Throwing Out the Baby With the Bath Water: A Comment on Green, Kim and Yoon

Nathaniel Beck[1] Jonathan N. Katz[2]

Draft of November 9, 2000. Word Count: 4300 (approx.)

[1]Department of Political Science; University of California, San Diego; La Jolla, CA 92093; phone (619) 807 1530; fax (503) 905 9476; beck@ucsd.edu

[2]Division of the Humanities and Social Sciences; California Institute of Technology; Pasadena, CA 91125; phone (626) 395 4030; fax (626) 405 9841; jkatz@caltech.edu

Nathaniel Beck

Professor

Department of Political Science

University of California, San Diego

La Jolla, CA 92093

beck@ucsd.edu

and

Jonathan N. Katz

Associate Professor

Division of the Humanities and Social Sciences

California Institute of Technology

Pasadena, CA 91125

jkatz@caltech.edu

Katz thanks the John. M. Olin Foundation for a Faculty Fellowship supporting his research

Abstract

Green, Kim and Yoon argue that many findings in quantitative international relations which use the dyad-year design are flawed. In particular, they argue that for the effect of democracy on both trade and conflict has been vastly overstated. They argue that researchers had ignored unobserved heterogeneity between the various dyads. They then argue that heterogeneity can be best modeled by "fixed effects," that is, a model which includes a separate dummy for each dyad.

We argue here that the use of fixed effects is almost always a bad idea for dyad-year data with a binary dependent variable like conflict. This is because conflict is a rare event, and the inclusion of fixed effects requires us to not analyze dyads that never conflict. Thus while the 90% of dyads that never conflict are more likely to be democratic, the use of fixed effects gives democracy no credit for the lack of conflict in these dyads. Green, Kim and Yoon's fixed effects logit can tell us little, if anything, about the pacific effects of democracy.

Their analysis of the impact of democracy on trade is also flawed. The inclusion of fixed effects almost always masks the impact of slowly changing independent variables; the democracy score is such a variable. Thus it is no surprise that the inclusion of dyadic dummy variables in their model completely masks the relationship between democracy and trade. We show that their preferred fixed effects specification does not out perform a model with no effects (when that model is correctly specified in other ways). Thus there is no need to include the masking fixed effects, and so Green, Kim and Yoon's findings do not overturn previous work that found that democracy enhanced trade.

We agree with Green, Kim and Yoon that it is important to model heterogeneity in time-series–cross-section data. We mention a number of alternatives to their fixed effect approach,

none of which would have the pernicious consequences of using dyadic dummies in their two re-analyses.

1 Introduction

Green, Kim and Yoon (hereinafter GKY) [NOTE TO COPY EDITOR: In this version GKY is not in our reference list and we never have a footnote to it. Should we. If so, what is title, year, etc? Please note we do NOT mention either King or Oneal/Russett] contribute to the literature on estimating pooled times-series–cross-section (hereinafter TSCS) models in International Relations (hereinafter IR). They argue that such models should be estimated with fixed effects when such effects are statistically necessary. While we obviously have no disagreement that sometimes fixed effects are appropriate, we show in this response that fixed effects are pernicious for IR TSCS models with a binary dependent variable (hereinafter BTSCS models) and that they are often problematic for IR models with a continuous dependent variable. In the binary case, this perniciousness is due to many pairs of nations always being scored zero, and hence having *no* impact on the parameter estimates; for example, many dyads never come into conflict. In the continuous case, fixed effects are problematic in the presence of the temporal stable regressors that are common IR applications, such as the dyadic democracy measures used by GKY.[1]

We focus here on what we feel are the critical defects of the GKY fixed effects approach for modeling typical IR applications. Since our response is critical, we do stress that sometimes fixed effects make sense for TSCS, although probably not for BTSCS, data. Like GKY, we believe it is always better to account for dyadic differences by theoretical variables, but this may not always be possible. Thus sometimes fixed effects are appropriate, but of course no one should be content to "explain" American-British trade by a dummy variable which corresponds to the dyadic name. Further, we agree with GKY that ignoring unmodeled heterogeneity, that is, dyadic differences that are not captured by the independent variables, *may* be a serious problem. But, for typical IR problems, and specifically for the analyses presented by GKY, we find their fixed effects model to be profoundly misleading in assessing the impacts of important independent variables. We stress that we are not simply talking about some minor changes in estimation efficiency, but, rather, estimates that are so far off as to be completely useless.

