Advance Access publication February 5, 2014

Political Analysis (2014) 22:457–463
doi:10.1093/pan/mpt051

Dyadic Analysis in International Relations: A Cautionary Tale

Robert S. Erikson
Department of Political Science, Columbia University, New York, NY, 10027
e-mail: rse14@columbia.edu (corresponding author)

Pablo M. Pinto
Department of Political Science, Columbia University, New York, NY, 10027
e-mail: pp2162@columbia.edu

Kelly T. Rader
Department of Political Science, Yale University, New Haven, CT, 06520
e-mail: kelly.rader@yale.edu

Edited by R. Michael Alvarez

International relations scholars frequently rely on data sets with country pairs, or dyads, as the unit of analysis. Dyadic data, with its thousands and sometimes hundreds of thousands of observations, may seem ideal for hypothesis testing. However, dyadic observations are not independent events. Failure to account for this dependence in the data dramatically understates the size of standard errors and overstates the power of hypothesis tests. We illustrate this problem by analyzing a central proposition among IR scholars, the democratic trade hypothesis, which claims that democracies seek out other democracies as trading partners. We employ randomization tests to infer the correct *p*-values associated with the trade hypotheses. Our results show that typical statistical tests for significance are severely overconfident when applied to dyadic data.

1 Introduction

In the quantitative literature on international relations, a common practice is to exploit dyadic data, where the units of analysis are dyads or pairs of nations. Typically, dyads are incorporated into panel data, with K dyads over T years, making an expanded $K \times T$ data set. This results in very large data sets, since N nations (or other base units) form $N(N-1)/2$ undirected dyads, that is, number of pairs of countries where the ordering of the pair does not matter. For instance, one hundred nations generate 4950 undirected dyads. One hundred nations over twenty years generate 99,000 undirected dyad-years, and so on.

Dyads present an ironic situation in that dyadic data sets, with 100,000 cases (or often considerably more), may seem ideal for hypothesis testing. Yet, the structure of dyadic data complicates the assessment to statistical significance. Because dyadic observations are not independent events, the usual tests of significance result in overconfidence, even when the model itself appears to be correctly specified.

The purpose of this article is to warn about the severity of the problem of overconfident standard errors in dyadic research. We proceed by applying "randomization tests" to research on the well-known hypothesis of democratic trade: Do democracies trade more with other democracies?

Authors' note: For comments on earlier drafts we are thankful to Donald Green, Soo Yeon Kim, Jeff Lax, Mark Manger, Yotam Margalit, Ken Scheve, Vera Troeger, Robert Walker, and participants at various academic presentations. Boliang Zhu provided excellent research assistance. Replication materials for this article are available from the *Political Analysis* dataverse at http://dx.doi.org/10.7910/DVN/23510UNF:5:YY9Ujh8lYTjBOtSWvyRtmw== IQSS Dataverse Network [Distributor] V1 [Version]. Supplementary materials for this article are available on the *Political Analysis* Web site.

Randomization testing is a nonparametric technique that is popular in biostatistics and increasingly in the social science literature.[1] Using randomization tests, we demonstrate that hypothesis testing with dyads using traditional methods can regularly lead to false positives, or Type I errors—claims of statistically significant findings when in fact the null hypothesis is true. The advantage of randomization inference over extant solutions is that the researcher need not rely on strong assumptions about the distribution of the errors, which is likely to be complex since the dyads are embedded, but not nested, within nations.

2 Modeling Democratic Trade

The intuition behind the democratic trade hypothesis is that democracies prefer trade partners that are democratic and other democracies would prefer to trade with them.[2] Empirically, numerous large-*N* dyadic analyses report statistical support for the democratic trade hypothesis (Dixon and Moon 1993; Bliss and Russett 1998; Mansfield et al. 2000).[3] Most famous may be the methodological symposium involving Green, Kim, and Yoon's (2001) "Dirty Pool" paper, which centered around whether to include fixed effects for dyads.

