estimates α_g for these groups as being as close to $-\infty$ as possible, subject only to the numerical precision of a computer. In this case, the covariates for these groups have no marginal effect on the probability of success, and so the likelihood is unaffected by the values of the covariates for these groups. (The same argument holds for the homogeneous groups of

2 Some readers may be surprised to see no mention of Chamberlain's (1980) conditional logit which removes the bias from the LogitFE specification when N is small. There are several reasons for this. First, most applications of LogitFE and LpmFE in political science are in situations where N is reasonably large (20 or more). A search of the most recent five years of political science/international relations articles in JSTOR found none using conditional logit for the very small N size case, and only about two or three per year use it in the larger N case. Second, as N grows, conditional logit and LogitFE converge. In practical terms, the results are almost identical for $N > 20$. The bias of LogitFE is very small in these cases (Katz 2001; Greene 2004; Coupé 2005) and the efficiency loss of LogitFE as compared to conditional logit is tiny (Beck 2018a). Third, conditional logit does not allow for computing marginal effects since it conditions out, rather than estimates, the group specific intercepts which are needed to estimate $P(y_{gi} = 1)$. Thus, in a world where showing marginal effects is standard, conditional logit cannot produce the usual quantities of interest. Finally, since conditional logit drops all the homogeneous groups, researchers should be aware that it changes the data set used for estimation in the same way that LogitFE does. Thus, all the issues discussed in this letter apply to conditional logit. For reasonably large group sizes, the LogitFE specification should be preferred to conditional logit since it provides similar estimates of the $\boldsymbol{\beta}$ while allowing for the computation of marginal effects.

3 For both the logit and linear specification, this simplifies. For the linear model, this is just $\hat{\boldsymbol{\beta}}$ for each observation and does not vary by observation. For the logit model, the average marginal effect differs by observation and is $\hat{\boldsymbol{\beta}}\hat{P}(y_{gi} = 1)\hat{P}(y_{gi} = 0)$, which is then averaged over the observed data to yield the average marginal effect.

4 Following convention, I refer to $y_{gi} = 0$ as a failure and the opposite as a success. Homogeneous groups may also show all successes for the group. As we see in the next section, this yields identical results, and so for simplicity and without loss of generality, this letter focuses on AllZero groups.

5 Maximum likelihood programs, such as Stata's **logit**, just drop the homogeneous groups. The workhorse logit program in R, **glm**, does not drop the groups since it uses iteratively reweighted least squares. While it estimates the $\boldsymbol{\beta}$ correctly, the estimates of the α_g are incorrect and yield the entire unit interval as the confidence interval for the predicted probability of success, leading to incorrect confidence intervals for marginal effects (of essentially the entire real line). This can be fixed by subsetting the data so that AllZero groups are dropped. The R package **bife**, which is for binary grouped data, automatically drops the AllZero groups and so is correct. This is documented in the replication results (Beck 2018b).

only successes, except that the group specific intercepts are estimated to be as close to ∞ as numerical precision allows.) I return in the conclusion as to how to think about marginal effects in the AllZero groups.

Interpreting the impact of the AllZero groups on the OLS estimation of Equation (2) is not quite so obvious. Here, it is simplest to work with group mean centered data to avoid putting the group intercepts in the specification. Let $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ be the group mean centered data and let $\tilde{\mathbf{X}}_0$ be the centered covariate matrix for the AllZero groups, with $\tilde{\mathbf{X}}_1$ being the corresponding matrix for the NotAllZero groups and $\tilde{\mathbf{y}}_1$ being the group mean centered vector of observations on y for the NotAllZero groups (and obviously the corresponding vector for the AllZero groups is $\tilde{\mathbf{y}}_0 = \mathbf{0}$).[6] Unsubscripted matrices and vectors refer to the complete data set.