The next section of this response shows that the use of fixed effects is clearly a bad idea for the binary dependent variable case. The following section considers the continuous dependent variable case. While each section focuses on the specific analyses offered by GKY and why fixed effects models are not appropriate for those analyses, we also offer positive suggestions on how IR researchers might estimate models with heterogeneous units. The concluding section deals with the general issue of the utility of fixed effects models.[2]

2 BTSCS and fixed effects

We have argued elsewhere that IR BTSCS data, such conflict data, is essentially event history data, where each dyad is observed to either still be at peace or to have begun a conflict in any given year.[3] While we argued there for grouped duration analysis ,[4] dyadic conflict data can be analyzed by any event history method (of which logit is one such, albeit flawed, method since it does not account for the temporal dependence of the data). The first thing we note is that event history analysis is a commonly used method in the social and biomedical sciences. GKY's argument is that these event history analyses should contain a dummy variable for each unit that is observed in the sample. However, we know of not a single event history analysis which uses a unit dummy variable. If GKY are correct, then *every* event history analysis that we know of is suspect.

The problem with fixed effects in event history analysis can be seen by considering GKY's attempt to model dyadic conflict presented in their Table 3. As can be seen from the table most dyads never conflict; in fact, over 93% of GKY's dyads — 2877 out of 3078 dyads – never do. The inclusion of fixed effects allows for perfect prediction of almost all the dyads;

as GKY agree, this means that over 90% of the dyads have no impact on the statistical estimates. Thus, a data set which contained only the 10% of the dyads which conflict would yield *identical* estimates to the full data set (including dyads which never conflict).

Why do the over 90% of the pacific dyads not effect the logit fixed effects estimates? For any such dyad, we would like the coefficients to be such that the probability of conflict is as close to zero as possible. To do this, choose the coefficient on the fixed effect for that dyad to be as negative as possible; this will drive down the estimated probability of conflict for all the yearly observations on that dyad to zero. Thus the other independent variables have no impact on the estimates, since no matter how they change, we can simply make the fixed effect more and more negative, ensuring the estimated probability of conflict remains near zero. Thus for these dyads, the independent variables other than the fixed effect tell us nothing about the probability of conflict. The fixed effects logit assumes that these pacific dyads do not conflict because of some unmodeled idiosyncratic feature of the dyad, and that the substantive independent variables for that dyad are thus irrelevant to explaining its lack of conflict.

To see why the GKY's approach is pernicious, let us start with a biomedical example where the intuition is easily developed. Suppose one wanted to assess the effect of the presence of some gene on the occurrence of some cancer. If we only observed the presence or absence of the gene, and whether the subject had cancer, we would conduct a standard logit analysis. Thus, for example, if 90% of the subjects without the gene were cancer free, whereas only 50% of those with the gene were cancer free, we would find that the gene is significantly associated with cancer (without of course having a clear causal inference). This is the equivalent of a cross sectional study asking whether democratic dyads are less likely to ever conflict then are non-democratic dyads.

Now let us add some longitudinal data. Suppose we follow subjects for five years, noting each year whether or not they got cancer; once a subject is observed with cancer, no further observations are made. We could analyze this data with various event history methods, including a logit (of course properly specified to take temporal dependence into account). 90% of the cases without the gene, by assumption, never get cancer. Thus, *using fixed effects*, these non-cancerous observation make *no* contribution to the statistical analysis (i.e., the likelihood). We would thus end up examining logit results based on only the 10% of cases without the gene but who got cancer and the 50% of cases with the gene who also got cancer. With such data we would likely conclude that the gene is unrelated to cancer, even though the gene is clearly related to cancer (by construction in this example).

