Is OLS with a binary dependent variable really OK?: Estimating (mostly) TSCS models with binary dependent variables and fixed effects[*]

Nathaniel Beck[†]

Draft of August 2, 2011

ABSTRACT

The paper uses Monte Carlo simulation to examine the small sample properties of linear regression with a binary dependent variable, and particularly time-series–cross-section data with both a binary dependent variable and fixed effects. It begins by examining the "folk' theorem" that linear regression is about as good as logit for estimating average marginal effects. This appears to be true for normally distributed data but not so true where the covariates show skewness or excess kurtosis; with more unusually generated data logit does outperform regression. Logit also appears to outperform regression for estimating binary (exogenous) treatment effects. The paper shows that the advantage of simple logit with group specific intercepts over Chamberlain's conditional logit is largely illusory. The main part of the paper shows that the whether one wants to drop groups with no successes before estimation (as required by any of the logit techniques) or to keep them (as regression allows) is largely a function of some non-empirical assumptions about the data generating process. Thus it may be true that regression with a binary dependent variable may not be "too bad," analysts must think carefully before employing the folk theorem.

[*] Chris Blattman and Shanker Satyanath started me down this path by telling me that development economists still liked the linear probability model for various reasons. Ethan Katz and Tom Coupé wrote the original Stata code which I shamelessly borrowed; errors in the code were most likely introduced by me. The codes have by now sufficiently diverged that anyone interested in replicating or extending my results should use my code; given the preliminary nature of this draft, anyone interested in seeing my code should just contact me. The coding would have been even more ungraceful without the kind help of Kristin MacDonald at Stata Corp., though she might not want her name associated with my code.

[†] Department of Politics; New York University; New York, NY 10003 USA; nathaniel.beck@nyu.edu

1. INTRODUCTION

Different subfields or disciplines appear to have different methodological conventions. In political science a binary dependent variable almost always leads researchers to use a limited dependent variable technique; applied economists seem more likely to still use regression appealing to a linear probability model ("LPM").[1] Applied economists working with time-series–cross-section (TSCS) data seem to believe that including fixed effects is required to rule out unmodeled heterogeneity.[2] This paper examines these choices via Monte Carlo analyses of the estimators.

This paper deals primarily with TSCS data. While one might wish for a unified treatment of all longitudinal data models, the issue of fixed effects is inextricably linked with the number of time periods under study. My own preference is to call longitudinal studies based on a small number of waves of respondents "panel" data and studies based on observing political units over a relatively long (roughly at least a decade) time period "TSCS" data. Obviously one can have studies on political units observed only twice, and one can have a large number of waves in a panel study. But such studies are rare.

I will assume a rectangular data set, that is, N units observed over T time periods. Given the nature of this paper, complicated issues that occur in real research, such as missing data or measurement error, will simply be ignored. If N and T are both small, we just have very little data. If T is small then models with fixed effects are subject to the famous "incidental parameters" problem first discussed by Neyman and Scott (1948). They showed that maximum likelihood estimators would be inconsistent if the number of incidental parameters gets large at the same rate as the number of observations. For present purposes, this means that maximum likelihood estimation of panel data with fixed effects will yield inconsistent estimates, since as N gets larger, the number of effects to estimate gets larger at exactly the same rate (assuming the number of panel waves is fixed).[3] In TSCS data we can think of both N and T growing large, so the problem is somewhat different. The results of this paper apply to studies where we can think of T as being large; N might or

[1] I will call these two approaches logit and regression and use LPM and regression interchangeably. There are, of course, alternatives to standard regression and logit. I do not think it matters much whether one uses logit or probit, and for the fixed effects issues that I am concerned with here, the standard treatment is conditional logit; it is harder to combine probit with fixed effects. On the regression side, one could do weighted regression or constrained regression. But since the principal appeal of the LPM is simplicity, both of estimation and interpretation, I deal only with simple linear regression (although I have results for weighted regression which indicate the weighting does not seem to improve the regression properties (error loss) that are under study in this paper.

[2] At this point I have not done a thorough enough reading of the applied economics literature, or the work in political economy done by applied economists, to be confident about any characterization of the methodology of that subfield or that set of researchers. My reading of the relevant recent articles in relevant journals, and my discussions with a highly non-random group of such researchers, indicates to me that my belief is at least not crazy. But without more systematic study of the literature there is no reason for anyone to be persuaded by my characterization. Fortunately nothing in this paper, other than the choice of topic, depends upon this characterization of a literature being correct. Thus I make no references to any specific applied articles or results.

[3] The problem is easy to avoid for linear models since it is easy to condition on the incidental parameters; this approach is not available for non-linear models.

might not matter, but in practice N is seldom a single digit number. None of the simulations presented are designed to study issues relevant to those interested in "panel" data.

Since this paper is concerned with the finite sample properties of estimators it relies completely on Monte Carlo methods. The simulations were designed to mimic the properties of TSCS data, as least insofar as N and T. As with any Monte Carlo study we can only be confident about results conditional on the Monte Carlo setup. Since the various simulations have many moving parts, I have simplified by limiting the analysis to models with a single independent variable of interest. In the applied economics literature interest almost always focuses on marginal effects; therefore I only compare estimators in terms of those marginal effects. All estimators (except as noted) are compared in terms of the absolute difference of the estimated and true marginal effect as a percentage of the true marginal effect; all reported results are the averages of these percent differences over the runs of the simulations. Obviously when the true marginal effect is small, a small absolute error will be a larger percentage error. At this point there is no assessment of standard errors or coverage rates.