Thus, the OLS estimate of $\boldsymbol{\beta}$ for the entire data set is given by

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1 + \tilde{\mathbf{X}}_0{}'\tilde{\mathbf{X}}_0)^{-1}(\tilde{\mathbf{X}}_1'\tilde{\mathbf{y}}_1), \tag{3}$$

whereas the corresponding estimate for the NotAllZero groups is given by

$$\hat{\boldsymbol{\beta}}_1 = (\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1)^{-1}(\tilde{\mathbf{X}}_1'\tilde{\mathbf{y}}_1). \tag{4}$$

We can also compare the variance covariance matrix of the two estimates. For the entire data set, this matrix is

$$(\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1 + \tilde{\mathbf{X}}_0'\tilde{\mathbf{X}}_0)^{-1}\widehat{\sigma^2}, \tag{5}$$

whereas the corresponding estimate for the NotAllZero groups is given by

$$(\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1)^{-1}\widehat{\sigma_1^2}, \tag{6}$$

where $\widehat{\sigma^2}$ and $\widehat{\sigma_1^2}$ refer to estimates of (the square of) the standard error of the regression in the full and restricted data sets, respectively.

It is immediately obvious that the two equations only differ by the $\tilde{\mathbf{X}}_0'\tilde{\mathbf{X}}_0$ portion of the $\mathbf{X}'\mathbf{X}$ matrix that is being inverted. Alternatively, it is also obvious that the OLS estimates using all the data are a weighted average of $\mathbf{0}$ and $\hat{\boldsymbol{\beta}}_1$—effectively, $\hat{\boldsymbol{\beta}}$ shrinks $\hat{\boldsymbol{\beta}}_1$ toward $\mathbf{0}$. The amount of shrinkage is a somewhat complicated function that depends on the relative scale of $\tilde{\mathbf{X}}_0'\tilde{\mathbf{X}}_0$ and $\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1$. As the proportion of AllZero groups goes up, $\hat{\boldsymbol{\beta}}$ goes to $\mathbf{0}$, but the path may not always be monotonic for all components of $\hat{\boldsymbol{\beta}}$.

The variance covariance matrix of the estimates has two components which move in different directions as we move from the entire data set to the NotAllZero subset. The estimated σ^2 will get smaller since we are eliminating the AllZero groups; but the $\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1$ matrix in the NotAllZero data will also be smaller in scale than the corresponding $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ matrix used to estimate the variance covariance matrix of $\hat{\boldsymbol{\beta}}$. Note, however, that the estimated standard error of the regression will be limited in how much it changes since the variance of $\tilde{\mathbf{y}}$ is constrained by its nature as a binary variable. Meanwhile, the $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ matrix is not similarly limited by any scaling and so could shrink considerably as the AllZero cases are dropped. Usually, the estimated standard errors of $\hat{\boldsymbol{\beta}}_1$ will be smaller than the corresponding estimates for $\hat{\boldsymbol{\beta}}$. The change in $\hat{\boldsymbol{\beta}}$ and the change in its estimated standard error offset, and so we usually see that the change of the t-ratio due to dropping the AllZero groups is smaller than the corresponding change of the estimate of $\boldsymbol{\beta}$. This smaller

6 Thus, the estimated $\boldsymbol{\beta}$ for the AllZero groups only is $\hat{\boldsymbol{\beta}}_0 = (\tilde{\mathbf{X}}_0'\tilde{\mathbf{X}}_0)^{-1}(\tilde{\mathbf{X}}'\mathbf{0}) = \mathbf{0}$ (assuming $\tilde{\mathbf{X}}_0 \neq \mathbf{0}$). Note that if we have a group of all successes, its centered y is also zero which is why we do not have to separately analyze the effect of homogeneous groups containing only successes.

change in the t-ratio may be one reason that applied researchers are content to conclude that the substantive results from LogitFE are similar to those of LpmFE. But we should go beyond simply inquiring as to the sign of a coefficient and whether its "significance" is beyond some standard threshold to actually looking at coefficients.[7]