Imposing fixed effects for dyad (and year) is increasingly the norm in dyadic research (e.g., Tomz, Goldstein, and Rivers 2007). This is sensible when testing the democratic trade hypothesis; the question of whether institutional features make democracies trade more with each other is best answered with the time-series aspect of the dyad panels: controlling for the specific dyad and for common year shocks, is there more trade in years when the two countries are more democratic?

3 The Challenge of Modeling Dyads

Estimating standard errors and statistical significance with dyadic data is a challenge because each dyad-year is not an independent observation. In our example of the democratic trade hypothesis, consider the initially nondemocratic nation *A* that shifts toward democracy in year *t*. This transition gets recorded in all 100+ dyads involving nation *A* for year *t*. Should each of these dyadic changes be treated as independent events, as one event, or as something in between?

Conventional OLS standard errors assume that observations are independently and identically distributed. It is well known that violating this assumption (as dyad-years do) causes significance tests to be overconfident (see Moulton 1990). Nonindependence, or clustering, in the data reduces the "effective" number of observations (Angrist and Pischke 2009). The more clustering in the data, the fewer the effective observations, and the more standard errors should inflate.[4]

4 Randomization Tests

Randomization testing does not rely on any assumptions about the shape of the disturbances in the data. It is a takeoff on Fisher's (1935) exact test. (For a modern overview, see Edgington and Onghena 2007.) In short, this test compares the observed democracy test statistic to a distribution of false test statistics obtained when the national identities are scrambled in a series of simulations. Using randomization tests allows the researcher who employs dyadic data for hypothesis testing to overcome the inference problems that attend data with such a complicated structure.

Randomization tests do, however, rely on the assumption of exchangeability of errors. In other words, conditional on the covariates in the model, that which is randomized is simply a label that could be assigned to any observation in the data without changing the value of the dependent

[1] Randomization tests have been used to challenge the certainty of claims about such matters as gun ownership deterring crime (Helland and Tabarrok 2004), capital punishment deterring murder (Donahue and Wolfers 2005), and the effectiveness of state voting laws (Erikson, Pinto, and Rader 2010).

[2] For more on the theory regarding the democratic trade hypothesis, see Supplementary Appendix II.

[3] We find that 192 articles published in nine prominent journals in the past two decades use dyadic analysis. Among them, twenty-nine were on international trade and had democracy as an explanatory or control variable. Supplementary Appendix II presents a summary of those studies.

[4] Clustering standard errors is a common correction but of little help with dyadic data. See Supplementary Appendix III for results and an extended discussion.

variable. This is a mathematically weaker (less stringent) assumption than the *i.i.d.* assumption on which the conventional test relies. The randomization test requires that we have selected the correct model for the democracy and trade relationship. To be sure, model selection is a contentious issue in the democratic trade literature, but conventional *t*-tests are inadequate regardless of one's preferred model specification. A researcher, having settled on a preferred specification, is less constrained by assumptions when proceeding with a randomization test than with a conventional test.

Because randomization tests do not rely on a theoretical reference distribution, such as the Student's *t*, their validity does not hinge on assumptions about the shape of disturbances in the data. Instead, the data themselves are used to create an empirically derived reference distribution, custom to the data set and its particular characteristics, assuming only exchangeability of errors. We first estimate the model coefficients and their associated test statistics in the typical way. To create a reference distribution, we randomly reshuffle the country democracy scores to break the systematic relationship between the democracy score and the observed trade level. Then, we rerun the models on the shuffled data and get a new estimate of the coefficient and test statistic, knowing that the true coefficient and test statistic should be zero on average because the data were randomly scrambled. We reshuffle and re-estimate a total of one thousand times.

