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

Nathaniel Beck and Jonathan N. Katz

Donald P. Green, Soo Yeon Kim, and David H. Yoon contribute to the literature on estimating pooled times-series cross-section models in international relations (IR).[1] 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 here that they are pernicious for IR time-series cross-section models with a binary dependent variable and that they are often problematic for IR models with a continuous dependent variable. In the binary case, this perniciousness is the result of 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 temporally stable regressors that are common IR applications, such as the dyadic democracy measures used by Green, Kim, and Yoon.[2]

We focus here on what we feel are the critical defects of Green, Kim, and Yoon's fixed-effects approach for modeling typical IR applications. Since our response is critical, we do stress that sometimes fixed effects make sense for time-series cross-section data, although probably not for binary time-series cross-section data. Like Green, Kim, and Yoon, we believe it is always better to account for dyadic differences with 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 that corresponds to the dyadic name. Further, we agree with Green, Kim, and Yoon that ignoring unmodeled heterogeneity, that is, dyadic differences that are not captured by the indepen-

We thank Matt Baum for research assistance. Jonathan Katz thanks the John. M. Olin Foundation for a Faculty Fellowship supporting his research.

1. Green, Kim, and Yoon 2001.

2. Obviously fixed effects do not work if there is an independent variable that varies only cross-sectionally, as does, for example, Green, Kim, and Yoon's distance variable.

International Organization 55, 2, Spring 2001, pp. 487–495

dent variables, *may* be a serious problem. But, for typical IR problems, and specifically for the analyses presented by Green, Kim, and Yoon, 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.

We show 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 Green, Kim, and Yoon 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.[3]

Binary Time-series Cross-section and Fixed Effects

We have argued elsewhere that binary time-series cross-section data, such as 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.[4] While we have argued for grouped duration analysis,[5] dyadic conflict data can be analyzed by any event history method (of which logit is one such method, albeit flawed, 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. Green, Kim, and Yoon'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 that uses a unit dummy variable. If Green, Kim, and Yoon 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 Green, Kim, and Yoon'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 percent of Green, Kim, and Yoon's dyads—2,877 out of 3,078 dyads—never do. The inclusion of fixed effects allows for perfect prediction of almost all the dyads; as Green, Kim, and Yoon agree, this means that over 90 percent of the dyads have no impact on the statistical estimates. Thus, a data set that contained only the 7 percent of the dyads that conflict would yield *identical* estimates to the full dataset (including dyads that never conflict).

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

4. Beck, Katz, and Tucker 1998.

5. Grouped duration analysis assumes that the timing of events is only observed discretely; the conflict datasets 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.

Why do over 90 percent of the pacific dyads not affect 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 that 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 Green, Kim, and Yoon'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 percent of the subjects without the gene were cancer free, whereas only 50 percent 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 that asks whether democratic dyads are less likely to ever conflict than are nondemocratic dyads.

Now let us add some longitudinal data. Suppose we follow subjects for five years, noting each year whether or not they developed 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). Ninety percent of the cases without the gene, by assumption, never develop cancer. Thus, *using fixed effects*, these noncancerous observations make *no* contribution to the statistical analysis (that is, the likelihood). We would thus end up examining logit results based on only the 10 percent of cases without the gene but who developed cancer and the 50 percent of cases with the gene who also developed 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 that takes each subject and models the time until cancer is observed (or whether no cancer is observed after five years).[6] 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

6. For more extended discussions of the theoretical equivalence of cross-sectional duration models with time-series cross-section models, see Alt, King, and Signorino 2001; Beck 1998; and Sueyoshi 1995.

that subject, and no independent variable could possibly have any impact. Since estimating binary time-series cross-section data using 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 binary time-series cross-section model (albeit with a loss of 90 percent 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.

Green, Kim, and Yoon are not unaware of this issue. They argue 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."[7] 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 Green, Kim, and Yoon's logic, we could do no analysis, because dyadic differences might be due to differences in their own intercept (the "baseline probability") rather than differences in democracy scores. Thus Green, Kim, and Yoon'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 research can only proceed using experimental studies.

In short, Green, Kim, and Yoon's conclusion, in Table 3, that variables such as democracy have no pacific impact, is simply nonsense. It is absurd to exclude over 90 percent 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 the result of idiosyncratic factors, but that is not a position taken in any other type of empirical analysis in political science. Because binary time-series cross-section data in IR frequently contain a lot of units that show no temporal variation on the dependent variable, Green, Kim, and Yoon's proposal to include fixed effects in these analyses is *never* a good idea.[8]

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 heterogene-

7. Green, Kim, and Yoon 2001, 455.

8. This is not to say that we accept the specification in columns 1 and 3 of their Table 3. These ordinary logits do not model temporal independence correctly nor do they 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.

ity.[9] 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 through the independent variables)—might prove better. Another alternative would be a split population model, where some dyads never conflict and others might eventually come into conflict.[10] All of these offer alternative estimation methods that allow for unmodeled heterogeneity without the serious side effects of fixed-effects estimation.

Time-series Cross-section Data with Continuous Dependent Variables

The fixed-effects estimator is not quite as problematic in the continuous dependent variable case. Green, Kim, and Yoon 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 2 and 4). Although it appears that fixed effects are clearly important in the static model (column 1), this is a highly misspecified model since it incorrectly ignores dynamics. The coefficient of 0.736 on the lagged trade variable in column 3 tells us that the static model in column 1 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 1 and 3.

Thus we agree with Green, Kim, and Yoon that column 1 of Table 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, by including 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 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 3 and 4.