The true DGP used in the simulations is a logit with a single independent variable, x, (and usually fixed effects). The simulations are similar to those used by Katz (2001) as corrected by Coupé (2005).[4] That code drew one set of x and the fixed effects which were held constant over all the runs of the simulation; because the draws of the small number of fixed effects can have huge consequences, I changed the design to redraw both new x's and new fixed effects for each iteration of the simulation. All results are based on 1000 runs of each simulation experiment. The results shown are the average percentage deviations from truth for each set of simulation parameters. These results are therefore not a sample from anything, but the 1000 individual results that make up each result are a sample given the DGP. More details on the various DGPs are given in the discussion of the results.[5]

It may seem odd not to discuss bias or other measures.[6] Unbiasedness is not a very interesting property, except insofar as increased bias increases the absolute error of an estimator. Since in simulations we know the truth, it makes sense to simply compare that truth with the estimation of said truth. I report percentage deviations because the truth varies over the course of the experiments. I chose to concentrate on mean absolute percentage deviation instead of the more typical root mean squared deviation because I see no reason to emphasize a few rather bad results. Results are clear enough the a mean squared approach would not have changed any conclusions.

The binary dependent variable, y is generated by a standard inverse logit function. This has important consequences for comparing regression and logit, since the DGP is the one assumed by the logit estimator. Obviously regression would do better if the DGP were the LPM, but the regression/logit comparisons were to deal with the question of the tradeoff of the simplicity of interpretation of regression as compared to the more reasonable but complex specification of logit. Obviously regression would do quite well if the LPM were the DGP

[4] Coupé corrected Katz's code to allow for x and the fixed effects to be correlated and to make sure that the fixed effects differed over units.

[5] All simulations were programmed in Stata/MP4 11.1. One set of 1000 simulations took about 5 minutes to run on a quad core MacPro.

[6] It is odd to not discuss accuracy of standard errors, but at this moment this seems like an important but secondary issue.

and the data were such that all predicted probabilities were admissible. Since such a DGP seems odd, and since the issue is often framed as whether logit is "worth it," this stacking of the deck in favor of logit does not seem overly wrong.[7]

The paper proceeds as follows. The next section compares logit and regression when there is no group structure to the data. While this comparison deals with a settled issue in political science, it appears that applied economists often still use regression and so these simulations are relevant. I then turn to TSCS data: Section 3 briefly discusses fixed effects with binary TSCS data and contributes to the comparison of simple logit with fixed effects to Chamberlain's (1980) "conditional logit" method. Section 4 returns to the comparison of logit and regression in the context of TSCS data with fixed effects.

2. LOGIT VS. REGRESSION

It may seem odd to begin by reopening what I thought to have been a settled debate: logit vs. regression for binary dependent variables model. I think that the current wisdom in political science is that logit (and related methods) and regression yield roughly similar t-ratios on the model coefficients in cases where regression is not predicting "probabilities" under zero or above one (or where most predicted probabilities are between roughly .3 and .7). Computational advances in the 1980's made it easy to estimate logit models, and by now most analysts are comfortable producing quantities of interest (marginal effects) in logit models. Thus it is rare to see a cross-sectional model with a binary dependent variable using regression methods.[8]

I was surprised to note that this is not the case in applied economics, even very current applied economics. While I had promised to provide no evidence on actual practice, the first three volumes of the *American Economic Journal: Applied Economics*, which began publication in 2009, contain about one LPM per issue. On conversing with applied economists, the source of the belief that regression is as good as logit for estimating marginal effects appears to be Angrist and Pischke's (2009) recent influential text, with the argument in the text based largely on Angrist (2001).

Angrist and Pischke (2009, 103) conclude that "[t]he upshot of this discussion is that while a nonlinear [probit, tobit] model may fit the CEF [conditional expectation function] for [limited dependent variables] more closely than a linear model, when it comes to marginal effects, this probably matters little. This optimistic conclusion is not a theorem, but ... it seems to be fairly robustly true.... Why, then, should we bother with non-linear models and marginal effects?" While this discussion is largely centered on marginal effects for (binary) treatment variables, and the ease of extending regression models to account for endogeneity, it may be that some applied economists, who are interested in estimating marginal effects in general, have taken this as a license to estimate the LPM instead of a more complicated non-linear limited dependent variable model.[9] Thus I begin the simulations by comparing

[7] For the analyses not involving regression there is no stacking of the deck.

[8] It is more common for time series data, where it is hard to combine logit with various time series issues. Time series analyses with a binary dependent variable are not common.

[9] I do not think that anyone believes that the LPM is the most likely data generating process, though

logit and regression for estimating marginal effects without any fixed effects.

To help assess Angrist's claim, I ran simulations which generated x as a standard normal and then generated y as a binary dependent variable using the inverse logit transformation of $\beta x + \alpha$ where both β and α were varied over small integers. For ease of writing, I refer to the situation where $y = 1$ as a "success" with the alternative being "failure'." Since I wanted to deal with situations where the probability of a success would be small rather than large, α varied from 0 to -4 and β varied from .5 to 3. [10] The number of observations was either 200, 400 or 1000.

I do not think that this is the setup envisioned by Angrist, but it seemed like a good place to start. I also compared variants of weighted regression (since with a binary dependent variable the regression model is heteroskedastic); weighting did not generally improve accuracy of estimation and so only results using simple regression are reported here.[11] The data appear very supportive of Angrist's claim, in that the percentage absolute percentage difference for estimating the average marginal effect of x via regression as opposed to logit was at worst 3.5% and in many situations regression outperformed logit (by similar puny margins). Only in one extreme case, where $\beta = .5, \alpha = -4, N = 200$ did OLS outperform logit by two percent; in eight other cases OLS was between 0.3% and .07% superior to logit.[12] All of these cases were ones with low values of β and small N's leading to insignificant or barely significant t-ratios on the estimate of β, $\hat{\beta}$. Obviously as β increases the logit response curve becomes more and more non-linear; Figure 1 shows that logit's superiority increases with β, though the maximal gain of logit never exceeds 3.5%. (Logit's superiority is also most marked with smaller sample sizes, as shown in the right panel.) So this section could presumably conclude that regression is pretty good for estimating marginal effects, and good enough so that the complexities of interpretation that come with logit are probably not worth it. But ....