This process gives us one thousand simulated regression coefficients with their standard errors. Our interest centers on the distribution of one thousand test statistics—the ratios of the simulated regression coefficients to their standard errors.[5] These empirical distributions provide the reference null distributions for the randomization test. We calculate a nonparametric *p*-value by locating the observed effect (the estimated test statistic) on this distribution and measuring what proportion of the one thousand test statistics are larger in absolute value than the absolute value of the observed test statistic. This yields an estimate of the probability that the democracy coefficient could have occurred by "chance" under the null, just as a typical *t*-test does but without relying on parametric assumptions about the shape of the disturbances.

When using a randomization test on observational data, one must base inferences on a reference distribution of test statistics, and not on a distribution of regression coefficients, as is common in experimental data. With observational data, actual values of the variable of interest may be correlated with other variables in the model. In the presence of such correlation, a coefficient is not a pivotal statistic because its distribution is a function of that correlation, which acts as an unknown nuisance parameter. The test statistic, on the other hand, is pivotal because it is explicitly adjusted for the correlation among variables in the data (Kennedy and Cade 1996; O'Gorman 2005).

The way in which we randomly shuffle the democracy scores reflects the underlying structure of the series. Specifically, we randomize the time series of country-level democracy scores and make sure that, in each shuffle, each country is randomly assigned the same democracy score in every dyad in which that country appears. To do so, we first scramble the nation labels for democracy scores, as if we drop the nation labels on the floor and reattach them to democracy scores randomly. For each simulation, each nation is assigned the democracy scores for one unique nation for every dyad in which it appears. Although scores are scrambled for nation, they are not scrambled for year. Thus, the time series of country-level democracy scores are preserved. For instance, Ecuador would be assigned Japan's democracy scores for all dyads in which Ecuador appears for all years in our data, and Belgium would be assigned Costa Rica's democracy scores for all dyads and all years. After the democracy scores are randomized, the minimum democracy score within the dyad, the key independent variable, is recalculated.

By scrambling the time series of democracy scores in this way, we are making an assumption about which errors are exchangeable—or, put differently, the level at which exchangeability holds. Specifically, we are assuming that, conditional on the given model specification, the time series of democracy score values can be randomly assigned to any country without inducing systematic correlation between them and the volume of trade within the dyads including that country.

[5]The test statistic is constructed like the *t-statistic* but does not follow a *t*-distribution.

In contrast, in addition to assumptions about the shape of the errors, conventional *t*-tests rely on independence (and thus exchangeability) at the level of the dyad-year.

5 The Randomization Test Data Set

Randomization testing requires a rectangular data set with no missing data, including state births or deaths. Accordingly, we use the bilateral trade data set created by Gleditsch (2002). From this data set, we are able to identify a rectangular matrix of 2346 undirected dyads comprising sixty-nine nations with complete coverage for the 1950–2000 period to use for our randomization tests. With the data pooled, this yields an N of 119,646 dyad-years.

Following the established convention in the literature (e.g., Green, Kim, and Yoon 2001), we measure democracy as the Polity score of the least democratic of the two trading partners. Each model includes the minimum democracy score within the dyad plus a standard set of controls.[6] The level of trade between the two dyad partners is measured as inflation-adjusted 1996 dollars. We include fixed effects for both dyad and year. Two models are estimated, one with a lagged dependent variable and one without. We estimate the democracy effect via OLS the usual way. Then, we compare the test statistic with that from the randomization test, as described below.

6 Results

Table 1 shows the estimates from standard parametric *t*-tests for two models, one with and one without dynamics. In each model, the estimated democracy effects are positive and highly statistically significant. "Highly statistically significant" is an understatement, as in both cases the *p*-value based on the conventional significance test is not 0.05, not even 0.001, but at most one chance in 2.2 trillion.