Note that we could alternatively estimate the same genetic effect by a standard event history method which takes each subject and models the time until cancer is observed (or whether no cancer is observed after five years).[5] In this case, we would clearly never think about adding one fixed effect for each observation, since the fixed effect for any subject would completely determine the predicted duration for that subject, and no independent variable could possibly have any impact. Since estimating BTSCS data via logit or cross-sectional event history methods is not conceptually different, one method should not be seen as allowing for fixed effects whereas the other clearly cannot allow it. In short, the ability to add fixed effects to a BTSCS model (albeit with a loss of 90% of the data) is illusory. In our hypothetical example analyzed with standard event history techniques we would correctly find an effect of the gene on cancer rates.

GKY are not unaware of this issue. They argue [NOTE TO COPY EDITOR: the quote is on page 25 in the ms, we need a better page number, here we really do need to figure out citations] that if we discovered new democratic dyads, that were always pacific, it would give us *no* information, because "we do not know the base probability (the intercept) of war for each of these new dyads...." Now we freely admit that it is logically possible that these new dyads might be pacific because of the name of the dyad (the fixed effects) or because both partners both grow green beans. But it seems odd to throw out the only theoretical explanation we have, that the dyad is pacific because it is democratic.

To see how odd this position is, let us go back to the simple logit data, where we have only one observation per dyad. Following GKY's logic, we could do no analysis, because dyadic differences might be due to difference in their own intercept (the "base line probability") rather than differences in democracy scores. Thus GKY's logic rules out any cross-sectional studies (with any type of dependent variable), unless they are done experimentally. While we certainly like experiments, we do not believe that IR can only proceed via experimental studies.

In short, GKY's conclusion, in Table 3, that variables such as democracy have no pacific impact, is simply nonsense. It is absurd to exclude over 90% of the cases from the analysis (or, equivalently, to allow them in the analysis but not allow them to affect any statistical results) and then conclude that some independent variable like democracy has the opposite effect of what every sensible study has shown. One could take the essentially nihilist position that any cross-sectional variation *could* be due to idiosyncratic factors, but that is not a position taken in any other type of empirical analysis in political science. Because IR BTSCS data frequently contains a lot of units which show no temporal variation on the dependent variable, GKY's proposal to include fixed effects in these analyses is *never* a good idea.[6]

Fortunately, it is not necessary to resort to fixed effects to model dyadic heterogeneity. There are many well-known ways to model heterogeneity in event history data, none of which are subject to the problems of the fixed effects solution. One popular model would be the Weibull duration model with gamma heterogeneity.[7] But given the nature of the data, a solution along the lines of adding frailty to the Cox proportional hazards model, that is allowing each unit to vary randomly in its probability of conflict (as well as varying systematically via the independent variables) might prove better. Another alternative would be a split population model, where some dyads never conflict whereas other might eventually come into conflict.[8] All of these offer alternative estimation methods which allow for unmodeled heterogeneity without the serious side effects of fixed effects estimation.

3 TSCS data with continuous dependent variables

The fixed effects estimator is not quite as problematic in the continuous dependent variable case. GKY use fixed effects to estimate a model on the political economy of trade (presented in their Table 2). No dyads have constant trade, and therefore no dyads are dropped in the fixed effects columns of Table 2 (Columns 3 and 5). While it appears that fixed effects are clearly important in the static model (column 2), this is a highly misspecified model since it incorrectly ignores dynamics. The coefficient of 0.736 on the lagged trade variable in Column 4 tells us that the static model in Column 2 is badly misspecified. Standard time-series arguments tell us that this misspecification has very serious consequences, which can be seen by comparing the estimates in Columns 2 and 4.