Comparing columns 3 and 4, we note that fixed effects explain very little additional variance. The 3,078 additional dummy variables increase the explained

9. This model, and the frailty model mentioned later, 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 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 Green, Kim, and Yoon's unmodeled heterogeneity problem, a solution that does not have the draconian consequences of Green, Kim, and Yoon's fixed effects. For a discussion of the heterogeneous Weibull model, see Greene 2000, 947. For a discussion of the frailty model, see Sargent 1998.

10. 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, whereas others will be recidivists. See Schmidt and Witte 1989.

variance from 73 percent to 77 percent. Green, Kim, and Yoon'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 Schwarz criterion (SC, also known as the Bayesian information criterion, or BIC). The SC, like other model selection criteria, judges models by their sum of squared residuals plus a penalty for lack of parsimony; the SC has a larger penalty than the common Akaike information criterion (AIC) (which is very similar to an *F*-test).[11] The SC clearly favors the dynamic model *without* fixed effects.[12] Thus on standard model selection grounds there is good reason to choose the model without fixed effects over Green, Kim, and Yoon's fixed-effects model.

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 3 of their Table 2 suffers from possible omission of fixed effects, the consequences of this omission cannot be great.

But why not include fixed effects? Why not, in other words, take the estimates in column 4 seriously? We should always be wary of statistical cures that may have serious side effects, especially when the illness being "cured" is not very serious. Green, Kim, and Yoon's fixed-effects "cure" for column 3 is akin to curing a cold with chemotherapy. Obviously, including fixed effects means that any independent variable that does not vary temporally cannot be used as an explanatory variable. Thus Green, Kim, and Yoon 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, that vary little from year are to year, are highly co-linear with the 3,078 fixed effects.[13] It is quite likely, then,

11. In terms of model selection, the AIC is equivalent to Green, Kim, and Yoon's *F*-test. Another common method of model selection, based on choosing a model with the larger $\bar{R}^2$, picks an 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 Green, Kim, and Yoon'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 SC having the largest penalty. The penalties for lack of parsimony decline with sample size, but the SC has the slowest rate of decline. (The penalty for the AIC is 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 SC is $k\log(N)/N$, which always exceeds the AIC penalty.) Applied researchers prefer the SC 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. For a discussion of model selection criteria, see Greene 2000, 306.

12. The SCs for the two models are 1.94 and 2.19.

13. 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

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 time-series cross-section 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 through a lagged dependent variable will generally make fixed effects much less relevant). Further, there clearly are models where the independent variables of interest shows year-to-year variation and so are not quite so highly co-linear with the fixed effects as in Green, Kim, and Yoon's trade model.

However, even where fixed effects are indicated, we agree with Green, Kim, and Yoon that fixed-effects models are never ideal.[14] We should clearly attempt to find substantive variables that explain dyadic differences; to simply allow for dummy variables that 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 that explains the fixed effects? One possible solution with none of the bad consequences of Green, Kim, and Yoon's fixed-effects model is the hierarchical- or random-coefficients model.[15] 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 Green, Kim, and Yoon feel is necessary (more than what the fixed-effects model allows for) without making it impossible to estimate coefficients for variables that are temporally stable. There is no doubt in our minds, however, that if one had to choose between the estimates of the dynamic trade model in column 3 and the fixed-effects model in column 4, the model *without* fixed effects is far superior for assessing the impact of variables like democracy on trade.

---

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. Green, Kim, and Yoon show the same thing with their Hausman tests. Since no one could doubt that the fixed effects radically change coefficient estimates, the Hausman test used by Green, Kim, and Yoon 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.

14. 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 binary time-series cross-section case, it is far superior to the full fixed-effects specification of Green, Kim, and Yoon. For a discussion of this approach, see Mansfield and Bronson 1997; or Beck and Tucker 1997.

15. This is a well-known model in statistics and econometrics. For a good introduction to this model in a political economy context, see Western 1998.

Conclusion

Green, Kim, and Yoon's logic 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? Green, Kim, and Yoon's position implies that only experimental study allows for any inferences, whether causal or not. They would overthrow not only much quantitative IR analysis but also every nonexperimental result ever obtained.

They do not, of course, go this far, since putting in fixed effects is clearly silly in simple cross-sectional analyses. Unfortunately, time-series cross-section 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 Green, Kim, and Yoon (and Edward Leamer and many others) that one should examine the robustness of findings to alternative specifications and methods. But to expect findings to be robust to odd specifications and or methods is a foolish expectation. While Green, Kim, and Yoon make a correct point, that *sometimes* fixed effects should be included in a time-series cross-section model (although it is probably incorrect to say they should ever be included in a binary time-series cross-section 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 Green, Kim, and Yoon 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 that allow for temporally or geographically dependent data, and 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.

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. Modeling Space and Time: The Event History Approach. In *Research Strategies in the Social Sciences*, edited by Elinor Scarbrough and Eric Tanenbaum, 191–213. Oxford: Oxford University Press.

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. Cambridge, Mass.: Harvard University. Available at <https://wwwc.cc.columbia.edu/sec/dlc/ciao/wps/tur01/>.

Donald P. Green, Soo Yeon Kim, and David H. Yoon. 2001. Dirty Pool. *International Organization* 55 (2):441–68.

Greene, William H. 2000. *Econometric Analysis*. 4th ed. Upper Saddle River, N.J.: Prentice-Hall.

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 (4):1486–97.

Schmidt, Peter, and Ann Dryden Witte. 1989. Predicting Criminal Recidivism Using "Split Population" Survival Time Models. *Journal of Econometrics* 40 (1):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 (4):1508–15.

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