It is very simple to see what is going on by looking at the scalar x case, where once again $\tilde{y}$ and $\tilde{x}$ have been group mean centered. The OLS estimate of β for the entire data set is given by

$$\hat{\beta} = \frac{\sum_{\text{NotAllZero}} \tilde{x}_{gi}\tilde{y}_{gi}}{\sum_{\text{AllData}} \tilde{x}_{gi}^2}, \tag{7}$$

whereas the corresponding estimate for the NotAllZero groups is given by

$$\hat{\beta}_1 = \frac{\sum_{\text{NotAllZero}} \tilde{x}_{gi}\tilde{y}_{gi}}{\sum_{\text{NotAllZero}} \tilde{x}_{gi}^2}. \tag{8}$$

These two equations differ only by an extra $\sum_{\text{AllZero}} \tilde{x}_{gi}^2$ in the denominator of Equation (7) which is nonnegative; hence, $|\hat{\beta}| < |\hat{\beta}_1|$. The standard error for $\hat{\beta}$ for the entire data set is given by

$$\sqrt{\frac{\widehat{\sigma^2}}{\sum_{\text{AllData}} \tilde{x}_{gi}^2}}, \tag{9}$$

whereas the corresponding standard error for the NotAllZero groups ($\hat{\beta}_1$) is given by

$$\sqrt{\frac{\widehat{\sigma_1^2}}{\sum_{\text{NotAllZero}} \tilde{x}_{gi}^2}} \tag{10}$$

where again the extra summation terms in the denominator must be positive.

For the scalar case, it is obvious that including the AllZero groups shrinks $\hat{\beta}_1$ toward zero (in absolute value), where the amount of shrinkage depends on the number of AllZero groups and on the variation of the centered x's in those groups. The estimated standard error of β_1 also gets smaller (in general) since the larger denominator due to $\sum_{\text{AllZero}} \tilde{x}_{gi}^2$ will almost always offset the increase in the estimate of the standard error of the regression due to the greater heterogeneity of y of the full data set. This again leads to offsetting effects in the t-ratio.

3 Examples

If readers need to be convinced of the mathematics, one example should do. Here, I reanalyze the Besley and Reynal-Querol (2011) results cited previously since the article is important and the replication data were provided by the authors. It is easy to compare the LpmFE and LogitFE results of the two estimates for the effect of democracy on whether a leader has a graduate degree. These results are presented in the authors' Table 1, with Column 1 being the LpmFE model and Column 3 being the LogitFE model. The regression results are based on 1146 country–year observations, 190 of which never had a leader with a graduate degree and are therefore dropped in the logit specification. Table 1 in this letter summarizes results of a slightly simpler specification, replacing the original conditional logit with LogitFE, so that this letter can focus on the issues of dropped cases. When comparable, the results here are similar to those of the original article.

7 Of course, it should not be surprising if marginally statistically significant LogitFE estimates become statistically insignificant using the LpmFE specification. This tells us nothing about the impact of the covariates on the probability of success in the heterogeneous groups!

Table 1. Reanalysis of Besley and Reynal-Querol (2011) using LpmFE and LogitFE.

	LpmFE/All		LpmFE/NotAllZero		LogitFE		LpmFE/AllZero
	$\hat{\beta}$	SE	$\hat{\beta}$	SE	$\hat{\beta}$	SE	$\hat{\beta}$
Democracy	0.260	0.043	0.294	0.046	1.748	0.269	0.000
AME (Democracy)					0.298	0.041	
Number of Observations	1146		956		956		190

Note: LpmFE and LogitFE estimates of effect of a democracy dummy variable on probability a leader has a graduate degree in a given country and year. AME is the average marginal effect estimated with the LpmFE. Mixed drops AllZero groups; LpmFE is only on AllZero groups. Data as in Besley and Reynal-Querol (2011); the data set used here is the relevant subset of this data (used to estimate their Table 1) which has 145 distinct countries possibly observed from 1872–2004, though few countries have complete data over that period. The log of GDP per capita is also included in the specification, as are country (but not year) dummy variables. Full regression results and replication data and Stata code are available at Beck (2018b).