Thus we agree with GKY that Column 2 dramatically overestimates the role of democracy in determining trade; this overestimate has *nothing* to do with ignoring fixed effects and everything to do with ignoring dynamics. Failure to correctly model the dynamics, either through generalized least squares or, better, via the inclusion of a lagged dependent variable, makes it appear that fixed effects are very important. This is because fixed effects essentially add a lagged dependent variable with a coefficient of one to the model; it may appear that such an fixed effects are necessary if the baseline model is the incorrect static model. We therefore focus on the impact of including fixed effects in a correctly specified dynamic model, that is a comparison of Columns 4 and 5.

Comparing Columns 4 and 5, we note that fixed effects explain very little additional variance. The 3,078 additional dummy variables increase the explained variance from 73% to 77%. GKY's F-test does, however, indicate that we can reject the null hypothesis that fixed effects can be ignored. This F-test is quite likely to reject the null hypothesis of no fixed effects, since with almost 90,000 degrees of freedom we have essentially perfect estimates of all coefficients. There are, however, other ways to choose between models. One popular method, common in applied time series analysis, is the BIC. The BIC, like other model selection criteria, judges models by their sum of squared residuals plus a penalty for lack of parsimony; the BIC has a larger penalty than the common AIC (which is very similar to an F-test).[9] The BIC clearly favors the dynamic model *without* fixed effects.[10] Thus on standard model selection grounds there is good reason to choose the model without fixed effects over GKYs fixed effects model.

But even if we think that the fixed effects model is superior, the similarity of performance of the two dynamic models, with and without fixed effects, means that estimating a model ignoring fixed effects simply cannot produce very biased estimates. So even if we concede that the dynamic model in Column 4 suffers from possible omission of fixed effects, the consequences of this omission cannot be great.

But why not include fixed effects, that is, why not take the estimates in Column 5 seriously? We should always be wary of statistical cures that may be have serious side effects, especially when the illness being "cured" is not very serious. The GKY fixed effects "cure" for Column 4 is akin to the proverbial "chemotherapy for a cold." Obviously including fixed effects means that any independent variable that does not vary temporally cannot be used as an explanatory variable. Thus GKY cannot assess the impact of geography on trade. Relatively few interesting independent variables are temporally constant, although many are almost constant. These variables, like democracy, which vary little from year are to year, are highly co-linear with the 3078 fixed effects.[11] It is quite likely, then, that the use of fixed effects will yield odd estimates of coefficients for variables like democracy, since the effect of democracy is then "controlled" for the fixed effects.

This is not to say that fixed effects never make sense for TSCS data with a continuous dependent variable. There clearly will be cases where the fixed effects have greater explanatory power than they do in the dynamic model of trade (though we suspect that modeling dynamics via a lagged dependent variable will generally make fixed effects much less relevant). Further, there clearly are models where the independent variable of interest show year to year variation, and so are not quite so highly co-linear with the fixed effects as in GKYs trade model.

Even, however, where fixed effects are indicated, we agree with GKY that fixed effects models are never ideal.[12] We should clearly attempt to find substantive variables that explain dyadic differences; to simply allow for dummy variables which indicate dyadic names to explain any dependent variable can hardly be very interesting. But what should analysts do if they do not know of any explanatory variable which explains the fixed effects? One possible solution, which again has none of the bad consequences of GKYs fixed effects model, is the hierarchical or random coefficients model.[13] The random coefficients model not only allows intercept terms to vary, it also allows the slope coefficients to vary from unit to unit (and this variation can be modeled as a function of other explanatory variables). This model allows for the dyadic variation that GKY feel is necessary (more than what the fixed effects model allows for) without making it impossible to estimate coefficients for variables which are temporally stable. There is however, no doubt in our minds that if one had to choose between the estimates of the dynamic trade model in Column 4 and the fixed effects model in Column 5, the model *without* fixed effects is far superior for assessing the impact of variables like democracy on trade.

