Best of All Plausible Worlds? Checking Robustness of Time-Series Cross-Sectional Models with Fictitious Plausible Alternate Treatments

Evangeline Reynolds
University of Illinois
Department of Political Science
David Kinley Hall
Champaign, IL 61820

Email – ereynol4@illiois.edu

Keywords: IGOs, Non-Parametric Methods, Time-Series Cross-Sectional Analysis, Permutation, Placebo

March 30, 2014

Contents

1 Abstract

This paper describes a new robustness framework for observational panel data analyses, motivated by the potential outcomes framework. Based on knowledge of the data generating process for the key explanatory variable, a set of plausible but fictitious alternate treatments is created. Estimating models with many fictitious data vectors in addition to the true independent variable yields a distribution of test statistics which could have resulted under the plausible alternate treatments. The researcher adjusts her confidence in rejecting the null according to the true treatment test statistic's position within this distribution. This approach does not afford us with a one-size-fits all routine, therefore I demonstrate the method using three examples from the literature.

2 Rationale

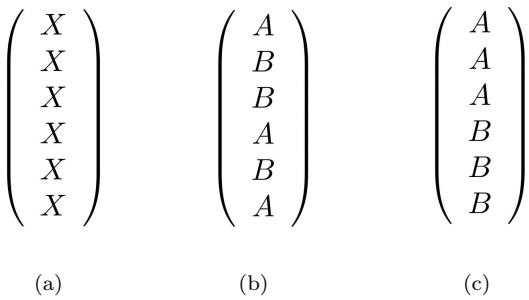
Political Science has faced great debates about the appropriateness of various modeling strategies in panel data regarding a variety of subjects such as pooling, (Beck & Katz 1995, Green, Kim & Yoon 2001, Beck & Katz 2001, King 2001, Oneal & Russett 2001), lagged dependent variables (Achen 2000, Wawro 2002, Kristensen & Wawro 2003), and other dynamic modeling decisions (Wilson & Butler 2007). This is due to the probably frequent violation of modeling assumptions which may lead to overconfidence in model estimates. Ultimately, it is not unlikely that a researcher chose a modeling strategy which may violate modeling assumptions. Concerns about misspecification is appropriate because it may lead to biased estimates of regression coefficients. What's more, scholars are increasingly concerned about reverse causality, selection bias, and endogeneity (Von Stein 2005).

What might assuage concerns about the violation of modeling assumptions and endogeneity is the approach that I present in this paper. My approach embraces the marrying parametric and non-parametric approaches. The parametric approach has convenient characteristics and well understood properties. Multivariate regression allows variables other than the explanatory variable of interest to soak up variance, addressing concerns about missing variable bias in the parameter of interest. Yet, non-parametric approaches do not depend on distributional assumptions to determine statistical significance.

A more simple non-parametric robustness check of a parametric model might to construct a reference distribution from all possible reassignments of the explanatory variable of interest or of the error term (Erikson, Pinto & Rader 2010, Erikson, Pinto & Rader 2014). However, I argue that constraining the set of allowable comparisons to create a reference distribution may be more appropriate, especially in the case of that the researcher is concerned with causal relationships and not just statistically significant relationships in an observational study.

In fact, the intuition for constraining comparisons in the observational setting might be made more concrete by thinking about experimental setting. First, take the rationale from Bailey's "Randomization, Constrained" (1986) for disallowing some comparisons in constructing a reference distribution after the results from a random experiment are collected. Bailey notes that some treatment configurations for an agricultural experiment – even though possible under complete randomization – would not be implemented by a researcher. For example, consider if three different treatments are to be administered to a field broken down into six plots, as shown in Figure 1(a). The treatments are A and B and half the plots will receive each treatment. Figures 1(b) and 1(c) contrast in that if a randomization procedure were applied the first would be accepted but the second would not. Spatially, there are too likely to be confounder for the outcomes. At the outset, Bailey argues, the researcher should think about which are the randomizations that he would never allow. The set of randomizations that would not be implemented should not be included in the reference distribution created by the permutation of the treatments, since there was not truly a chance that it would be implemented. The logic for constraining comparisons in observational studies is similar in that, I argue, we should limit the comparisons the relevant comparisons, i.e. treatment assignments that are plausible.[1]

[1]For observational studies, contrasting with Bailey's example, it not is likely that the most random looking assignments will necessarily be plausible ones. Rather treatment assignments might be a function of a covariate.



(a) (b) (c)

Figure 1: The above represent (a) plots of land to receive random assignment, (b) acceptable randomization, and (c) a randomization that would be rejected by the researcher.

Another analogy from the randomized experiment experience provides intuition for constraining comparisons in observational studies. Specifically consider a case where treatment assignment is not equally likely among groups. Consider a treatment that was disproportionately administered to a vulnerable group for policy reasons, given that the treatment was costly, there is greater justification to provide it to an economically vulnerable group. Within the vulnerable group and the other group treatment assignment was random. If this information is ignored, incorrect inferences would be made about the effect of the treatment. For example, if there is no true effect, an analysis with that does not account for the data generating process for the treatment could conclude that the treatment actually leads to worse economic outcomes. If the researcher creates a potential outcomes reference distribution fulling randomizing the treatment then the researcher might conclude that there is a treatment effect when none exists. However, if the potential outcomes distribution is constructed taking the treatment assignment into account, then the false positive is not likely to result.

The risks associated with 'allowing' comparisons (implicitly using econometric tools or explicitly when the nonparametric analysis allows for it) in the reference distribution which are not possible or very implausible are also present in observational studies. As in experimental studies, it may be appropriate to disallow treatment comparisons that are implausible, especially when looking for evidence of a causal relationships and when modeling data structures where assumptions are harder to meet.

3 Method

The approach explained within builds on the econometric approach that scholars already take with parametric models. First scholars make theorize about the world and make hypotheses about relationships in observational data based on their hypotheses. Then they collect data with the aim of testing these hypotheses. Then begins the modeling process. Scholars, based on intuition and statistical tests settle on a preferred specification, possibly using an alternate specification or an alternate operationalization of a variable as a test of robustness.

My approach advocates doing this much but then going further in thinking about how the treatment was assigned. I ask the question, "What statistics and estimates would this modeling specification have yielded had the treatment been different?" It is the question of the potential outcomes framework. We do not know the exact data generating process in observational studies as we would with experimental studies. However, this is not an excuse to *ignore* the data generating process. Researchers know that the distribution of their treatment is not random. At the outset, before intimately getting to know the data, there are some vectors of data that the researcher would reject as implausibly being the explanatory variable of interest, i.e. the treatment data. For example, this would be data that, if the true data were replaced unbeknownst to the researcher, the researcher might go back, looking for an error in the data source or putting the data set together. Yet there will be sets of vectors that the researcher would not reject as a the treatment vector

were the true data surreptitiously replaced. While these were not actual realizations of the treatment, these might be considered plausible treatments.[2]

So my approach is to *create* a set of these plausible treatment assignments, with albeit limited knowledge about the data generating process for the treatment of interest. The appropriate method for doing this might vary according to the explanatory variable at hand. I will go into greater detail about specific methods a subsequent section. Using the set of plausible treatment, the researcher would then take his preferred modeling specification and replace the realized treatment with plausible treatments one at a time, collecting the model results.

This method models the outcomes using these plausible treatment assignments instead of the realized treatment create a set of relevant comparisons for the model performed with the realized treatment. The test statistics associated with the alternate treatments can be collected and used as a reference distribution for the test statistic associated with the realized treatment. If the test statistic is outlying in this distribution, then the researcher will be more confident of a true causal effect of the treatment. Conversely, if the test statistic is not outlying in this reference distribution, the analysis casts doubt on the strength of the evidence for a causal relationship being identified in the panel data analysis.

Steps for Constrained Comparisons Analysis

1. Collect data for testing hypothesis.
2. Based on econometric literature, choose and implement appropriate model for evaluating if there is a statistically significant relationship between the outcome of interest and explanatory variable of interest.
3. Consider the data generating process. (In its simplest form, this might even take into account researcher's coding scheme for the)
4. Mimic the data generating process - obviously, the data will diverge from the actual data. Randomization will be one way to introduce divergence.
5. Use the data generated through mimicry to consider how unusual the result is holding everything else constant.
6. Compare the statistical significance of the true estimate to the set of test statistic generated by plausible treatments.

4 Specific Methods Overview

I demonstrate three ways of generating plausible comparisons. They are constrained permutation, shifting interventions, and modeling interventions and shuffling the error.

4.1 Constrained Permutation

One method to constrain comparisons to a relevant set is via a constrained permutations test. This method has parallels to the panel permutation tests method recently explained in Erikson et al. (2014). In Erikson et al., a full permutation test in the TSCS context uses randomization to break the relationship between the dependent variable and independent variable, but keeps the time-series structure in tact. Then rerunning the model with many permutations of the independent variable, the researcher can compare the actual test statistic to the test statistic that would have been generated if the same time series for the explanatory variable had been assigned at random to the cross sectional units; this can highlight unacceptably high false positive (type I error) rates, and the research may adjust her confidence in a true relationship. I

[2]It should be noted that these ideas are not restricted to the bivariate case. Even though the simplest case of treatment and control are indicators 1 and 0, continuous treatments might have plausible and implausible assignments.

propose using permutation of the independent variable of interest, as they also advocate, except that certain permutations would not be allowed. In some circumstances, knowledge of the data generating process will make it appropriate to constrain the possible permutations – allowing reassignment of cross-sectional unit (country) data to *only* a subset of cross-sectional units.