The *p*-values obtained from our randomization tests are shown in the table, immediately below the conventional *p*-values. They are calculated by asking what proportion of the one thousand test statistics calculated in the random democracy score shuffles is larger in absolute value than the absolute value of the observed test statistic from the unshuffled data. By this test, the estimates of the democracy effect are statistically significant, that is, outside the realm expected by chance from random assignment of country democracy profiles. But ominously, the conventional *t*-test used in dyadic analyses produces *p*-values that are too small. The randomization test *p*-values are 10 *billion* times larger than the parametric *p*-values with the dynamic model and 36 *sextillion* times larger without dynamics. Thus, conventional significance tests dramatically overstate the confidence of results obtained with dyad-year data.[7]

We can use the one thousand randomization test runs to illustrate just how likely a Type I error is in data with this structure. In each run, we recorded the parametrically derived *p*-values associated with the minimum democracy score coefficients calculated using the conventional *t*-test on the scrambled data. Because the time series of minimum democracy score was randomly shuffled, we know that it has no meaningful association with trade within the dyad. Thus, if the conventional test is appropriate for these data, we should (incorrectly) reject the null hypothesis only 5% of the time at the 95% confidence level, 10% of the time at the 90% confidence level, and so on. That is, the distribution of *p*-values we obtain should be uniform between 0 and 1.

Figure 1 shows the distributions of the one thousand conventional *p*-values calculated during the randomization tests. The shaded areas cover *p*-values that are 0.1 or smaller, small enough to reject the null hypothesis at the 0.10 confidence level. What we find is quite different from the uniform distributions we would expect if the conventional test were appropriate. Even though these *p*-values were calculated using data in which no systematic relationship exists between the minimum democracy score within the dyad and trade between the two countries, 60% or 84% of the conventional *p*-values were less than or equal to 0.1. In other words, using conventional significance tests, this

[6] For a data description and sources, see Supplementary Appendix I.

[7] The quantity 36 sextillion is greater than the number of grains of sand on Earth. The overconfidence is even more extreme when modeling without fixed effects. And clustering the standard errors by dyads offers little improvement. See Supplementary Appendix III.

Table 1 Regression analysis of bilateral trade, 1950–2000

	Dyad and year fixed effects	*Dyad and year fixed effects with dynamics*
GDP	0.918 (0.007) *p=.000*	0.224 (0.005) *p=.000*
Population	0.252 (0.012) *p=.000*	0.103 (0.007) *p=.006*
Alliance	0.384 (0.024) *p=.000*	0.064 (0.015) *p=.000*
Democracy	0.013 (0.001) $p=1.0 \times 10^{-25}$ *rand p=.036*	0.003 (0.000) $p=2.2 \times 10^{-12}$ *rand p=0.022*
Lagged bilateral trade		0.791 (0.002) *p=.000*
Constant	−17.312 (0.259) *p=.000*	−5.161 (0.178) *P=.000*
R^2	*N*=2,346 *T*=51 0.87	*N*=2,346 *T*=50 0.95

Note. Democracy is the lower value within the dyad. GDP and bilateral trade are in real 1996 dollars and are natural log transformed. Population is natural log transformed. Except for those on Democracy, all $p < 0.0009$ are rounded to zero.

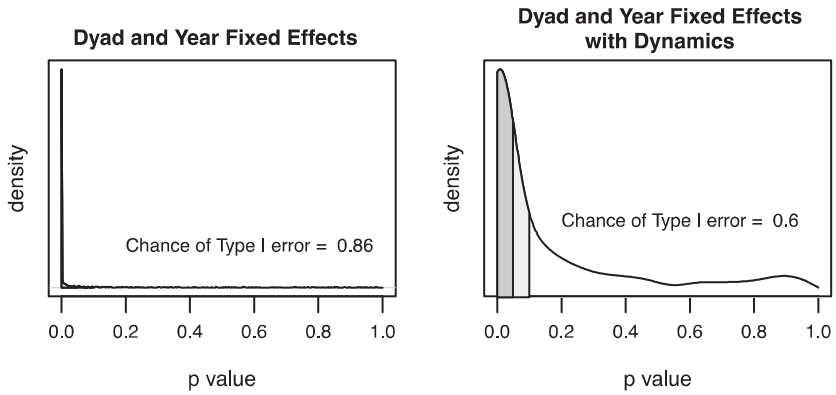

Fig. 1 Rate of false positives in conventional *T*-test. This figure shows the densities of the one thousand conventional *p*-values calculated during the randomization tests for the models in Table 1. The shaded areas cover *p*-values that are 0.1 or smaller. The dark gray shaded areas cover *p*-values of 0.05 or smaller.

dyadic data would cause one to falsely infer a significant effect most of the time instead of 10% as we would expect at the 90% confidence level.