4 Conclusion

The logic of GKY is that all cross-sectional analyses are suspect, because unit specific baselines are not included. The logic of their argument holds for all cross-sectional analyses, including the garden variety regressions we see run on surveys every day. It is possible that two respondents differ in their preferences because of idiosyncratic features, but would we not prefer to explain these differences by differences in explanatory variables such as social class? The GKY position implies that only experimental study allows for any inferences, whether causal or not. GKY would not only overthrow much quantitative IR, they would overthrow every non-experimental result ever obtained.

Of course they do not go this far, since putting in fixed effects is clearly silly in simple cross-sectional analyses. Unfortunately, TSCS data allows analysts to propose almost silly estimators, because the repeated observations allow such estimators to produce results that might appear meaningful at first glance.

We certainly agree with GKY (and Leamer and many others) that one should examine the robustness of finding to alternative specifications and methods. But to expect findings to be robust to odd specifications and or methods is a foolish expectation. While GKY make a correct point, that *sometimes* fixed effects should be included in a TSCS model (although it is probably incorrect to say they should ever be included in a BTSCS model), there is nothing in their analyses of trade or conflict that should be seen as challenging any currently standard estimates.

We close by agreeing with GKY that the assumption of complete homogeneity of data, across both units and time, is usually suspect. Our own work has attempted to provide some estimation methods which allow for temporally or geographical dependent data, and in Sections 2 and 3 we have provided some citations for useful ways of attempting to model heterogeneity. While there may be some cases where fixed effects are appropriate, these other avenues appear to us to be both more promising and less likely to produce useless estimates than does the fixed effects model.

Notes

[1]Obviously fixed effects do not work if there is an independent variable that varies only cross-sectionally, as, for example, GKY's distance variable.

[2]For reasons of space, we focus our discussion entirely on the consequences of fixed effects estimation, and do not discuss other issues.

[3]Beck, Katz and Tucker 1998

[4]Grouped duration analysis assumes that the timing of events is only observed discretely; the conflict data sets only tell us whether conflict occurred in some year. Grouped duration data is distinguished from continuous time duration data, where the timing of events is known exactly. The difference between the two types of data is not critical for our argument here.

[5]See Alt, King and Signorino 2001, Beck 1998 or Sueyoshi 1995 for more extended discussions of the theoretical equivalence of cross-sectional duration models with BTSCS models.

[6]This is not to say that we accept the specification in Columns 1 and 3. These ordinary logits do not model temporal independence correctly and do not model the dynamics correctly. But these problems can be addressed without recourse to fixed effects. See Beck and Tucker 1997 and Beck, Katz and Tucker 1998.

[7]This model, and the frailty model mentioned below, allow units to be more heterogeneous

than would be allowed by their simpler variants. Both models add randomness to each unit's underlying propensity to fail, either as a simple stochastic term or as a function of some explanatory variables combined with a stochastic term. In the biomedical medical literature, this is called "frailty," since some individuals (those that are more frail) are more likely to die regardless of the values of the observed independent variables. Frailty is a solution to GKY's unmodeled heterogeneity problem, a solution which does not have the draconian consequences of GKY's fixed effects. See Greene 1999 947 for a discussion of the heterogeneous Weibull model and Sargent 1998 for a discussion of the frailty model.

[8]This has been investigated in the biomedical world, where such models are called cure models. In these models, some patients are cured whereas others will eventually suffer relapse if we wait long enough. See Tsodikov 1998. In the criminological world, some ex-prisoners will never return to prison, while others will be recidivists. See Schmidt and Witte 1989.