In general, permutation tests offer an important step toward jointly exploiting parametric and nonparametric tests. (Kennedy 1995, Kennedy & Cade 1996) These tests have strong intuitive appeal as a check for over confidence in model estimates. The procedure involves breaking the relationship between the dependent variable and independent variable which may expose the propensity of the modeling strategy towards false positives due to violations of the modeling strategy. The procedure draws strongly from the intuitive logic of Fisher's Exact test (a non-parametric test). Fisher suggested that when trying to determine if the means of two groups are different, instead of making assumptions about the normality of data distribution, the data could be randomly reassigned to the two groups; the differences in means of the permutations of data could be compared to the actual difference in means. If all of the possible permutations were made, then it could be *exactly* known how likely or unlikely the *observed* difference in means is given the overall distribution of the data, as opposed to estimated via application of normality assumptions. The permutation procedure extends Fisher's exact test, using the logic of breaking the relationship between treatment and outcome, it but compares results of multivariate regressions' test statistics rather than simple comparisons of means. (Kennedy & Cade 1996). In so doing, researchers make the assumption of *exchangeability* which researchers justify (Erikson, Pinto & Rader 2010, Erikson, Pinto & Rader 2014) because at least this is a weaker assumption than the usual parametric modeling assumption that errors are independently and identically distributed (IID):

> Exchangeability means that if the null hypothesis is true, if the variable of interest indeed has no effect, then observed outcomes across individuals would be similar (conditional on confounding covariates) no matter the level of the variable of interest. In other words, if exchangeability holds, then under the null hypothesis, the variable of interest is merely a label that can be applied to any observation without changing the expected outcome (Erikson, Pinto & Rader 2010).

Permutation of the data is especially useful in the time-series cross sectional data setting where analysis is plagued with problems, such as serial auto-correlation, in exposing the severity of these problems (Bertrand, Duflo & Mullainathan 2004). In this paper, I build upon the permutations test strategy in the time-series cross-sectional data, but argue for constraining reassignments. In Erikson, Pinto, and Rader (2014), the authors advocate for creating a new reference distribution of test-statistics after breaking the relationship between the explanatory variable of interest, Z (level of democratization), and the outcome variable, y (level of trade) in the TSCS context. Taking into considerations the advise of Kennedy and Cade (1995, 1996), their random reassignment is of Z on the units of analysis, country dyads. For the time-series cross-sectional setting they argue that the cross-sectional units (country dyads) should be taken as the principle observation. Therefore their permutation procedure involves repeated random reassignment of the *entire* dyad time-series of the theoretically key independent variable (level of democracy), to any other dyad. So the entire time-series for the independent variable should be taken as a treatment that may or may not affect outcomes. This approach is appropriate given likely unmodeled dependencies in the data, and consistent with the urgings of scholars to take the cross sectional unit as the primary observations (Wilson & Butler 2007).

4.2 Model the explanatory variable of interest

Another approach to constrain the comparisons made is to model the explanatory variable of interest using an important explanatory variable and use the remaining unexplained variance to create comparable data, randomizing its assignment, but using the predicted values as a baseline. I implement this in a simple case in which the explanatory variable of interest is an intervention and much of the variation in the timing of the intervention can be explained by country wealth. I model the intervention year with wealth. The strength of the this relationship is statistically significant, yet there is substantial remaining variance usable for creating a set of comparisons. The remaining variance (the residuals) are reassigned to create a plausible

set of comparisons. The comparisons bear some resemblance to the actual data in that much of the variance in each of the generated data can be explained by wealth.

4.3 Temporally shifting the explanatory variable of interest independent

Finally, I use a set of comparisons generated directly from the explanatory variable. In the case of an explanatory variable that is an intervention, coded as a dummy variable, flipping the zero to a one a few years prior to the actual intervention serves as a good comparison (see Dube, Kaplan & Naidu 2011 and Luechinger and Moser 2012). Here, anticipation of the intervention might be of concern, but if the researcher specifies in advance and provide a rationale for the number of years he or she might plausibly observe anticipatory behavior, and willing to attribute the observed effect to the intervention of interest. This approach has been used by Dube, Kaplan and Naidu (2011) and Luechinger and Moser (2012) use a placebo variable generated *directly* from the intervention of interest itself, shifting the intervention a fixed number of years earlier for each cross sectional unit.

5 Implementation

Table 1: Data and Model Specification Sources

Source	Cross-Sectional Unit	Outcome	Treatment Variable	Model Specification Tested	Constraint Justification
Neumayer 2005	Country	Human Rights Outcomes	Convention Against Torture	Country Fixed Effects Lagged Dependent Variable Robust Standard Errors	Intervention Timing is Predicted by Country Wealth
D'Souza 2012	Directed Dyads	Exports	OECDAnti-Bribery Convention	Directed-Dyad Fixed Effects, Year Dummy Variables,Time Trend Control,Robust Clustered Standard Errors	Participant Countries are Wealthiest Countries
Greenhill 2010	Country	Human Rights	Human Rights Context	Ordered Probit Model with Lagged Dependent Variable	Region Predicts Assignment

I implement this technique using existing time-series cross-sectional models from the literature. Two of my sources model human rights outcomes, with explanatory variables of interest being membership in international human rights treaties (Neumayer 2005) and networks of international organizations (Greenhill 2010). A third data/model source models international trade, with economic treaty membership interacted with a domestic economic variable as the explanatory variable of interest (D'Souza 2012). I am grateful to the authors for providing their data by posting online or providing it upon request.

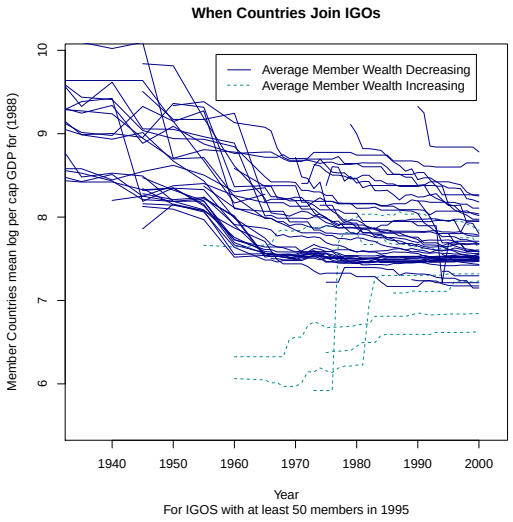


Figure 2: Data demonstrates that IGO membership usually begins with wealthier countries.

5.1 The Effect of Human Rights Treaties

Neumayer explores the conditional effect of human rights convention on human rights outcomes. He hypothesizes that human rights outcomes will improve in the presence of these conventions in countries where the number of international non-governmental organizations is relatively high per capita and for countries with strong democracies. These are proxies for civil society. He interacts the conventions with the country's polity score as well as with the number of NGOs per capita. He models various a human rights outcome and a civil rights outcomes in his paper as a function of various human rights treaties. Neumayer estimates the effects using OLS country fixed effects model as well as an ordered probit model without fixed effects, both with robust standard errors, justifying these modeling decisions in his paper.

A bare-bones non-parametric test of this model might be to randomly assign intervention years to countries. However the further constraint for which I advocate derives from the fact that a general pattern for timing of IGO membership is related to wealth. Figure 2 suggests the pattern of membership accession for large IGOs begins typically with countries with larger GDPs, joined by countries with smaller GDPs. Here I used the "IGOs 2.1" data compiled by Pevehouse and Nordstrom. Such data might suggests that not all patterns of membership accession are equally likely.

To create plausible alternate data to run in Neumayer's models, I model the year of signature to the Convention Against Torture as a function of country wealth - per capita GDP in 1995. This is simply a linear OLS model. The data is censored and countries that do not sign the Convention Against Torture during dataset time range are assigned a membership year of 2003 a year after the completion of the dataset.[3] In the case of the Convention Against Torture wealthier countries join first. For the countries that sign onto the CAT, the correlation for the year of membership and the year is -0.277.[4] The model and residuals are recorded.

With the model and the collected residuals I generate a plausible alternate treatment assignments. Specifically, I take the fitted values for the model of the CAT membership year, then reassign residuals randomly. I merge this data onto the Neumayer dataset and create a dummy variable coded one (1) if the country joined the fictitious convention and zero (0) otherwise. I run the model with this data.

This process is repeated 1000 times collecting the estimates, test statistic and p-values associated with each run of the model with a different plausible treatment. I do this for both modeling specifications presented

[3] In future work, I would like to have a look how sensitive the results are to decisions like this one

[4] This is when the countries that do not sign on over the time period of Neumayer's data are assigned a join year of 2003.

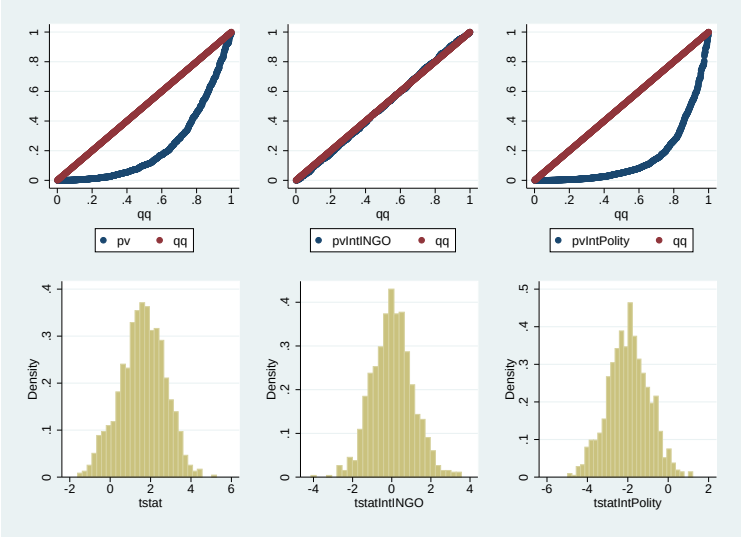


Figure 3: Test statistics and p-values generated by simulated conventions - following wealth predicted joining - with OLS fixed effects specification. The values 2.18, -2.82, -2.74 are the respective test statistics for the intervention, the intervention interacted with INGOs, and the intervention interacted with polity.

in Neumayer Figures 3 and 4.

The researcher might be surprised to find that many of these not closely related interventions produce quite large test statistics for the direct effect coefficient and one of the interacted effect coefficient (that is the interaction with Polity), as is the case for many of the real conventions that Neumayer tests. The case of INGOs shows good distribution for pvalues and test statistics when shuffling. However, the distribution of INGOs per capita for countries is very right skewed. It might be appropriate to take the log of the international nongovernmental organizations.