Of course, the fact that the parametric tests have unacceptably high Type I error rates does not necessarily mean that there is no real democracy effect on trade. Indeed, despite the marked overconfidence of the standard hypothesis tests, the coefficients on the minimum democracy score remain statistically significant at conventional levels according to our randomization tests.

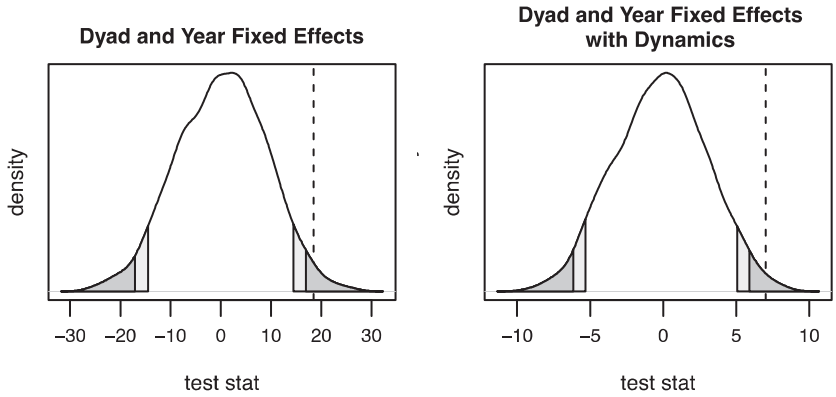

Fig. 2 Randomization test results. This figure shows the densities of the one thousand test statistics on the democracy variables estimated using the randomization test. The dark gray shaded areas represent the 5% most extreme test statistics, and the entire shaded areas cover the 10% most extreme test statistics. The dotted lines indicate the magnitudes of the test statistics estimated using the observed data, the ratio of the coefficients and standard errors in Table 1.

The coefficient on the minimum democracy score, though substantively small, is significant according to randomization testing, both without dynamics ($p = 0.036$) and including dynamics ($p = 0.022$).

Figure 2 graphs the randomization test results. The graphs show the densities of the one thousand test statistics on the democracy variables when each country's democracy score over time is from a random shuffle. The dark gray shaded areas represent the 5% most extreme test statistics, and the entire shaded areas cover the 10% most extreme test statistics. The dotted lines indicate the test statistics estimated using the observed data, the ratio of the coefficients and standard errors in Table 1. For each model, the estimated democracy effect is sufficiently large to be outside the 0.05 bound.

To summarize, the randomization results show that conventional *t*-tests are inappropriate for testing hypotheses on dyad-year data because they rely on highly overconfident standard errors. Although the randomization tests do provide statistical support for the democratic trade hypothesis in this case, researchers seeking to test hypotheses using dyad-years cannot ignore the complex error structure in the data without incurring a high risk of a false positive.

Consider the case where the independent variables of interest (such as domestic democracy scores) rarely change and in truth have no effects on the dyadic relationships of interest. In such circumstances, and even with perfect model specification, the null hypothesis (in this case true) has a strong likelihood of rejection with conventional hypothesis testing apparatus. The danger then is building a science in an attempt to understand a cascade of findings that in fact are nothing more than false positives. Randomization tests provide one antidote to that possibility by making correct inference with dyadic data possible, even though the structure of dyadic data is quite complex.