[9]In terms of model selection, the AIC is equivalent to GKY's F-test. Another common method of model selection, based on choosing a model with the larger R^2. picks even less parsimonious model than does the AIC, though as the sample size becomes large, maximizing $\bar{R}^2$ becomes equivalent to the AIC (or GKY's F-test). The various criteria differ only in their penalty for lack of parsimony, with the $\bar{R}^2$ having the smallest penalty, followed by the AIC and F-tests, and the BIC having the largest penalty. The penalties for lack of parsimony

decline with sample size, but the BIC has the slowest rate of decline. (The penalty for the AIC is $\frac{k}{N}$ where k is the number of model parameters and N is the sample size; this penalty becomes trivial as N gets large. The penalty for the BIC is $\frac{k \log N}{N}$ which always exceeds the AIC penalty.) Applied researchers prefer the BIC because it picks more parsimonious models than do the other criteria; parsimonious models, in general, have better out of sample forecasting properties than do more complex models, even if the latter show better in sample fit. See Judge, et al. 1985 for a discussion of model selection criteria.

[10]The BICs for the two models are 1.94 and 2.19.

[11]We know this must be so, since the inclusion of the fixed effects changes the coefficient of democracy enormously. Another way to see this is to note that the fixed effects model first regresses the independent variables of interest and the dependent variable on the dyadic dummies, and then estimates the parameters of interest by regressing the residuals from these regressions on each other. If independent variables like democracy are very stable for any dyad, then the dyadic dummy will explain democracy quite well. The use of fixed effects implies that we only care about whether the small part of democracy that is temporally unstable explains trade. GKY show the same thing with their Hausman tests. Since no one could doubt the fixed effects radically change coefficient estimates, the Hausman test used by GKY tells us nothing that is not obvious. The purpose of our response is to inquire whether

the new estimates, based on the fixed effects model, are in any way more useful than the estimates obtained without using fixed effects. No one could doubt that the use of fixed effects radically changes all estimated impacts.

[12]If one were committed to fixed effects, then for dyads we prefer the vastly more parsimonious specification that models the dyadic fixed effect as the sum of its two component fixed effects. While this is not a good solution for the BTSCS case, it is far superior to the full fixed effects specification of GKY. See Mansfield and Bronson 1997 or Beck and Tucker 1997 for a discussion of this approach.

[13]This is a well known model in statistics and econometrics. See Western 1998 for a good introduction to this model in a political economy context.

References

Alt, James E., Gary King and Curtis Signorino. 2001. Estimating the Same Quantities from Different Levels of Data: Time Dependence and Aggregation in Event Process Models. *Political Analysis*. 9(1):21–44.

Beck, Nathaniel. 1998. Modelling Space and Time: The Event History Approach. In *Research Strategies in Social Science*, ed. Elinor Scarbrough and Eric Tanenbaum. Oxford University Press pp. 191–213.

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(4):1260–88.

Beck, Nathaniel and Richard Tucker. 1997. "Conflict in Time and Space." Center for International Affairs Working Paper 97-8, Harvard University, `https://wwwc.cc.columbia.edu/sec/dlc/ciao/wps/tur01/`.

Greene, William. 1999. *Econometric Analysis*. Fourth ed. Upper Saddle River, N.J.: Prentice Hall.

Judge, George, W. E. Griffiths, R. Carter Hill, Helmut Lütkepohl and Tsoung-Chao Lee. 1985. *The Theory and Practice of Econometrics*. Second ed. New York: Wiley.

Mansfield, Edward D. and Rachel Bronson. 1997. "Alliances, Preferential Trading Arrangements, and International Trade." *American Political Science Review*. 91(1):94–107.

Sargent, Daniel J. 1998. "A General Framework for Random Effects Survival Analysis in the Cox Proportional Hazards Setting." *Biometrics*. 54:1486–97.

Schmidt, P. and A. Witte. 1989. "Predicting Criminal Recidivism Using Split Population Survival Time Models." *Journal of Econometrics*. 40:141–59

Sueyoshi, Glenn T. 1995. “A Class of Binary Response Models for Grouped Duration Data.” *Journal of Applied Econometrics*. 10(4):411–31.

Tsodikov, A. 1998. “A Proportional Hazards Model Taking Account of Long-Term Survivors.” *Biometrics*. 54:1508–15.

Western, Bruce. 1998. “Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach.” *American Journal of Political Science*. 42(4):1233–59.