Table 2: Neumayer Constrained Comparison adjusted P-values, unadjusted in parentheses

	Neumayer Fixed Effects	Ordered Probit	Revised Fixed Effects
Intervention	.313	.530	.258
	(0.029*)	(0.029*)	(0.061*)
Interaction with INGO	.002*	.422	.088
	(0.005*)	(0.317)	(0.111)
Interaction with Polity	.222	.482	.191
	(0.006*)	(0.066)	(0.022*)

I do estimate a slightly different specification than Neumayer's Fixed Effects specification. Neumayer perhaps incorrectly uses the "absorb" option in STATA, which does not properly reduce the degrees of freedom which the dummy variables contribute. Estimating the correct fixed effects model, the test statistics are 1.89, -1.61, and -2.31 for the intervention, the interaction with INGOs per capita, and the interaction with Polity respectively.

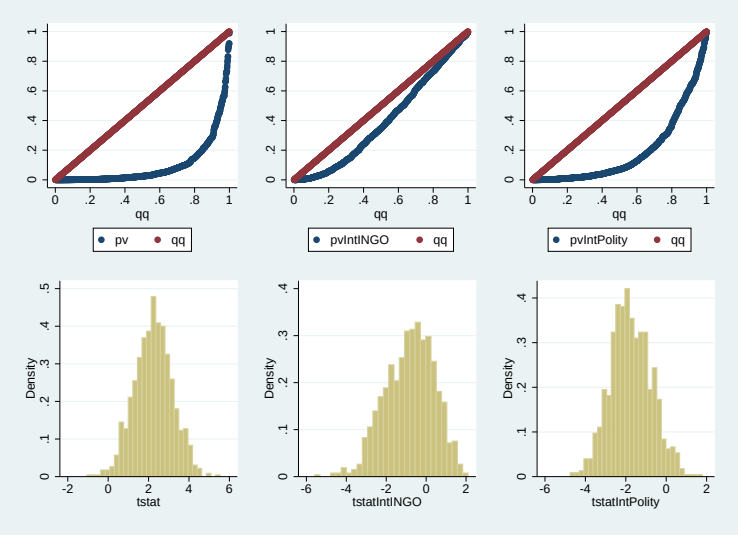


Figure 4: Test statistics and p-values generated by simulated conventions - following wealth predicted joining - with Ordered Probit specification. The values 2.19, -1.00, and -1.84 are the test statistics for the intervention, the intervention interacted with INGOs, and the intervention interacted with polity.

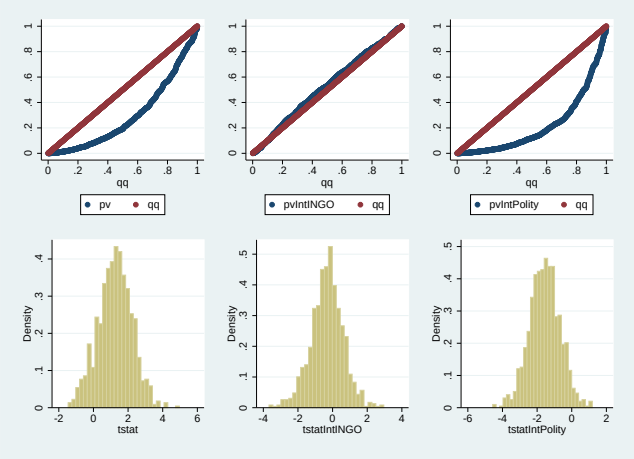
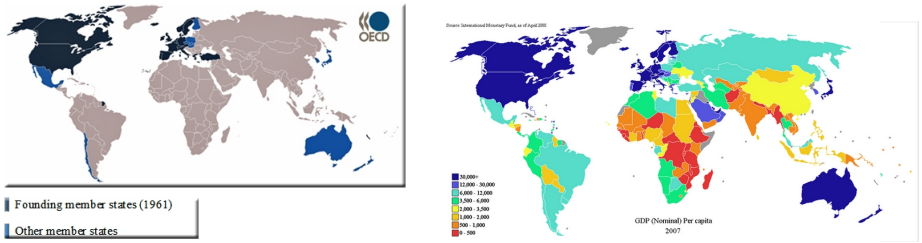


Figure 5: Test statistics and p-values generated by simulated conventions - following wealth predicted joining - with my own OLS fixed effects specification. The values 1.89, -1.61, and -2.31 are the respective test statistics for the intervention, the intervention interacted with INGOs, and the intervention interacted with polity.



(a) (b)

Figure 6: Subfigure a) shows which are the members of the OECD and subfigure b) shows per capita GDP in 2007. Countries that are not members of the OECD but which signed the convention anyway include Argentina, Brazil, Bulgaria, Colombia, Russia and South Africa.

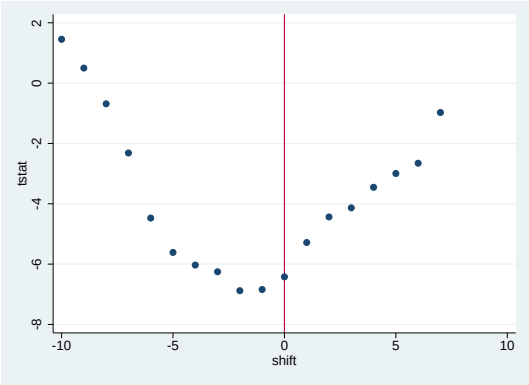


Figure 7: Test statistics generated by intervention shifted by the specified number of years.

5.2 Anti-Bribery Convention's Effect on Aggregate Trade Flow

D'Souza models exports from one country to another as a function of membership to the OECD Anti-Bribery Convention.[5] She also interacts this with a measure of corruption for the importing country. She anticipates that the country that is a member of this convention will suffer export losses to countries that are more corrupt.

Yet the treatment assignment is far from random, as shown in Figure 6. The set of countries treated are the wealthiest in the world. This raises the question, would other similar treatment assignments - but not this specific OECD convention assignment - also produce similar statistically significant results? I employ two strategies for creating comparable treatments. First I shift the true intervention year all by the same interval. Then I create the interaction variable, multiplying on the corruption level for the trade partner with the shifted vector, and estimate the model with the new 'treatment' and interacted 'treatment'. A summary of the results for the interaction (the vector which interests D'Souza) are show in Figures 7 and 8. These are the test statistics for the interaction term in model 2 of her paper.

In addition I use a constrained permutation to reassign the treatment to create comparable treatments. Specifically, I take the true years of intervention that D'Souza uses, the year of signature to the Convention, but shuffle them only among the countries that sign the Convention. Figure 9 shows that twenty-seven percent of the time, the shuffles yield a test statistic greater in absolute value than DSouza's test statistic.[6] The actual test statistic is -6.42.

[5]This is the OECD Convention on Combating Bribery of Foreign Public Officials in International Business Transactions.
[6]I do not shuffle the U.S. consistent with D'Souza's treatment of it as a special case.

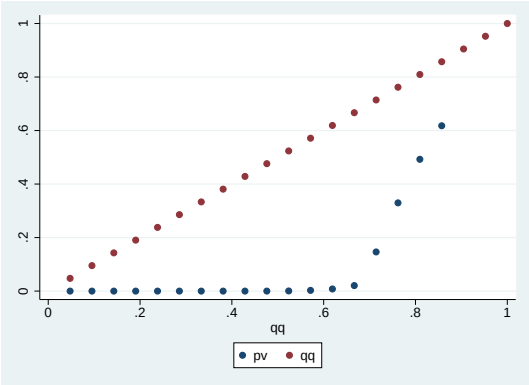


Figure 8: The qq plot for the p-values associated with the intervention shifts.

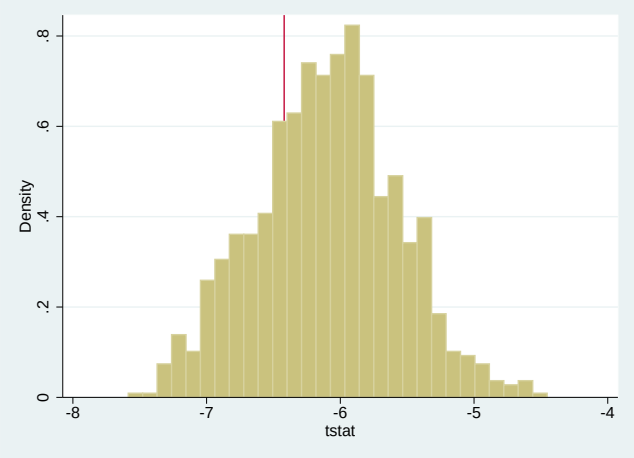


Figure 9: Distribution of test statistics for 1000 constrained permutations

Table 3: Intervention Year for D'Souza Data

ARG	2000	AUS	2000	AUT	1999	BGR	1999
BRA	2002	CAN	1999	CHE	2000	CHL	2003
CZE	1999	DEU	1999	DNK	2000	ESP	2000
EST	2004	FIN	1999	FRA	2000	GBR	2002
GRC	1999	HUN	1999	IRL	2002	ISL	1999
ITA	2001	JPN	1999	KOR	1999	MEX	1999
NLD	2001	NOR	1999	NZL	2001	POL	2001
PRT	2001	SVK	2000	SVN	1999	SWE	1999
TUR	2003	USA	1977*				

5.3 IGO Common Membership's Effect on Human Rights

I apply the constrained permutation approach to Brian Greenhill's (2010) analysis.[7] He hypothesizes that socialization via IGOs will allow for diffusion of human rights norms, ultimately affecting human rights behavior. In his study, geographic explanations for the human rights outcome is treated as a potentially confounding variable for measuring the relationship of interest between a country's IGO networks' human rights record and its own human rights record. Greenhill's use of control variables leads to results which supports his IGO/human rights hypothesis. However, using a constrained permutation test, in which the time-series data of the theoretically key independent variable (IGO human rights context) is allowed to be reassigned within its own region, leaves the finding in doubt.

To apply the permutation procedure in this case, Erikson, Pinto and Rader (2014) would think of the cross-sectional units as Country A, Country B, Country C, Country D, etc. and these country labels can simply be exchanged for the explanatory variable of interest. The exchange of labels is done many (typically 1000) times, and then the model is run with each permutation of variable Z, and with the set of test statistics for Z a new reference distribution is generated.