Simulations of the properties of estimators usually do not worry much about how single exogenous variables are generated: x is commonly generated by a uniform or normal distribution. But more complicated distributions of the x may have interesting consequences for non-linear estimators (and particularly for estimating marginal effects). I there repeated the first experiment, but generated x with some negative skewness (so as to make successes rarer) and some excess kurtosis.[13] Such an x was generated as a (standardized) log normal,

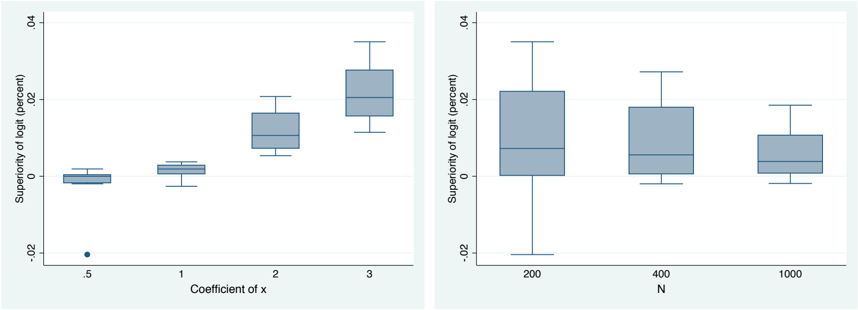
of course regression does have its usual optimal properties. But I do not think that anyone uses regression over logit because they believe that the former more accurately represents the underlying DGP. The average marginal effect for the regression model is simply the estimated coefficient; for the logit model it is the average (over the data) of $\hat{p}_i(1 - \hat{p}_i)\hat{\beta}$ where $\hat{p}_i$ is the estimated probability of success for unit i and $\hat{\beta}$ is estimated logit coefficient.

[10] The estimated probability of observing a one varied from .02 to just about .50 The t-ratio on β varied from just about 1 to 16. And yes, .5 is not an integer! 0 would not have been an interesting choice for β.

[11] OLS of course is consistent in the presence of heteroskedasticity and our sample sizes were relatively large; it may be the case that heteroskedasticity has more consequences of correct estimation of standard errors, but that issue is not discussed in this paper.

[12] Full tables of results will eventually be in an appendix.

[13] A normal has a kurtosis of three; the excess kurtosis of a distribution is the fourth central moment minus three. All data were standardized to have zero mean and unit variance.



(a) By β

(b) By N

Figure 1: Comparison of gain of logit over OLS by β and N

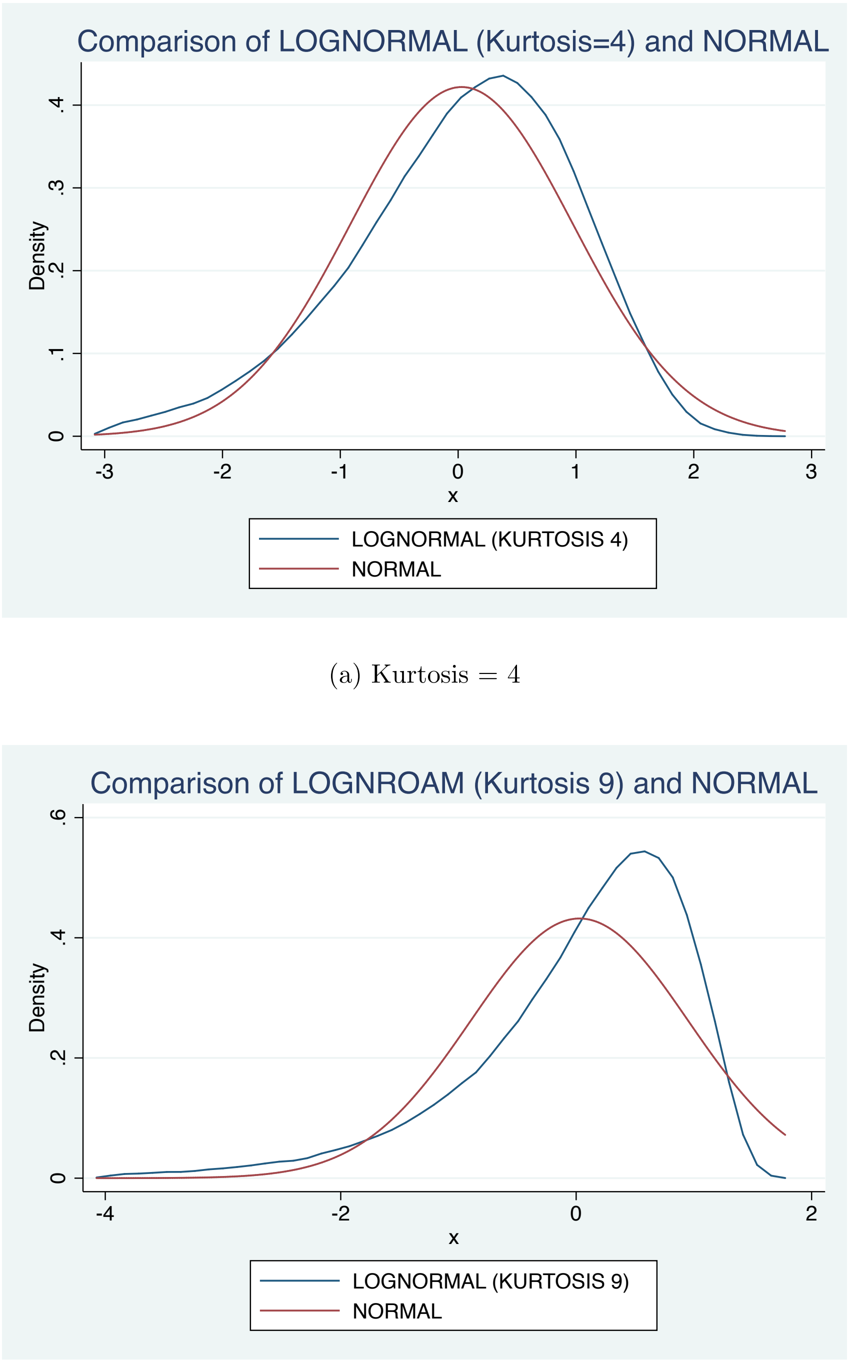
with the exponentiated normal having a standard deviation of either .5 or .25 (giving a kurtosis of 9 or 4 and skewness of -1.4 and -.6). Graphs of the two lognormal densities (compared to normals) are in Figure 2. While these densities clearly are not normal, they are also not extremely "wild."