7 Discussion and Conclusion

We have seen the perils of analyzing large-*N* dyad-year relationships, specifically in the context of the democratic trade hypothesis. Conventional tests of significance, we show, are wildly overoptimistic with dyad-year data. Each time we scramble the national identities of the democracy scores in our randomization analysis and generate a nonsense coefficient for democracy, it is more likely than not to be "statistically significant." The implication is that if significance tests were evaluated heedlessly in dyadic research, causal relationships would often be deemed to be "significant" even when the null hypothesis is correct. Moreover, when the data suggest a substantively plausible relationship, as found with democracy and trade, the calculated significance level can be off by factors of trillions, leading to extreme overconfidence in the results.

The traditional tests of significance are wrong because their usual underlying assumptions are not true in the case of dyad-years. Unmeasured causes of trade between dyad partners tend to persist from year to year, defying the assumption that errors are independent over time. When a nation undergoes a pro-democratic revolution or, alternatively, when democratic leaders are deposed in a coup, the change ripples through all the nation's many dyads, artificially inflating the number of relevant cases. Further, there is not as much variance in the minimum democracy score within the dyad. Most nations (and their trade partners) tend to maintain the same democracy score from year to year.

What is to be done? One constructive solution for dyadic analysis would be to perform randomization testing of the sort presented here. The requirements are a rectangular data set and the ability to randomize data and perform multiple simulations. In some cases, such as degree of democracy, the independent variable is an attribute of the nation rather than a relational dyad-level variable. In these instances, it is permissible to shift the unit analysis from the dyad-year to the nation-year. For instance, one can ask whether countries trade more with democratic countries when their home institutions are at their most democratic. The analysis would be free of the problems noted here that are specific to dyadic analysis.

Our meta-lesson is that it is possible to be lulled into falsely supporting a research hypothesis by the force of a large *N*, which, in an imperfect nonexperimental setting, is prone to false positives. When the model is wrong, citing significance levels based on the reported standard errors in computer output only offers the illusion of success.

References

Angrist, Joshua D., and Jorn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Bliss, Harry, and Bruce Russett. 1998. Democratic trading partners: The liberal connection, 1962–1989. *Journal of Politics* 60(4):1126–47.

Dixon, William J., and Bruce E. Moon. 1993. Political similarity and American foreign trade patterns. *Political Research Quarterly* 46(1):5–25.

Donahue, John, and Justin Wolfers. 2005. Uses and abuses of empirical evidence in the death penalty debate. *Stanford Law Review* 58:791–845.

Edgington, Eugene S., and Patrick Onghena. 2007. *Randomization tests*. 4th ed. Boca Raton, FL: Taylor and Francis Group.

Erikson, Robert S., Pablo M. Pinto, and Kelly T. Rader. 2010. Randomization tests and multi-level data in U.S. state politics. *State Politics and Policy Quarterly* 10:180–98.

Fisher, R. A. 1935. *The design of experiments*. Edinburgh: Oliver and Boyd.

Gleditsch, Kristian S. 2002. Expanded trade and GDP data. *Journal of Conflict Resolution* 46:712–24.

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

Helland, Eric, and Alexander Tabarrok. 2004. Using Placebo Laws to test "more guns, less crime." *Advances in Economic Analysis and Policy* 4:1–7.

Kennedy, Peter E., and Brian S. Cade. 1996. Randomization tests for multiple regression. *Communications in Statistics-Simulation and Computation* 25:923–29.

Mansfield, Edward D., Helen V. Milner, and B. Peter Rosendorff. 2000. Free to trade: Democracies, autocracies, and international trade. *American Political Science Review* 94(2):305–21.

Moulton, Brent R. 1990. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *Review of Economics and Statistics* 72(2):334–38.

O'Gorman, Thomas W. 2005. The performance of randomization tests that use permutations of independent variables. *Communications in Statistics Simulation and Computation* 34:895–908.

Tomz, Michael, Judith Goldstein, and Douglas Rivers. 2007. Do we really know that the WTO increases trade? Comment. *American Economic Review* 97:2005–18.