In contrast to their work, I argue that confounding variables might be used to productively limit complete randomization at the country level to a set of relevant comparisons. The exchangeability assumption, in the case of Greenhill's explanatory variable, is more likely to be true when constrained. Given this, I allow the region to determine what I consider exchangeable units. The "labels" are still the letters (or the names of the country) but the region of the country is recognized. So only European Country A and European Country C could be reassigned to one another, but neither could be assigned to Latin American Country B and South Asia Country D. That is, reassignment of the data would only be allowed within the regional group. If we truly think regions might strongly predict an outcome, and if we observe outcomes for the constrained procedure that look much like that observed outcome for the actual data, then, I argue, it is difficult to reject the null hypothesis of no effect with confidence.

Greenhill (2010) uses time-series cross-sectional data to explores his hypothesis that states may change their internal behavior because of their external relationships. His dependent variable is *Physical Integrity Rights Score*, which is created by summing up the Cingranelli and Richards' scores for each of four types of human rights violations which result in cumulative scores of 0 to 8 (zero being most egregious level of violations) for each country and in each year in the dataset. The key explanatory variable is the *IGO Human Rights Context*, which, for each country, is created from the *Physical Integrity Rights Score* of the other countries in its IGO network. This variable is created in two steps. First, for each IGO in which a country is a member and for a given year, the human rights score of the remaining members is averaged. Second, the average is taken of all the results from the first step to yield the *IGO Human Rights Context* for that country in that year; only IGOs in which the country is a member are included, so this aggregation procedure yields a measure of the human rights environment created by the IGO cohort. For the additional explanatory variables included in his models (control variables) the author provides a thorough rationale for why each was included and discusses the sources and scaling of the data (see pages 133-137).

Greenhill's data encompasses 137 countries, and the years from 1980 to 2004. The models are ordered probit regressions with robust clustered standard errors by country. In each of the five regressions, the author lags the dependent variable one year. For all the other explanatory variables including the variable of interest, IGO human rights context, the lag is indicated by the column label for the model. The rationale for his modeling decisions are supported by citations from methodological literature.For each of the five models, the IGO Human Rights Context variable is positive and statistically significant. It also appears that the three year lag of the IGO context variable best explains the domestic human rights outcome (although the estimates are not statistically different from each other). Based on these results, a the researcher would naturally reject the null hypothesis of no effect for the IGO Human Rights Context variable.

My motivation for taking a second look at Greenhill's results was the hunch that region might be driving the result. In his models, Greenhill recognizes the potential for proximate states to confound the analysis, including the neighborhood effect variable. This is an average of the *Physical Integrety Rights Score* for all of

[7]The piece is titled "The Company You Keep: International Socialization and the Diffusion of Human Rights Norms" recently published in *International Studies Quarterly* (2010).

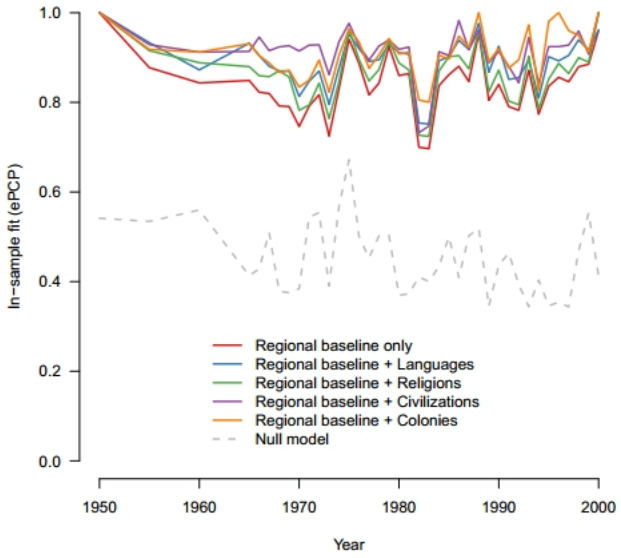


Figure 10: Lupu and Greenhill's figure. Their analysis shows how determinative region is for common IGO memberships

the contiguous neighbors. Figure 11 depicts Greenhill's belief that Human Rights Behavior might be effected by geographically proximate states.

However, my perspective contrasts with Greenhill's as shown in Figure 12; I believe region might also largely *determine* the IGO Network, and the relationship that Greenhill uncovers might be spurious. The "treatment" of a given IGO cohort (and therefore IGO Human Rights context) is assigned largely according to region. The intuition that this is indeed the case is supported by the Lupu and Greenhill empirical study (2011) which shows that the greatest predictor of having similar IGO networks is region as show in their figure which I reproduce here, Figure 10.[8]

Thinking about the process that generated the data, I would say that the possible assignments of the treatment of interest (IGO networks human rights profile) is limited. Using classical permutation tests allows us to consider what kind of result the modeling strategy would have returned if the assignment of the "treatment" data had been otherwise and constraining the permutation test asks the question but looks at a subset of permutations, which can be more realistically imagined.

We would not imagine that the IGO network for Paraguay could possibly have been assigned the IGO network assigned to Poland. The region is too deterministic for it to be so. The constrained permutation test would not allow for the reassignment of the Paraguay data to Poland. By contrast, the constrained permutation test *does* allow Paraguay's IGO network data to be reassigned to Argentina and Bolivia. [9]

They are likely to look somewhat similar because they are a member with Argentina and Bolivia in many regional organizations. However, they are obviously not identical. Therefore, the shuffle allows us to compare

[8]The authors use the network-analytic tool of modularity maximization to construct dynamic "IGO Communities." Then they evaluate which factors explain the IGO network structure.

[9]Lupu and Greenhill's discussion of their findings is related: "In the case of certain country pairs, the degree of continuity among their memberships in IGO communities is especially high. For example, the Italy-Denmark, Chile-Bolivia and Venezuela-Mexico dyads share a common IGO community membership in more than 98the years for which data are available. On the other hand, a large number of dyads very rarely – if ever – share membership in the same IGO community. Examples of dyads whose rate of shared membership is less than 5Russia-Saudi Arabia. Between these two extremes is a large number of dyads whose patterns of shared community membership is less stable; in some years they belong to the same IGO community and in many others they do not. Examples of these include Azerbaijan-Belarus, China-Sri Lanka, and Hungary-Indonesia, all of which have a continuity score of between 0.45 and 0.55."

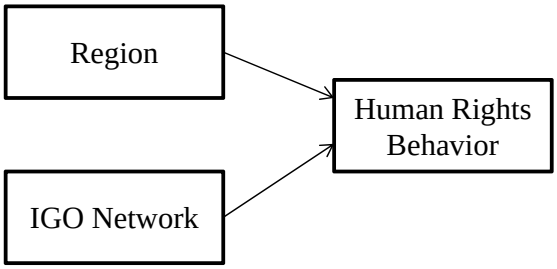


Figure 11: Greenhill's approach to controlling for neighborhood effects

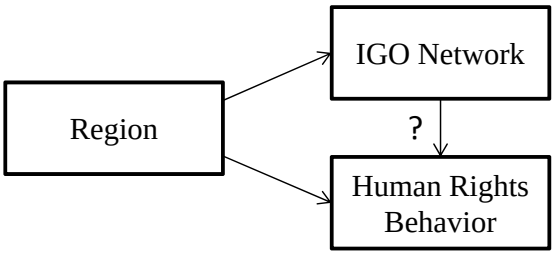


Figure 12: The region may be determining the IGO Network

how the modeling strategy would "react" (in terms of test statistics and estimates) to potential, believable inputs (i.e. IGO networks that are experienced within one's own region) versus the real-life assignment of IGO networks.

The regional characteristics become more of a baseline under this procedure, and the idiosyncrasies of countries' networks (for example non-regional valence component) of the IGO Networks becomes more prominent. Though procedurally different, the objectives of the constrained permutation approach are similar to the inclusion of a control variable. If region is the real driver of the correlation between IGO Network (lagged) and Human Rights Behavior, the new reference distribution that is created from the constrained randomization procedure should make the outcomes of the true data look not unusual. That is, the true test statistic might fall right in the middle of the constructed reference distribution, calling into question a rejection of the null hypothesis. Alternatively, if the null hypothesis is still rejected after the constrained permutation, the author would have greater confidence that the results is not spurious due to modeling misspecification. Compared to the test statistics returned for the constrained permutations, the actual test statistic may be unremarkable (consistent with the null hypothesis) or may be very outlying (leading the researcher to reject the null hypothesis).

As described above, the *complete* permutation test decouples the independent variable of interest from the dependent variable. For example, the IGO context variable time-series associated with Spain could be reassigned to India in the case of *complete* permutation.[10] However, with a constrained permutation test, not all reassignments of the data are allowed. Reassignment of the data, instead, is constrained to countries within a country's own region. This is theoretically motivated by the belief that the IGO variable's relationship with the country's human right outcome might be driven primarily by region. If human rights outcomes move together regionally this might effect the modeling outcomes. Since countries are likely to have the highest common memberships with countries in their region, their IGO network is likely to be highly determined by the region. Therefore, I regionally constrain the reassignment of the time series. This amounts to mimicking the data generating process, in so far as the likely regional assignments of IGO network data. I generate a set of potential outcomes which, I argue, are better comparisons for the true outcome since we want to control for regional causality.

The regions are taken from the 'rworldmap' package in R (Scutt-Phillips, Bivand, Foster & South 2012) (See also South (2011)), which are derived from 2008 Environmental Performance Index (EPI) downloaded from `http://epi.yale.edu/`. It is clear that the regional classifications were not created with human rights outcomes in mind. It is likely, however that cultural factors do determine where the lines are drawn. I think that the procedure is still not problematic for this analysis as Greenhill wants to control for cultural factors as well such as common language or common colonial history. The EPI Variable considers 8 different regions as shown in Figure 13 : Central and Eastern Europe (19 countries in data set, 19! = 1.216451e+17 possible permutations), East Asia and the Pacific (15 countries, 15!=1.307674e+12), Europe (21 countries, 21! = 5.109094e+19 possible permutations), Latin America and Caribbean (22 countries, 22! = 1.124001e+21 possible permutations), Middle East and North Africa (14 countries, 14! = 8.7178291e +10 possible permutations), North America (2 countries, 2! =possible permutations), South Asia (6 countries, 6! = 720 possible permutations), and Sub-Saharan Africa (38 countries, 38!= 5.230226e+44 possible permutations). For some regions the the chances that country-data is "correctly" assigned to the actual country is quite high, (as in North America - only Canada and the US are included in the category so the chance is 50/50), but in most cases it will be a rare occurance.