For these lognormal x, I considered only large sample estimation, where $N = 1000$. α varied from -1 to -4 and β varied from 1 to 3. In none of the runs was regression better than logit; logit's advantage varied from about almost nothing to just over 50%. Logit's advantage was much more marked in the more kurtotic case, but even in the less kurtotic case logit's advantage over regression was between just barely positive and 30% (Figure 3).[14] Are our data more likely to resemble a normal or a lognormal, and in the latter case, how kurtotic are they likely to be? Of course no one knows, but there is no reason to believe that our data have a nice normal shape. Since, in this simple setup, the only advantage of regression is ease of interpretation, such ease *may* come at a non-trivial price.

Turning to the situation closer to that envisioned by Angrist, where y is generated as an inverse logit transformation of $\beta x + \delta D + \alpha$ where D is an exogenous dummy variable (so the simplest treatment variable). x and D were generated as correlated.[15] For these simulations I only studied the large N case, with all simulations generating 1000 "observations." Because these simulations required the variation of two coefficients (β and δ), the constant term varied over $0, -1, -2, -3$ and the coefficients varied over $1, 2, 3$. The quantity of interest in these simulations was the treatment effect, that is, $P(y = 1|D = 1, x) - P(y = 1|D = 1, x)$. Obviously for non-linear DGPs this quantity depends on x, with the most common quantity being the average of this difference over all the "sample" values of x (the average marginal

[14] As in the previous case, logit's superiority also grew with β.

[15] D was generated by taking a normal and then cutting it at either the 75th or 90th percentile to make the treatment increasingly rare; the latent for D and x were generated as the sum of two normals; correlation was ensured by their having one of these two normals in common. If we simply regressed y on D we would just have a difference of percentages test which would of course be correct no matter whether done by logit or regression. If x and D were not correlated it is likely that regression would still perform well.



(a) Kurtosis = 4

(b) Kurtosis = 9

Figure 2: Comparison of two lognormals with a standard normal

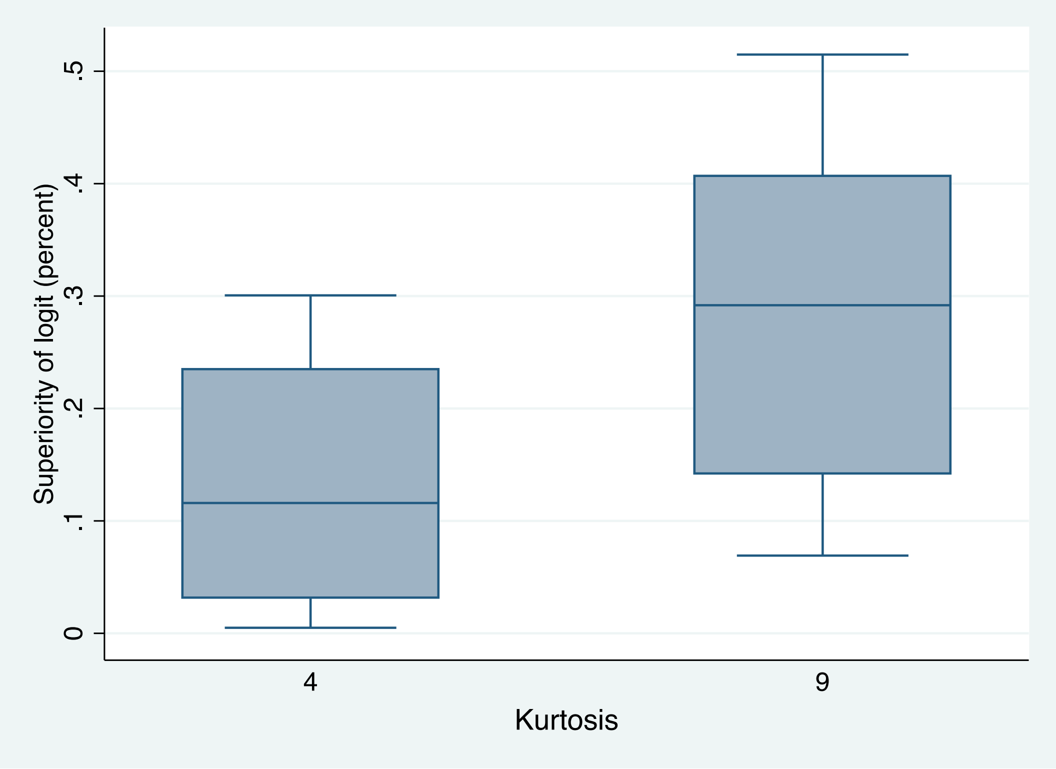


Figure 3: Difference in percentage accuracy of logit and regression by kurtosis of x

effect).[16] With our simple specification, the true marginal effect of x averaged over the data is $\frac{\sum_i p_i(1-p_i)\beta}{N}$, where p_i is the true probability of success for observation i. Regression is simpler; the marginal impact defined in any manner is just β. This, of course, is why regression is attractive to those who like to study marginal effects.To get a sense of scale, the average *true* marginal effect over all the experiments was between .08 and .49.[17]

Unlike the previous situation, logit typically outperforms regression even with nice normal data.[18] In about a quarter of the experiments logit and regression were within about 3% of each other; in some of these cases regression was a bit better than logit, but the differences here were less than 1%. While there is no simple relationship between the various coefficients and the percentage improvement of logit over regression, it is basically the case that the performance of regression degrades as the probability of success gets smaller. This is seen in Figure 4. This of course is consistent with our received wisdom that logit and regression are

[16] It is less common to compute the marginal effect at the mean (median) of x (the marginal effect at the mean (median)). The marginal effect at the mean may vary considerably in smallish samples.

[17] Since I am looking at absolute percentage deviations from truth, a small deviation from truth is large when the true value is small. Whether small deviations from small true values are as serious as large deviations from large values is an interesting question that does not have a simple answer. The same issues arises as we move from risk to relative risk with the same caveats applying.