With all data, the effect of the democracy dummy on the linear probability of a leader having a graduate degree is 26% (with a standard error of 4.3%). Restricting the data set to countries with at least one leader having a graduate degree increases this coefficient to 29.4% (with a small increase in the standard error). This is a 14% increase in the estimated coefficient when 17% of the data are dropped. The LogitFE, which automatically drops the AllZero groups, shows the average marginal effect of a country being a democracy on having a leader with a graduate degree of 29.78%, almost the same as the corresponding LpmFE estimated marginal effect *dropping the AllZero groups*. For those who doubt the algebra of the previous section, I also report the regression results including only the AllZero groups. The estimated coefficient is, of course, zero (to 17 decimal places).

4 Conclusion

The takeaway from this article is fairly simple. Researchers often require fixed effect specifications to account for unmodeled heterogeneity. Such researchers often either choose LogitFE or LpmFE without justification or present the results of both as a "robustness" check. What such researchers must remember is that LpmFE and LogitFE are estimated on different data sets, with the latter being a subset of the former obtained by dropping all the AllZero groups. LpmFE, in keeping all the groups, estimates the average marginal effect of a covariate as a linear combination of zero and the estimated coefficients using only the NotAllZero group data. Depending on how many groups are AllZero (which is observable), this effect may be large.

The correct comparisons to the average marginal effects generated by the LogitFE specification are the LpmFE estimates calculated using only the NotAllZero group data set. Researchers reporting only LpmFE results should report estimates both keeping and dropping the AllZero groups, always remembering that the former is a linear combination of the latter and zero. Of course, these two estimates may differ only slightly if there are few AllZero groups, but both estimates should still be reported.

Which of the two estimates is "correct?" This question cannot be settled empirically. It can be argued that since the covariates in the AllZero groups do not lead to any successes in that group, the marginal effect of those covariates in such groups is zero. Alternatively, it may be the case that these covariates in such groups could have an effect, but there is some unmeasured factor (estimated as the group specific intercept) that so depresses the probability of a success in that group such that no measured covariate can offset this factor. But either way, the existence of AllZero groups, with known zero marginal effects, violates the assumption of the LpmFE specification of constant marginal effects. Reporting separate estimates for all the data and for only the NotAllZero groups goes some way to remedying this. It is then easy to assess how

important is the move from the linear to the logit specification by comparing the estimates of the average marginal effects from the LpmFE and LogitFE specification on the same data.

It is up to researchers (and readers) to think about whether using all the groups or only the NotAllZero groups is closer to what is the right marginal effect to estimate. Since this cannot be dealt with empirically, the best that can be recommended here is to report both estimates and to never compare average marginal effects estimated using different data sets. In grouped data with fixed effects, the marginal effect of covariates may differ nontrivially between AllZero and NotAllZero groups; this must be part of a summary of the data.

References

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

Beck, N. 2018b. "Replication Data for: Estimating Grouped Data Models with a Binary Dependent Variable and Fixed Effect via Logit vs Ols: The Impact of Dropped Units." https://doi.org/10.7910/DVN/SVAONZ, Harvard Dataverse, V1, UNF:6:kLjEynIuw96KpHaOHc6tJA== [fileUNF].

Besley, T., and M. Reynal-Querol. 2011. "Do Democracies Select More Educated Leaders?" *The American Political Science Review* 105(3):552–566.

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

Coupé, T. 2005. "Bias in Conditional and Unconditional Fixed Effects Logit Estimation: A Correction." *Political Analysis* 13(3):292–295.

Greene, W. 2004. "The Behaviour of the Maximum Likelihood Estimator of Limited Dependent Variable Models in the Presence of Fixed Effects." *Econometrics Journal* 7(1):98–119.

Katz, E. 2001. "Bias in Conditional and Unconditional Fixed Effects Logit Estimation." *Political Analysis* 9(4):379–384.

Wright, J., E. Frantz, and B. Geddes. 2013. "Oil and Autocratic Regime Survival." *British Journal of Political Science* 45:287–306.