A complete permutation test, given the 138 countries in the data set, would have yielded 5.012889e+234 possible permutations of the data. The constraining of the permutation test limits the data reassignment by many many orders of magnitude. Still, there is a great deal of freedom with 5.997856e+128 ($19! * 15! * 21! * 22! * 14! * 2! * 6! * 38!$) possible permutations. Running so many permutations would not be practical, and the literature suggests that with a random subset of about 1000 of permutations, researchers can get a good sense of potential outcomes (Manly 2007, Moore, McCabe, Duckworth & Alwan 2008).

As discussed above, the procedure is performed holding the time-series in tact. I shuffled the IGO context variable, but only allowing reassignment within the EPI region. To make a broad set of comparisons repeat this procedure 1000 times and then rerun with the constrained permuted data.

In Table 4 I compare the statistics generated by modeling with the true data to the set of statistics generated by modeling with 1000 permutations of the data, using the constrained permutation procedure I outline above. Figures 14, 15, and 16 show the distributions from which the values in Table 4 are derived. After collecting the z-statistics for 1000 constrained permutations, and creating the new reference distribution, we see that the z-statistic for the actual data is *not* very outlying in comparison. At most, we see that the z-statistic is greater than 674 of the permutations of the data. In all other cases, it is not even greater than half of the permutations, and lies closer to zero than more than 800 of the permutations (lag 2 years to lag 5 years). I also make this comparison for the p-value and the estimate to illustrate how unsurprising the result is when compared to modeling results of the randomized data.

Figure 15 contains the p-value QQ plots for the constrained permutation test which compares the quantiles

[10]For more discussion of the alternate permutation methods see Kennedy (1995). Kennedy and Cade (1996) "suggest that the simple method of shuffling Z [the independent variable of interest] is sufficient in the multivariate context.... Further Monte Carlo analyses in O'Gorman (2005) confirms that the simple shuffle Z method performs as well in terms of size (falsely rejects the null hypothesis at an acceptable rate) and power (correctly rejects the null hypothesis at an acceptable rate), even in the presence of non-normal error structures and high correlation between Z and the other covariates X." From Rader 2011 "Randomization Tests and Inference with Grouped Data"

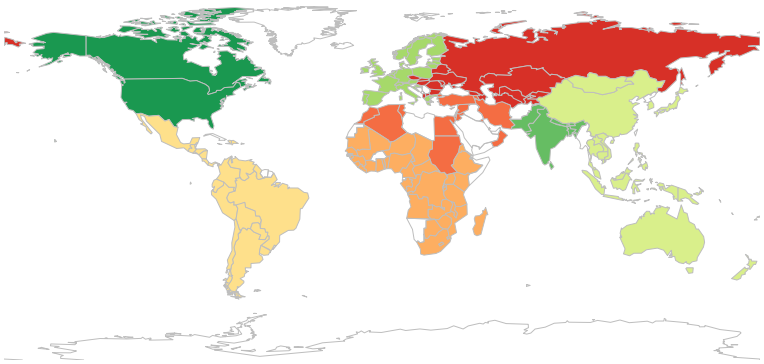


Figure 13: Environmental Performance Index (EPI) regions which are used in the analysis

of the p-values for the permutations and the quantiles we would expect for p-values of random data (an even distribution from zero to one). We see that for the vast majority of the shuffled data the null will be rejected with 95 percent confidence. I would assert that this indicates that the false positive rate is too high. The z-distribution in the regress framework which is used to translate into the p-values simply does not account for the data generating process. The horizontal line in the plots is at y=.05 as a reference.

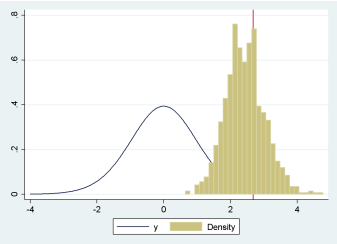
I include Figure 16 to suggest how inappropriate it might be to think about rejection of the "null hypothesis" as hypothesis that the estimate for the IGO Human Rights Context variable is different from zero. For all 5000 estimates generated from the constrained permutation tests, (1000 for each regression) none of estimates is negative! If it is true that the data generating process is such that the time series data could only feasibly be assigned to another country within the same region, and given that all of the feasible permutations of the data lead to positive estimates,[11] it is not logical to talk about the estimate's difference from zero as a measure of confidence for rejecting the null. Instead we should be concerned with how unusual the true estimate is in the distribution of *potential* outcomes. (This also is why I consider the position of the true z-statistic being greater or less than a that created via a permutation. I do not need to look at the z-statistic from zero as an absolute value, but rather distance from zero.)

Finally, and most importantly for interpreting the results of the original regression, I create a distribution of z-statistics generated from rerunning the model with 1000 permutations of data, shown in the histograms in Figure 14. The reference distribution contrasts with the z-statistic distribution that is assumed in the model. I adjusted p-values using the z-statistic and report them in the bottom of Table 4. This is consistent with the literature, "The results from Monte Carlo analyses in Kennedy and Cade (1996) suggest that a simple method of shuffling Z is sufficient to the multivariate context so long as inferences are based on the distribution of test statistics and not on the distribution of coefficients." The p-values are unimpressively large values ranging from 0.33 to 0.91, suggesting that it is inappropriate to reject the null hypothesis of no effect for IGO human rights context on the human rights behavior of individual countries. The difference in the reference distribution that Greenhill is using and which I use explain the difference in our p-values. The position of the z-statistic for the true configuration of the data very outlying for the assumed distribution, while the position of the z-statistic for the true configuration of the data is *not* very outlying for the distribution created by the randomization procedure.

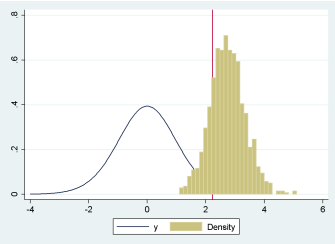
[11] Granted this is a subset of the total set of what I would call feasible permutations, since it is only 1000 of the permutations. Still 1000 permutation is a good distribution of the total set of permutations

Table 4: How do Greenhill's model results compare to 1000 constrained permutations of IGO Human Rights Context Variable?

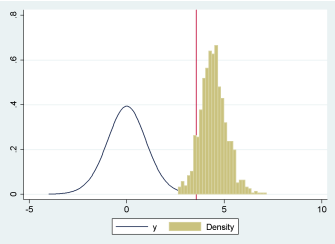
Variable	**(1-Year lag)**	**(2-Year lag)**	**(3-Year lag)**	**(4-Year lag)**	**(5-Year lag)**
Z-statistic outlyingness (True value is bigger than how many z-statistics of 1000 permutations?)	674	185	87	178	181
Greenhill z-stat	2.68	2.23	3.56	2.70	2.06
p-value's true 'shock value' (true value smaller than how many in 1000)	693	185	82	178	184
Greenhill p-value	0.007	0.026	0.0004	0.007	0.039
Number of times that p-value is <.05	789	912	1000	983	864
Estimate outlyingness (True value is bigger than how many estimates of 1000 permutations?)	891	524	699	564	574
Greenhill Estimate	.2564	.2269	.3753	.2315	.2058
Adjusted p-value (derived from z-stat distribution)	.326	.815	0.913	0.822	0.819



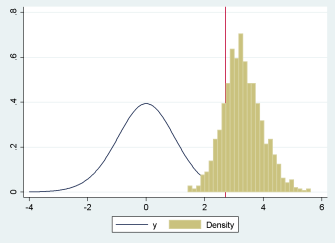
(a) Lag 1-Year



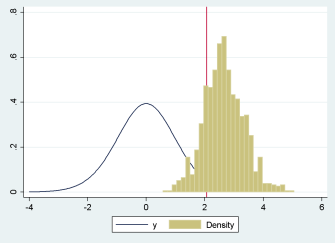
(b) Lag 2-Year



(c) Lag 3-Year

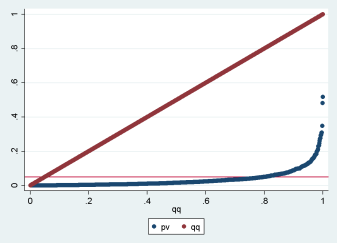


(d) Lag 4-Year

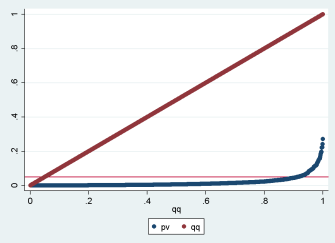


(e) Lag 5-Year

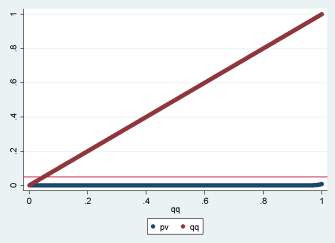
Figure 14: Z-statistic Histograms (a), (b), (c), (d), and (e) are th new reference distribution created through the constrained permutation procedure. The vertical line indicates the position of the z-statistic for the observed data. The bell curve centered around zero is the z-statistic distribution that are used to derive p-values in Greenhill's analysis. The z-statistics produced do appear outlying considering these theoretical distributions, but not to the reference distributions resulting from the constrained permutation test.