[18] I do not report simulations with lognormal x. The difference between the logit and regression estimates of the marginal impact of D with lognormal x is similar to that with normal x; in these simulations D is generated in the same manner as in the normal simulations.

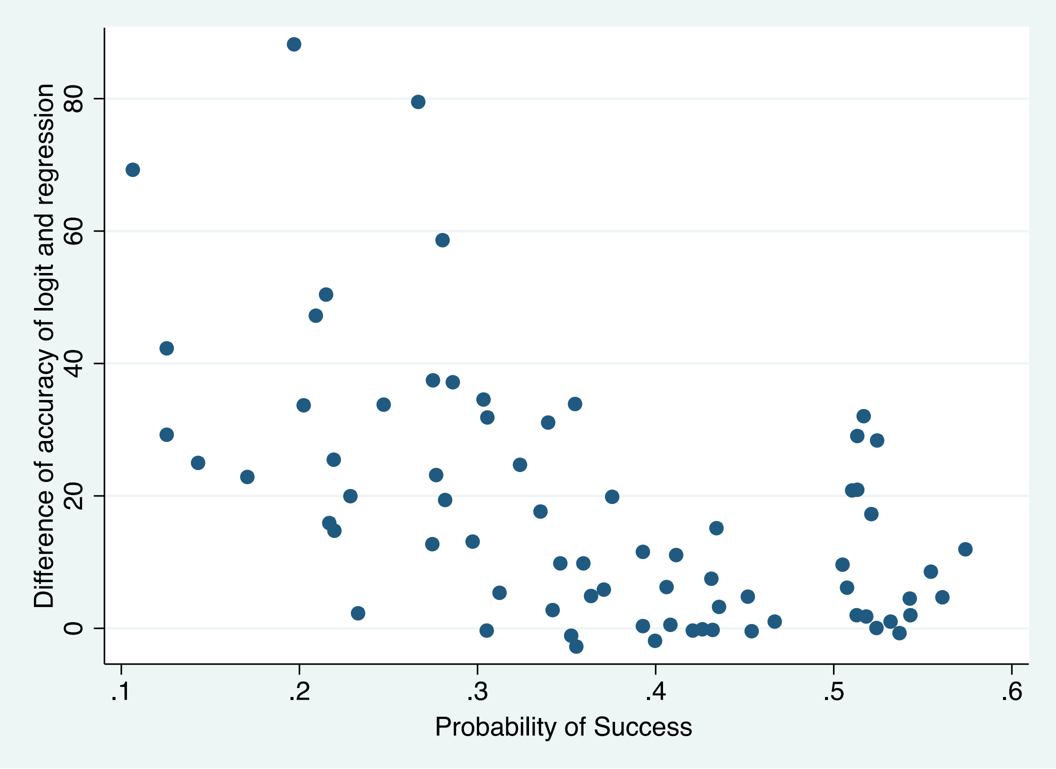


Figure 4: Difference in percentage accuracy of logit and regression in estimating the marginal effect of a treatment variable D where treatment is a function of $c + bx + dD$ and x is normally generated, x and D are correlated. Each dot represents the average of 1000 simulation runs with a given configuration α, β and δ. Y-axis is in percentage points and X-axis is average probability of success in the that sample.

similar where the probabilities of success and failure are about equal on average. Thus both low values for the constant term or rare successes both led to increase superiority of logit over regression. In half the cases this superiority was over 12%, and in a quarter of cases 30% or more.[19] Thus Angrist's sanguine position with respect to regression being adequate for estimating marginal effects is perhaps a bit too ruddy.

3. FIXED EFFECTS FOR TSCS DATA

I now turn to the real issue of this paper, what is a good way to estimate time-series–cross-section data models with a binary dependent variable and fixed effects (unit specific

[19] If one is worried that percentage error makes small differences look big, an analysis of absolute percentage differences yields similar results. While the absolute differences are, of course, smaller than the percentage differences, in about a quarter of cases the error in estimating the mean marginal effect using regression exceeds that using logit by .05 percentage points, reaching as much as .1; these are not trivial figures given that the average marginal effects were between .10 and .30. Qualitatively the absolute analysis shows that logit is better than regression in exactly the cases for the percentage analysis.

intercepts)? In such models we can write[20]

$$P(y_{it}=1)=h(\beta x_{it}+\alpha_i), i=1\dots N, t=1,\dots T \tag{1}$$

where h is some function (for our purposes either linear or inverse logit).

It may seem odd to study the estimation of this model in that I have argued elsewhere that estimating a model fixed effects with a binary dependent variable can have pernicious consequences (Beck and Katz, 2001) and that one should think of such data as event history data (Beck, Katz and Tucker, 1998) where fixed effects would be a bit odd. Fixed effects are most problematic where a large number of units never achieve a success (or always achieve a success). The simulations here allow for only a limited number of such units. But, more importantly, whatever my position on this issue, it appears that applied economists regard the inclusion of fixed effects as *de rigeur* so as to rule out a variety of other explanations for the data. Thus it seems interesting to discuss how to estimate Equation 1 without believing that this specification is inherently problematic.[21]

It also may seem odd that Equation 1 does not contain the dynamic or Markov transition structure that I have argued is superior (Beck, Epstein, Jackman and O'halloran, 2001). The Markov transition model is dynamic in that the current probability of success is different following success or failure. In this model

$$P(y_{it}=1|y_{i,t-1=1})=h(\beta_1 x_{it}+\alpha_1), i=1\dots N, t=1,\dots T \tag{2}$$
$$P(y_{it}=1|y_{i,t-1=0})=h(\beta_0 x_{it}+\alpha_0). \tag{3}$$

While to my mind this model makes sense for a wide variety of TSCS models with a binary dependent variable, this model neither lives well with fixed effects nor does it seem to be the common applied economics model. Since this paper is about logit versus regression and not specification of binary dependent variable models, I study the non-dynamic specification here, noting only that this is, at best, odd.