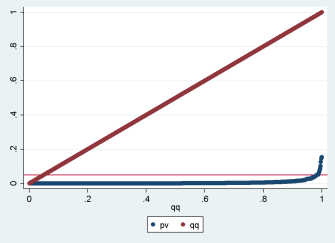
(a) Lag 1-Year



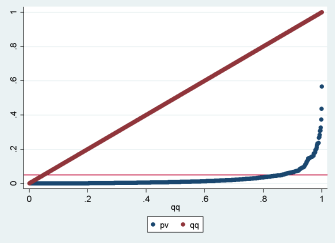
(b) Lag 2-Year



(c) Lag 3-Year

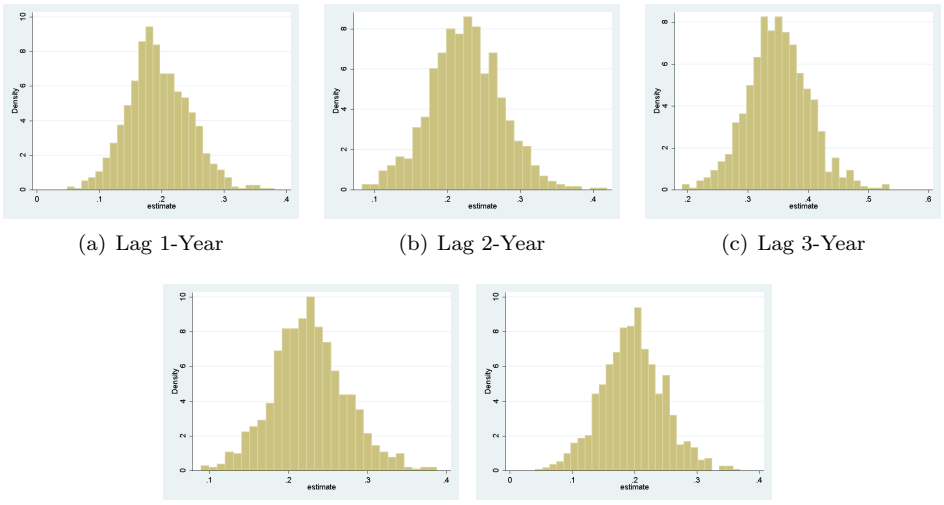


(d) Lag 4-Year



(e) Lag 5-Year

Figure 15: QQ Plots (f), (g), (h), (d), and (e)



(a) Lag 1-Year (b) Lag 2-Year (c) Lag 3-Year

(d) Lag 4-Year (e) Lag 5-Year

Figure 16: Estimate Histograms (f), (g), (h), (d), and (e)

6 Conclusion

This paper explores a new procedure for making time-series cross-sectional analyses more rigorous. The pitfalls of time-series cross sectional data analysis are becoming more and more well-known. My approach is to hold most of the architecture of the analysis in place (control variables and specification), I introduce plausible alternate treatments to get a sense of what the model could have estimated given likely data generating processes for the independent variable of interest - i.e. the true treatment. When the realized treatment variable does not produce outlying statistics compared with plausible treatments, it is hard to assert that the estimate represents a *treatment effect* of the treatment variable on the outcome variable as is the typical implied if not explicit interpretation.

But does the constrained permutation test really expose a problem that could not have been exposed another way? The answer is, maybe, yes. There are countless variables that might be added to these regressions and different procedures used that could improve the models' consistency with modeling assumptions, such as differencing, including more lags, detrending etc. However, researchers have settled on these specifications as their preferred specification which peer reviewers also have found acceptable.[12] Using non-parametric tests asks the researcher to be explicit about the comparisons that he is willing to make given his beliefs about the data generating process for the treatment - which also becomes explicit in the process. In some cases, researcher will find that they are not able to make such strong claims as they would have without this procedure, but the field benefits as a whole by standing on the stronger footing by making relevant comparisons.

7 Appendix: Even more on IGO context

In this section, I explore different proceedures that could have been implemented with the Greenhill data and modeling specification. First, as a point of comparison with the *constrained* permutation test, I show the results of a *full* permutation test. That is the treatment data is reassigned shuffling among all countries. This implements exactly the protocol in Erikson et al. (2014).[13] Running the *full* permutation *only* would not challenge Greenhill's conclusions. Second, I show the procedure of subsetting the data by region and then running *full* permutation tests on each model in the subset. For the subsets, this procedure follows Erikson et al. (2014) exactly but does lead scholars to question the validity of the Greenhill result. Third, I include an additional lag of the dependent variable according to how much the independent variable of interest is lagged. This is because the dependent variable is introduced on the left-hand side by shuffling the IGO context variable. Concerned about this, I control with a lagged dependent variable and rerun the model. In future versions of this paper, I should remove the part of the shuffled independent variable that is the dependent variable. For example, when Indonesia's IGO context is reassigned to Malaysia, I need to remove the part of Indonesia's IGO context comes from being in IGOs with Malaysian.

7.1 Constrained Permutation Test Contrasted to Full Permutation Test

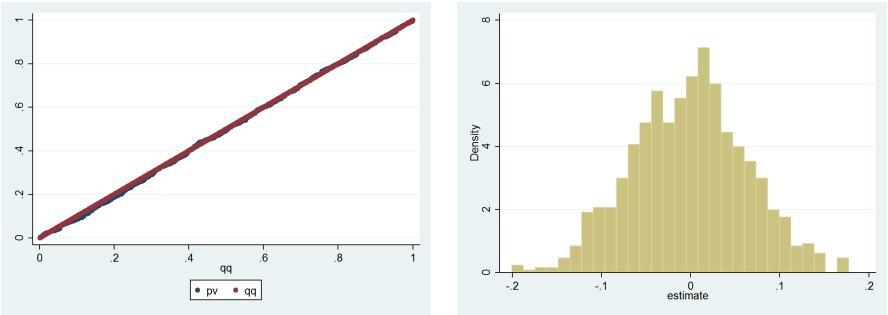
Here, I take a bit of space to show what we would have seen if we had run the *unconstrained* permutation test (allowing reassignment of the IGO human rights context time series of a country to *any* other country). Would the variable still have been deemed statistically significant? Figures 17 and 18 suggest that the data is well behaved. It is centered around zero. The estimates are very outlying within the distributions, suggesting that the high level of confidence communicated by the p-value is appropriate.

[12] This is actually not true for D'Souza. In future research I will look at her prefered specification which interestingly includes exporter and importer time specific dummies as well as pair fixed effects. This results in a computationally intensive specification. I have not yet done the comparisons for this specification. D'Souza notes that it is not typical so it might not be very representative of what Political Scientists are doing anyway.

[13] Of course their data is dyadic, so in fact implementing a complete permutation with the Greenhill dataset is simpler.

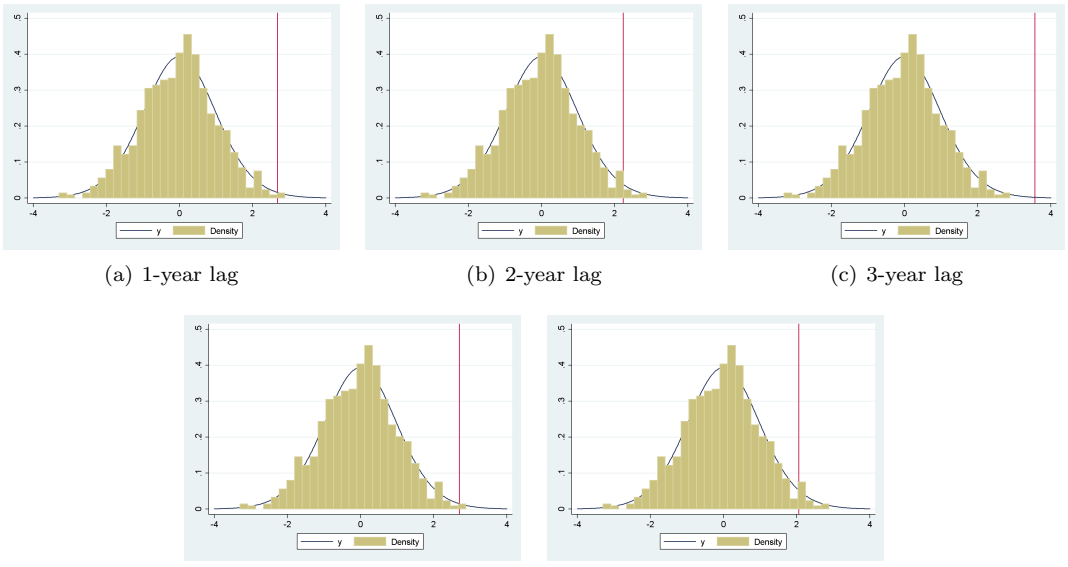
Table 5: After the *Unconstrained* Permutation, How do True Outcomes Compare?

Variable	**(1-Year lag)**	**(2-Year lag)**	**(3-Year lag)**	**(4-Year lag)**	**(5-Year lag)**
Z-statistic outlyingness (True value is bigger than how many z-statistics of 1000 permutations?)	997	990	1000	1000	976
Greenhill z-stat	2.68	2.23	3.56	2.70	2.06
p-value's true 'shock value' (true value smaller than how many in 1000)	993	975	1000	993	953
Greenhill p-value	0.007	0.026	0.0004	0.007	0.039
Estimate outlyingness (True value is bigger than how many estimates of 1000 permutations?)	1000	1000	1000	1000	1000
Greenhill Estimate	.2564	.2269	.3753	.2315	.2058
Adjusted p-value (derived from z-stat distribution)	.003	.01	<.001	<.001	.004



(a) QQ-plot (b) Histogram Estimates

Figure 17: Each of the model's Z-statistic histograms for complete permutation test which allows reassignment of the IGO Human Rights Context of a country to any other country in the data-set. The observed value is shown with a vertical red line.