As noted in the introduction, maximum likelihood estimation of Equation 1 will be inconsistent for the case where $N\to\infty$ so long as T is finite or $\lim_{N\to\infty}\frac{N}{T}$ is finite. Neyman and Scott (1948) showed that one could get around this problem if one could estimate β conditional on the α_i. Chamberlain's conditional logit (denoted "clogit" here) does just this, and hence provides a consistent estimate of β (but no estimates of the α_i). While estimators are either consistent or not, the degree of error in estimating Equation 1 with a standard logit declines as T grows. Two recent papers (Coupé, 2005; Katz, 2001) have shown that the bias in standard logit (denoted "felogit" here) becomes small when T is above 16 or 20. However, these papers did not address how felogit compares to clogit in terms of absolute error properties. Since it seems easier to interpret felogit results it *may* seem easier felogit results it is important to know the price of this interpretative ease.[22]

[20] This is for a single scalar x; moving to a vector of x's is trivial and irrelevant for the purposes of this paper.

[21] The Markov transition model, by conditioning on previous success, will make it less likely that the unit specific intercepts are needed, though of course that is an empirical question which can be tested.

[22] The interpretative ease of felogit is that one can compute marginal effects of x for any of the units, since

The data setup for the simulation is as before, with the $y_{i,t}$ being generated via an inverse logit function and Equation 1. The simulation set N to 20 or 50 (N has little if any impact of the results, except in relation to T) and T was set at either 20, 30 or 50 (since I have no interest in the panel data issue here). Since the only case of interest is where x and the α_i are correlated, the α_i were drawn from a normal distribution and then offset[23] by a negative number so as to reduces the probability of success, another NT normal variates were drawn and then x was created as some constant times this normal plus the fixed effects. The constant here determines the degree of correlation between x and the fixed effects. Since for these simulations the estimates β are directly comparable, I simply compared the average mean absolute percentage (as a percentage of truth) errors in estimating β by felogit and clogit. The offset term for the fixed effects was either -1, -2, -3 or -4, β was set at either 1, 2 or 3 and the multiplier to control correlation of the x and the fixed effects was set at 1 or 2 (yielding R^2s in a regression of x on the fixed effects of either about .20 or .50).

One might conclude that over all the simulations the difference between clogit and felogit is small; while clogit always produces more accurate estimates of β, the worst difference between the two estimators is 7.5%. Not surprisingly the two estimators converge in accuracy as T increases; when $T = 50$ the difference between the two estimators ranges from 0.3% to 2%, when $T = 20$ the corresponding range is 1.4% to 7.6%. The advantage of clgoit over felogit also increase with β and declines with α, though the effect is small; similarly clogit's relative performance is slightly better when x and the effects are less correlated. But these differences are small when compared to the effect of group size.[24]

Should we just use felogit over clogit even though clogit might be, say, 7% more accurate (at best), that is when there are only 20 observations per group (see Figure 5). Clearly when we have 50 observations per group the difference between the two methods does become trivial (and I will leave 30 per group to the eye of the beholder). The choice cannot be based on computational time; with $T = 50$ and $N = 20$, 100 clogits took 30 seconds on my MacBookPro, while 100 took 12 seconds. There is the supposed interpretative simplicity of felogit, since marginal effects can be computed for any of the groups (taking the relevant intercept into account) while clogit can only produce a non-group specific marginal effect (conditioning only a group having at least one success). I postpone further discussion of this "advantage" until the next section. Now normally a 5% error is not worrisome (in a world where other assumptions lead to much larger errors). But the 5% increase in accuracy in clogit as compared to felogit is equivalent to clogit requiring 10% fewer observations to get the same level of accuracy. Before we throw away 10% of our observations we should be sure this is for more than saving 200 milliseconds of computing time. While obviously there is

one can estimate the α_i. In clogit the α_i are conditioned out and not estimated, so one can only compute the marginal effect of x conditional on at least one observation for a unit being a success. As we shall see in the next section, some of this interpretative difference is illusory.

[23] This offset is captured by a constant term in the logits or regression, but in the notation I prefer to keep fixed effects for all groups and suppress the overall constant.

[24] I also ran simulations where x was generated lognormal. Because clogit and felogit allow for non-linearities, I did not expect that the difference between clogit and felogit to be affected by how the x were generated. This proved to be the case, and the simulations for the lognormal x were almost identical to those for the normal x.

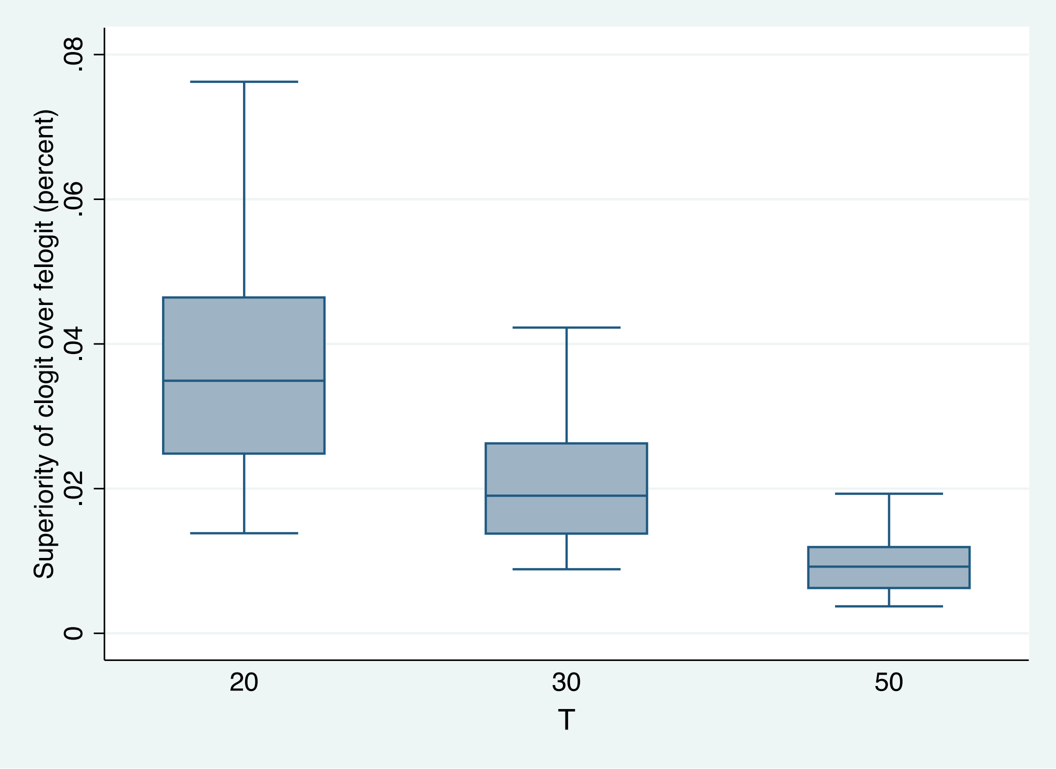


Figure 5: Difference in percentage accuracy of clogit and felogit by number of observations per group

no hard and fast cutoff for T, it does appear that for 20 observations in a group that the advantage of clogit over felogit is non-trivial. I return to this issue in the next section.

4. LOGIT VS. REGRESSION FOR TSCS DATA

For this section I return to comparing felogit and regression for estimating Equation 1. The reason for doing this is that I want to compare the estimated marginal effect of x and clogit does not allow this.[25] Why reconsider regression in this context? Because felogit drops any group which has no success while regression does not. When success is relatively rare this may lead to many groups being dropped from the estimation; regression uses all the groups in the data set. But what are the consequences of this for estimating the marginal effect.

First, though this is all well known, it is worthwhile to go over why felogit (and clogit) drop any group without a success (see any econometrics text for the derivation or Green, Kim and Yoon, 2001 or Beck and Katz, 2001 for a discussion related specifically to political science data). The basic idea is that the felogit model attempts (forgive the anthropomorphisms

[25] As noted, clogit only allows for the estimation of the marginal effect of x conditional on a group having at least one success. I return to the clogit/felogit issue below.

here) to match the successes and failures to the predicted probabilities of such. For any group with all failures, the probability of failure can be made as close one as desired by making the estimated fixed effect as small as possible. At that point the marginal effect of x for those observations goes to zero.

Another way to look at this is that if we use Equation 1 as the DGP, and observe some groups with no successes, then, for those group, any smaller values of α_i for those groups are equally consistent with the data; in other words, the estimates of α_i are unidentified for any group with no successes. Whether this is a bug or a feature is a bit unclear, but regression does not suffer from either this bug or fails to have this feature. By taking the zeros and ones as real data, regression simply tries to minimize the average of $(\beta x_{it} + \alpha_i - 1)^2$ for the successes and $(\beta x_{it} + \alpha_i)^2$ for the failures. Thus it always provides estimates of the fixed effects, and all sample data are used for estimating β. Whether these estimates are good estimates for estimating marginal effects is another question. And what are the "true" marginal effects?

Once thought about "correctly" these questions become easy, but it might be helpful to address these issues in the roundabout way that these simulations led me to think about these issues. I began by using the same simulation as before, this time generating the data using Equation 1 as the DGP. Data were generated as for the clogit/felogit comparison; coefficients were chosen so as to make probabilities low enough so that some groups had zero successes. Since zero successes in 20 independent draws is of course more likely than in 50 independent draws, I worked only with the 20 observations per group. Since the number of groups appears irrelevant, I worked with 20 groups only. The normal case suffices for this discussion. Since I wanted to make sure that about 5 groups on average had no successes, α_1 was set at either -3 or -4, β was set at 1 or 3 and the constant multiplier for correlation was set at 1 or 2.

I was puzzled by the initial results: regression did a much better job of estimating the average marginal effect of x than did logit. Over the 8 scenaria, regression was anywhere between 10 and 107 percentage points better than logit; as before, regression performed best when β was small (1). However, the puzzling result was simply an artifact of comparing lemons and limes (though not apples and oranges).

As the data generator I knew the true probability of success; thus the true average marginal effect was computed over all 400 observations, including groups with no successes. But felogit drop all groups with no successes and so was computing the average marginal effect over observations where a group had a least one success. On average about 7 groups were dropped over the course of the various scenaria. And because groups were more likely to be dropped if β were small (since a large β gave more chance for at least one observation in a group to have a non-trivial probability of success), regression's superiority to felogit when $\beta = 1$ simply reflects this difference.[26] In the dropped cases the average probability of success was about .02. Averaging the marginal effect of x in these dropped cases clearly leads to felogit overstating the average marginal effect in the whole sample. Note that when we compare felogit and regression on the same subset of the data, as before logit and regression

[26] When $\beta = 1$ about 9 groups out of 20 were dropped; the corresponding figure when $\beta = 3$ was about 5 groups dropped.

perform similarly (never deviating by more than a few percentage points).[27]

The DGP in this section has relied on the fixed effects to produce groups with no successes. What happens if we intervene in this process, and radically reduce the α_i (to -8 for the dropped groups, which yields a probability of success in those groups of .0003 when $x = 0$.[28] Such a change has no effect on the felogit results, since these groups are dropped in either case. Interestingly regression copes well with this case also, outperforming felogit a bit more markedly than in the previous case for estimating the average marginal effect in the full sample and essentially tying felogit for the average marginal effect in the restricted sample.