(a) 1-year lag (b) 2-year lag (c) 3-year lag

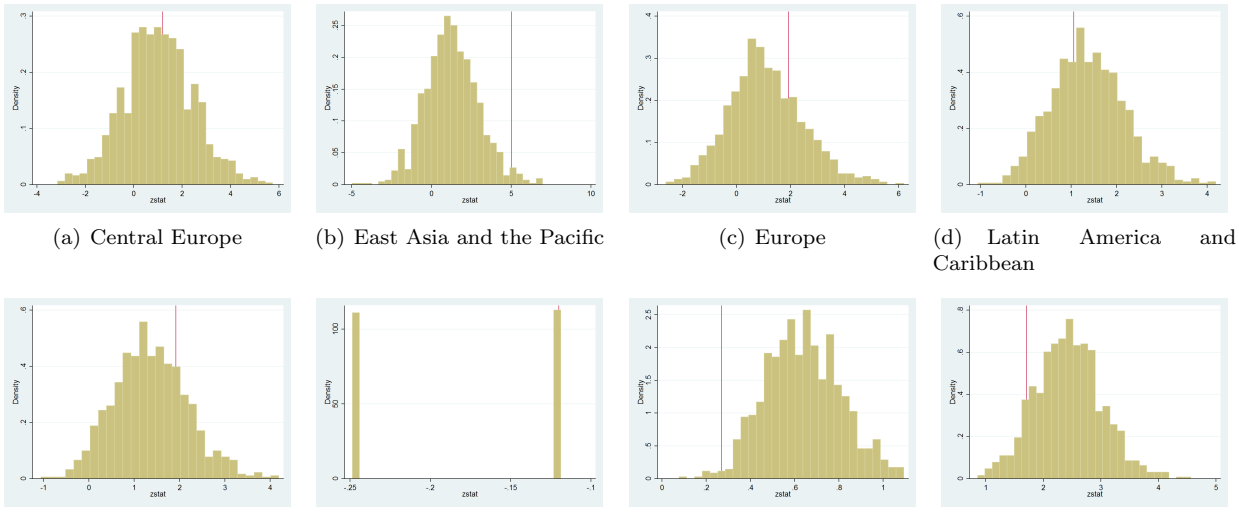
(d) 4-year lag (e) 5-year lag

Figure 18: Three-year lag Z-statistic Histograms for complete permutation test which allows reassignment of the IGO Human Rights Context of a country to any other country in the data-set.

7.2 Full Permutations in Models of Subsetted Data

If the reader is uncomfortable with the constrained permutation test, another approach is to use the full permutation test, after subsetting on the potentially confounding variable of concern. It is typical for researchers to subset data and see if their result hold as a robustness check. I implement this procedure for the 3 year lag, where the coefficient on *IGO Human Rights Context* is the greatest in the original model. I note that for most of the regions, the estimate is still positive (with the exception of North American, which only includes Canada and the U.S.), though the statistical significance has dropped in most cases (the p-values increase for all cases except East Asia and the Pacific).

Another way of looking at the constrained shuffle is subsetting the data by region, and then applying the permutation procedure *post hoc*. Even after subsetting the data, we notice that in every case, except for two country region North America, the estimates are still positive. Stopping the analysis at this point, the researcher might feel that he is still on strong grounds for saying that IGO human rights context is likely to change domestic human rights outcomes. However, running an *unconstrained* permutation test on each data set, would lead us to question this position of confidence.



(a) Central Europe (b) East Asia and the Pacific (c) Europe (d) Latin America and Caribbean

(e) Middle East and North Africa (f) North America (g) South Asia (h) Sub-Saharan Africa

Figure 19: Three-year lag Z-statistic Histograms for shuffled regional subset (a), (b), (c), (d), and (e)

Table 6: Subsetting the Data, How do Z-statistics compare with distribution of Z-statistics generated by full permutation test (1000 permutations)

Variable	**(True Coef.)**	**(Z)**	**(p-value)**	**(no. of Zstats<true)**
Central and Eastern Europe	0.9284125	1.2	0.231	559
East Asia and the Pacific	1.096528	5.03	$5.3e^{-7}$	976
Europe	0.8234178	1.93	0.054	741
Latin America and Caribbean	0.3602877	1.06	0.29	364
Middle East and North Africa	0.6427928	1.92	0.055	757
North America	-0.7975935	-0.12	0.906	496*
South Asia	0.1872536	0.27	0.788	13
Sub-Saharan Africa	0.2222008	1.71	0.087	96

The researcher might initially be convinced that she has uncovered that the IGO context hypothesis seems

to be supported in the East Asia and Pacific category. However, from a multiple comparisons perspective, this is not as clear. Consider that as outlying as as the text statistic is for East Asia and the Pacific, the position of the z-statistics for South Asia and Sub-Saharan Africa are about as *inlying* - they are surprisingly close to zero given the reference distribution.

7.3 Additional Lags of Dependent Variable

The reassignment of the data may be a cause for concern because the lagged dependent variable is introduced as a component of the shuffled data. For example, when Argentina's network is reassigned to Peru, the human rights rating for Peru will be be a component of that variable, as Peru is in organizations with Argentina. Thus a component of the reassigned variable will simply be the lagged dependent variable. This is probably a minuscule component, many other countries' data is also included in the variable and should greatly mask it's presence.

Still, this issue can be addressed by actually including the lagged dependent variable as an additional predictor variable. The number of years lagged will be exactly the number of years that the IGO context independent variable is lagged in the model. For the case of the one-year lag of the IGO variable, in fact, the constrained randomization as I have performed it above, does just this, as Greenhill originally includes a one-year lag of the dependent variable and the IGO context variable is lagged one year. For the other models (2-year, 3-year, 4-year, and 5-year lags), I add the lag of the dependent variable accordingly.

(Ordering the outcomes of the permutations from 1 to 1000 - after which outcome)

It should be noted that since the variation on the IGO context variable is small, we see a greater magnitude in the estimate than for the neighbor variable, which has greater overall variation.

IGO memberships do correlate with region. The neighborhood variable that Greenhill includes in his analysis is a very limited interpretation of neighborhood. In table 2 of his article, the neighborhood of variable is not significant in any of the models (at the 90igo context variable from the model, the neighborhood variable jumps to a high level of significance. The p-value drops to .001, and the estimates size almost doubles.

He only includes contiguous states and nearby islands as neighbors for his *Neighborhood Effect* variable. The variable "provides a measure of the average PIR scores of each individual state's contiguous neighbors and nearby island states." He also writes "Data on the levels of geographical connectedness between pairs of states were obtained from O'Loughlin, Ward, Lofdahl, Cohen, Brown, Reilly, Gleditsch, and Shin (1998)."

Table 7: Ordered Probit Model of Countries' Physical Integrity Rights - (Adding Lags)

Variable	(1-Year lag)	(2-Year lag)	(3-Year lag)	(4-Year lag)	(5-Year lag)
Key Independent Variable					
IGO Human Rights Context	0.256**	0.116	0.216*	0.107	0.069
Domestic Controls					
Lagged dependent variable (1-year)	–	0.427***	0.434***	0.480***	0.491***
Lagged dependent variable (1,2,3,4,5 years)	0.531***	0.211***	0.220***	0.160***	0.158***
Democracy	0.014**	0.009*	0.009	0.010*	0.010
Trade	0.004***	0.003***	0.002**	0.002**	0.003***
FDI	0.000	0.002*	-0.003***	-0.003	-0.002
GDP per capita (logged)	0.100**	0.118***	0.105***	0.123***	0.124***
Civil War	-0.402***	-0.153	-0.130	-0.047	0.006
International Controls					
International War	-0.511***	-0.518***	-0.721***	-0.545**	-0.406*
Regime durability	0.003*	0.002	0.001	0.001	0.001
Population Density	-0.000	-0.000	-0.000	-0.000	-0.000
Hard PTA membership	-0.087	-0.109	-0.146	-0.173*	-0.214**
Soft PTA membership	0.115	0.205**	0.180**	0.163**	0.155*
Neighborhood effect	0.035	0.040*	0.031	0.035*	0.045*
Common language	0.023	0.013	0.006	0.008	0.007
Common colonial history	-0.022	-0.020	-0.014	-0.004	-0.003
Log likelihood					
N					

Note: Significance levels for two-tailed tests are marked as follows: + p<.1, * p < .05, ** p < .01, and *** p < .001.

Table 8: After the Constrained Permutation with Lagged Dependent Variable according to year IGO human Rights Context is lagged

Variable	**(1-Year lag)**	**(2-Year lag)**	**(3-Year lag)**	**(4-Year lag)**	**(5-Year lag)**
z-statistic outlyingness (True value is bigger than how many z-statistics of 1000 permutations?)	674	363	248	263	357
z-stat (Greenhill)	2.68	1.28	2.26	1.36	0.76
p-value's true 'shock value' (True value smaller than how many in 1000)	693	360	246	273	352
p-value (Greenhill)	0.007	0.202	0.024	0.172	0.448
Estimate outlyingness (True value is bigger than how many estimates of 1000 permutations?)	891	535	632	414	477
Estimate (Greenhill)	.256	.116	.216	.107	.069
Adjusted p-value (derived from z-stat distribution)	0.326	0.637	0.752	0.727	0.648

References

Achen, Christopher H. 2000. "Why lagged dependent variables can suppress the explanatory power of other independent variables." *Ann Arbor* 1001:48106–1248.

Bailey, Rosemary A. 1986. Randomization, Constrained. In *Encyclopedia of Statistical Sciences*, ed. Samuel Kotz & Norman L. Johnson. Vol. 7 John Wiley New York pp. 524–530.

Beck, N. & J.N. Katz. 1995. "What to do (and not to do) with time-series cross-section data." *American Political Science Review* pp. 634–647.

Beck, N. & J.N. Katz. 2001. "Throwing out the baby with the bath water: A comment on Green, Kim, and Yoon." *International Organization* 55(2):487–495.

Bertrand, M., E. Duflo & S. Mullainathan. 2004. "How Much Should We Trust Differences% in% Differences Estimates." *Quarterly Journal of Economics* 119(1):249.

D'Souza, Anna. 2012. "The OECD anti-bribery convention: changing the currents of trade." *Journal of Development Economics* 97(1):73–87.

Dube, A., E. Kaplan & S. Naidu. 2011. Coups, corporations, and classified information. Technical report National Bureau of Economic Research.

Erikson, Robert S., Pablo M. Pinto & Kelly T. Rader. 2010. "Randomization Tests and Multi-Level Data in US State Politics." *State Politics & Policy Quarterly* 10(2):180.

Erikson, Robert S, Pablo M Pinto & Kelly T Rader. 2014. "Dyadic Analysis in International Relations: A Cautionary Tale." *Political Analysis* pp. 1–7.

Green, D.P., S.Y. Kim & D.H. Yoon. 2001. "Dirty pool." *International Organization* 55(2):441–468.

Greenhill, Brian D. 2010. "The Company You Keep: International Socialization and the Diffusion of Human Rights Norms." *International Studies Quarterly* 54(1):127–145.

Kennedy, Peter E. 1995. "Randomization tests in econometrics." *Journal of Business & Economic Statistics* 13(1):85–94.