It should be remembered that that the consequence of changing fixed effects is different in the LPM and logit worlds. For the LPM, a unit change in a fixed effect has enormous consequences, moving probabilities outside the unit interval. For logit, of course, the effect is a function of average probabilities. Thus, for example, moving a fixed effect from zero to -4 (when all other covariates are at zero) decreases the probability of success from .5 to just under .02. A similar decrease to -8 drops that probability to .0003. But even with a fixed effect of -4 about 70% of the time 20 observations will be all zero; that rises to 99% as the fixed effect drops to -8. But once a group has no successes, it is observationally equivalent (in terms of constants) to any other group with no successes, regardless of the true DGP.[29]

One consequence of this is that regression does worse (in terms of mean absolute error of estimating the average marginal effect) as the fixed effects of a few group are made more and more extreme (negative). Thus moving from changing a quarter of the groups to having a fixed effect of -4 to a fixed effect of -8 increased the mean absolute error by 50%.[30] It is also the case that probabilities of success for a given observation are not well estimated. While averaged over all the dropped groups, the probability of success was zero to about 10 decimal places, even for large special fixed effects (-8) the mean absolute predicted probability was .08, meaning that lots of predicted probabilities were below -.08. This error rose, *for the dropped groups* as the large special effects were set to more reasonable values; when the five extreme special effects were set to -2 the mean absolute prediction was .13 even though the average prediction error for those groups was still zero (to 10 decimal places).[31] Thus there were a large number of predicted negative probabilities (of .1 or more) for observations in groups with no successes.

If one wants the average marginal including groups with no successes, it appears that regression does not do a terrible job. It is hard to know whether regression is doing a good job, since all that can be done is to compare estimators. But the simulations do seem to indicate that regression does less well as the fixed effects are set to smaller and smaller values.

Note that the marginal effect of x in the dropped cases varies in the two scenaria. As the

[27] This of course is with a normally generated x. The lesson from Section 2 is that logit may outperform regression for less nicely generated x.

[28] To keep the simulations a bit realistic, the five smallest fixed effects were reduced; originally the fixed effects were generated as a standard normal, so in general the smallest generated effect was not much under -2 before intervention.

[29] Observational equivalence given the data is not the same thing as not being identified.

[30] These simulations covered relatively few combinations of the various parameters, and so the 50% should just be taken as an indication and not the average of a wide variety of parameter settings.

[31] This increase was smooth over the intermediate settings of these fixed effects.

fixed effects get smaller, the marginal effect in the dropped groups get smaller (a fact that can only be known by the data generator). Given that the $\hat{\alpha}_i$ estimated by regression are a bit notional (estimable only because regression takes zero as being zero and not just an indication which part of the real line $\alpha_i + bg$ lies on), perhaps we should restrict ourselves to estimating conditional effects only in groups with at least one success. But only an analyst can make a guess as to whether the groups with no successes are that way because the fixed effects for these groups are only a bit small (so that average probabilities are large enough so that the true marginal effects of the covariates in these groups are small but non-trivially above zero) or whether these effects are extremely small (so the true marginal effects of the covariates in the dropped groups are zero to several decimal places). This guess can be somewhat informed by data, that is, examining whether the x's in the dropped cases are markedly lower than in the non-dropped cases. But in the end the analyst who decides that it makes sense to estimate the average marginal effect of x over all the cases must base that decision on some knowledge outside the data.[32]

Returning to felogit versus clogit, it is clear that the advantage of felogit over clogit for interpretative purposes is illusory. While we may think of felogit as estimating the average marginal effect, it really estimates the average marginal effect of x given that a group has a least one success. This is exactly the marginal effect that can be computed with clogit. Since felogit can only estimate *any* marginal quantity conditional on a group having at least one success, for most purposes it has no interpretative advantage over clogit, which estimates the same thing.

5. CONCLUSION

Obviously it is hard to be confident about generalizing from a simulations based on a small number of scenaria. I trust that the scenaria presented are not too unrealistic, though of course it is hard to know what it means for a scenario to be realistic. But at least a few tentative conclusions can be suggested.

Angrist has argued that regression is just about as good as logit for estimating average marginal effects. In the simple case where the probability of success is a simple function of x, this appears to be true when x is normally generated. When skewness and kurtosis were added to x the result was not so favorable to regression. The treatment effect scenario was also not so favorable to regression. We must remember that the DGP process was logit, so perhaps the comparison is unfair. But it is hard to think of a plausible DGP for a binary dependent variable that does not resemble a logit. Obviously one can find what to me are implausible DGPs where regression wins over logit, but it is hard to do that in a world of a linear index model, that is, where $P(y_i = 1) = h(\beta x)$. It surely is the case that the linear probability model is not a plausible DGP. Since the only advantage of regression over logit is interpretative,[33] it seems to me that logit is safer than regression, even for estimating

[32] I have no idea if a Bayesian approach would help here, but surely this sentence does beg for some Bayesian analysis.

[33] This is for exogenous right hand side variable; linear models have other advantages for endogenous right hand side variables.

marginal effects.

The second conclusion, which I think is on firmer ground, is that clogit is only a bit better than felogit for larger group sizes (20 or more), there is no reason to prefer felogit on interpretative grounds. Both clogit and felogit are estimating the same quantity of interest, the marginal impact of a variable in groups with at least one success.

Do we want to ignore the marginal effects in groups with no successes. If not, regression is needed (at least as compared to logit), and the results indicate that regression is not terrible (with the degree of terribility varying in the usual way). It is up to applied researchers to think about whether marginal effects in groups with no successes are part of any quantity of interest.

REFERENCES

Angrist, Joshua D. 2001. "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice." *Journal of Business & Economic Statistics* 19:2–16.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Beck, Nathaniel, David Epstein, Simon Jackman and Sharyn O'halloran. 2001. "Alternative Models of Dynamics in Binary Time-Series–Cross-Section Models: The Example of State Failure." Paper presented at the Annual Meeting of the Society for Political Methodology, Emory University.

Beck, Nathaniel and Jonathan N. Katz. 2001. "Throwing Out the Baby With the Bath Water: A Comment on Green, Kim and Yoon." *International Organizations* 55:487–95.

Beck, Nathaniel, Jonathan N. Katz and Richard Tucker. 1998. "Taking Time Seriously: Time Series Cross Section Analysis with a Binary Dependent Variable." *American Journal of Political Science* 42:1260–1288.

Chamberlain, Gary. 1980. "Analysis of Covariance with Qualitative Data." *Review of Economic Studies* 47:225–38.

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

Green, Donald, Soo Yeon Kim and David Yoon. 2001. "Dirty Pool." *International Organizations* 55:441–68.

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

Neyman, J. and E. L. Scott. 1948. "Consistent Estimation Based on Partially Consistent Observations." *Econometrica* 16:1–32.