Kennedy, Peter E. & Brain S. Cade. 1996. "Randomization Tests for Multiple Regression." *Communications in Statistics-Simulation and Computation* 25(4):923–936.

King, G. 2001. "Proper nouns and methodological propriety: Pooling dyads in international relations data." *International Organization* 55(02):497–507.

Kristensen, I.P. & G. Wawro. 2003. Lagging the dog? The robustness of panel corrected standard errors in the presence of serial correlation and observation specific effects. In *annual meeting of the Society for Political Methodology, University of Minnesota.*

Luechinger, S. & C. Moser. 2012. "The Value of the Revolving Door: Political Appointees and the Stock Market.".

Lupu, Yonatan & Brian D. Greenhill. 2011. "'Clubs of Clubs': A Networks Approach to the Logic of Membership in Intergovernmental Organizations." *Annual Political Networks Conference* .

Manly, Bryan F.J. 2007. *Randomization, Bootstrap and Monte Carlo Methods in Biology.* CRC Press.

Moore, David, George McCabe, William Duckworth & Layth Alwan. 2008. *The Practice of Business Statistics.* W.H. Freeman and Company.

Neumayer, Eric. 2005. "Do international human rights treaties improve respect for human rights?" *Journal of conflict resolution* 49(6):925–953.

Oneal, J.R. & B. Russett. 2001. "Clear and clean: The fixed effects of the liberal peace." *International Organization* 55(2):469–485.

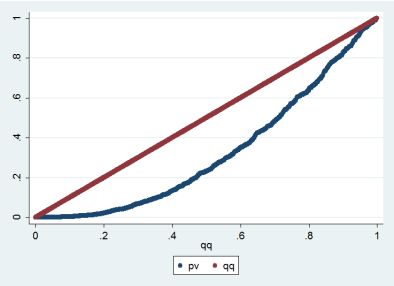
Scutt-Phillips, B.R., R. Bivand, P. Foster & M.A. South. 2012. "Package rworldmap.".

South, A. 2011. "rworldmap: A New R package for Mapping Global Data." *R Journal* p. 35.

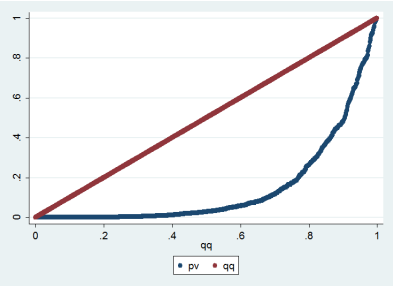
Von Stein, J. 2005. "Do treaties constrain or screen? Selection bias and treaty compliance." *American Political Science Review* 99(4):611.

Wawro, Gregory. 2002. "Estimating dynamic panel data models in political science." *Political Analysis* 10(1):25–48.

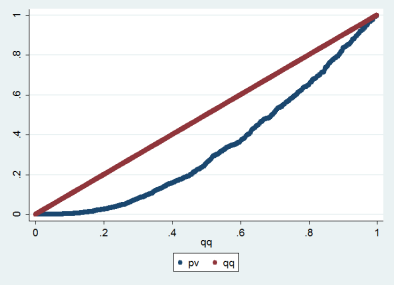
Wilson, S.E. & D.M. Butler. 2007. "A lot more to do: The sensitivity of time-series cross-section analyses to simple alternative specifications." *Political Analysis* 15(2):101–123.



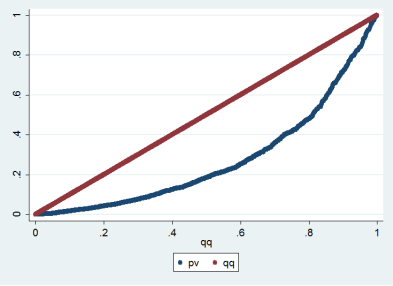
(a) Central Europe



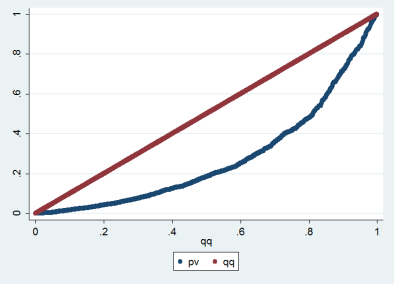
(b) East Asia and the Pacific



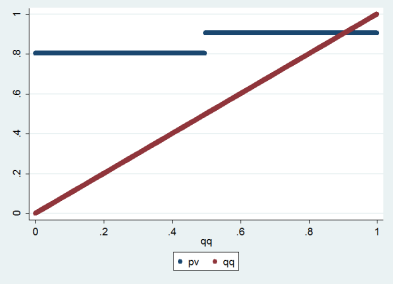
(c) Europe



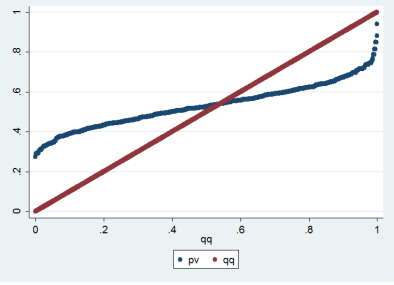
(d) Latin America and Caribbean



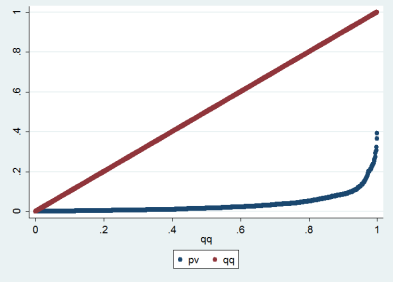
(e) Middle East and North Africa



(f) North America

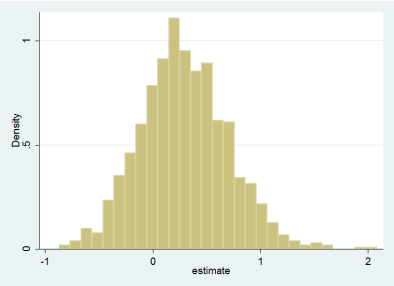


(g) South Asia

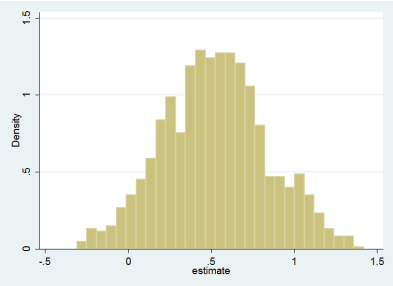


(h) Sub-Saharan Africa

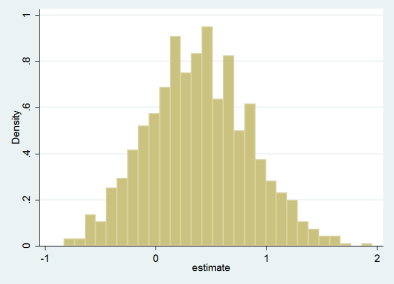
Figure 20: Three-year lag QQ-plots for shuffled regional subset (f), (g), (h), (d), and (e)



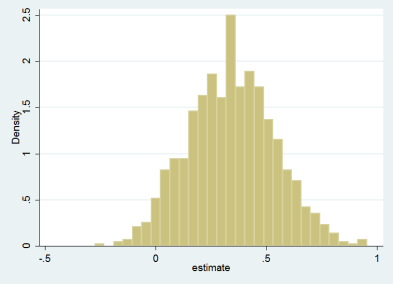
(a) Central Europe



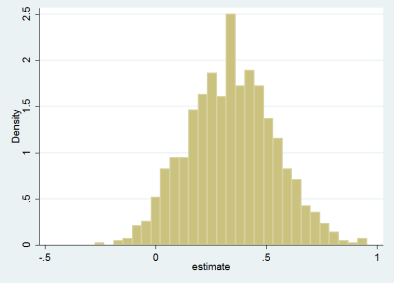
(b) East Asia and the Pacific



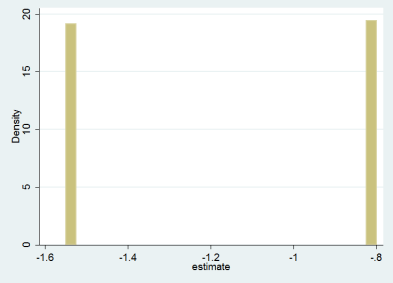
(c) Europe



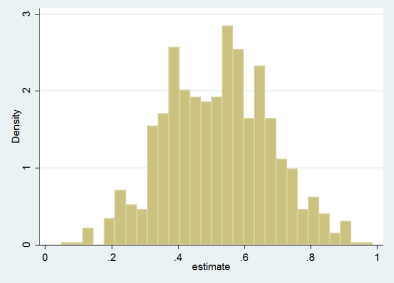
(d) Latin America and Caribbean



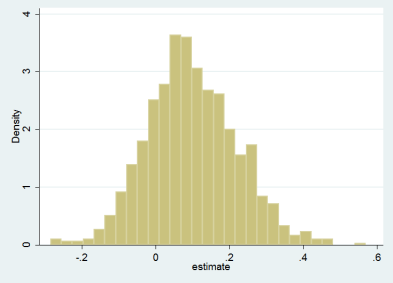
(e) Middle East and North Africa



(f) North America



(g) South Asia



(h) Sub-Saharan Africa

Figure 21: Histograms of estimates for shuffled regional subset (a), (b), (c), (d), (e), (f) and (g)

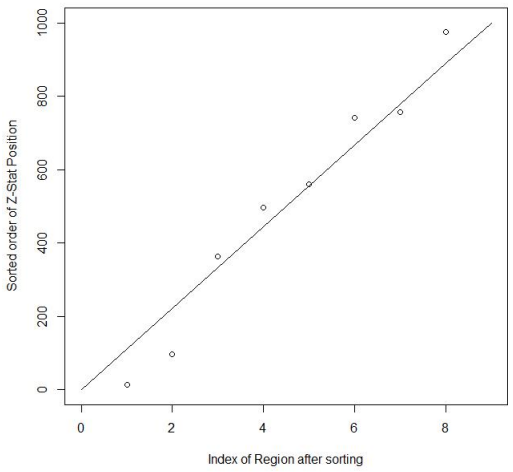


Figure 22: Plot of where Z stats Fall for 8 regions, ordering by position in reference distribution