The Political Methodologist

Newsletter of the Political Methodology Section
American Political Science Association
Volume 22, Number 2, Spring 2015

Editor:
Justin Esarey, Rice University
justin@justinesarey.com

Editorial Assistant:
Ahra Wu, Rice University
ahra.wu@rice.edu

Associate Editors:
Randolph T. Stevenson, Rice University
stevenso@rice.edu

Rick K. Wilson, Rice University
rkw@rice.edu

Contents

Notes From the Editors

The editors are proud to present the Spring 2015 print edition of *The Political Methodologist*!

This issue starts off with the Data Access and Research Transparency (DA-RT) statement that has been endorsed by a large number of leading journals in political science. *The Political Methodologist* is proud to be a part of this movement, which we joined in 2014 with a statement by Associate Editor Rick Wilson. Colin Elman and Arthur Lupia asked journals that have signed on to the statement to simultaneously print it as a strong signal to the discipline that reproducible and transparent research is a priority for political science and must be a part of the practice of researchers working today. Christopher Gandrud's contribution, correcting and refining aspects of an existing data set, underscores how making data and replication code freely available enables the continuous improvement of research.

We also have two papers about the use of causal inference methods in quantitative political research. Natália Bueno and Guadalupe Tuñón write about why presenting graphical robustness checks for regression discontinuity results should be a part of our practice, and helpfully provide examples with R code to facilitate the wider use of this technique. Tom Pepinsky uses a provocative question—"why can't we just make up instrumental variables?"—to make an instructive point about what makes an instrumental variable valid and how we can be clearer in our research and pedagogy about how we choose these variables.

Finally, an update on *TPM*'s circulation. I am pleased to report that our on-line platform continues to be a very successful supplement to this print format: between January and June of this year, our on-line website has received 18,807 unique visitors. I'm hopeful that our upcoming special issue on peer review (note the Call for Papers noted in this issue) will attract even more interest in *TPM* and in our shared project of improving the practice of political methodology.

Data Access and Research Transparency (DA-RT): A Joint Statement by Political Science Journal Editors

Colin Elman
Syracuse University
celman@maxwell.syr.edu

Arthur Lupia
University of Michigan
lupia@umich.edu

In this joint statement, leading journals commit to greater data access and research transparency, and to implementing policies requiring authors to make as accessible as possible the empirical foundation and logic of inquiry of evidence-based research. Please visit http://www.dartstatement.org/ for more information.

Transparency requires making visible both the empirical foundation and the logic of inquiry of research. We agree that by January 15, 2016 we will:[1]

1. Require authors to ensure that cited data are available at the time of publication through a trusted digital repository.[2] Journals may specify which trusted digital repository shall be used (for example if they have their own dataverse).
    - If cited data are restricted (e.g., classified, require confidentiality protections, were obtained under a non-disclosure agreement, or have inherent logistical constraints), authors must notify the editor at the time of submission. The editor shall have full discretion to follow their journal's policy on restricted data, including declining to review the manuscript or granting an exemption with or without conditions. The editor shall inform the author of that decision prior to review.
2. Require authors to delineate clearly the analytic procedures upon which their published claims rely, and where possible to provide access to all relevant analytic materials. If such materials are not published with the article, they must be shared to the greatest extent possible through institutions with demonstrated capacity to provide long-term access.
3. Maintain a consistent data citation policy to increase the credit that data creators and suppliers receive for their work. These policies include using data citation practices that identify a dataset's author(s), title, date, version, and a persistent identifier. In sum, we will require authors who base their claims on data created by others to reference and cite those data as an intellectual product of value.
4. Ensure that journal style guides, codes of ethics, publication manuals, and other forms of guidance are updated and expanded to include improved data access and research transparency requirements.

Corrections and Refinements to the Database of Political Institutions' yrcurnt Election Timing Variable

Christopher Gandrud
Hertie School of Governance
gandrud@hertie-school.org

The **yrcurnt** variable in the Database of Political Institutions (Beck et al. 2001, updated in 2013)[1]–DPI–is a regularly used measure of government election timing. For example, Alt, Lassen, and Wehner (2014) use the variable in their recent study of fiscal gimmickry in Europe. They find that fiscal gimmickry–straying from accepted accounting standards–is more common directly before elections (and in countries with weak fiscal transparency).

Because the DPI's **yrcurnt** variable is so regularly turned to for testing how election timing affects governments' choices, it is especially important that it be reliable and valid. However, the variable in the current (2013) DPI release has a number of issues that this note aims to correct.[2]

[1] Part of this list draws on language used in "Research Transparency, Data Access, and Data Citation: A Call to Action for Scholarly Publications," Inter-university Consortium for Political and Social Research, March 16, 2014.

[2] See `http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/preservation/trust.html`, `http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf`, and `http://public.ccsds.org/publications/archive/652x0m1.pdf`.

[1] See: `http://go.worldbank.org/2EAGGLRZ40`. Accessed February 2015. All corrections and refinements discussed here refer to this version of the database.

[2] Note that Alt, Lassen, and Wehner's (2014) substantive findings hold up when using the correct data presented here, though the estimated magnitudes of the effects are reduced. See the replication repository for more details: `https://github.com/christophergandrud/Alt_et_al_2014_Replication`.

Variable definition

Before looking at problems in the **yrcurnt** variable, let's reiterate how the variable is defined. The DPI Codebook[3] (p. 4) defines the **yrcurnt** variable as the years left in a country's chief executive's current term such that:

> "a '0' is scored in an election year, and n-1 in the year after an election, where n is the length of the term. In countries where early elections can be called, YRCURNT is set to the de jure term limit or schedule of elections, but resets in the case of early elections."

Issues

The original variable has a number of issues that make it problematic for studying how election timing impacts government policymaking. The first set of concerns is about clear errors that make the variable a less reliable measure of the concept defined above. These errors can be straightforwardly corrected. The second set of concerns has to do with issues that bring into question the variable's validity in a number of cases for studying how election timing shapes policymaking. I argue that simple refinements can be made to improve the variable's validity in these cases.

In this note I focus on the EU-28 countries (all 28 current European Union member states). Problems with the variable may exist for a wider range of countries. Given the depth of problems with the **yrcurnt** variable discussed here, it would certainly be worthwhile in future work to reevaluate the variable for the full breadth of countries covered by the DPI.

I used the European Election Database[4] to find correct election years. This information was cross-checked and supplemented with individual country-election articles on Wikipedia.[5]

Errors

There are many instances in the DPI where election years are not recorded as 0. In other cases non-election years were mis-coded as 0. In some cases incorrect statutory election timing was also given. The following table lists these errors:

Country	Errors in the DPI **yrcurnt** variable for the EU-28
Belgium	Missing the 2010 election.
Croatia	Incorrect election timing for the 1995 and 2000 elections. Also, the DPI incorrectly classifies 1991 as 4 years from the next election. Croatia gained independence in 1991 and its first election as an independent country was scheduled for the following year. So, 1991 should be coded as 1 year from the next election.
Denmark	Missing the 2001 and 2007 elections.
Estonia	Incorrect election timing for the 1995, 1999, 2003, 2007, and 2011 elections. Also counting originally started at 4, but should start at 3 as there is a 4 year term limit (not 5).
Germany	Missing the 2005 election.
Greece	Missing the 2007, 2009, 2012 election years.
Ireland	Missing the 2011 election.
Italy	Missing the 2008 election.
Latvia	Missing the 2006, 2010, 2011 election years.
Netherlands	Missing the 2003 and 2006 elections.
Portugal	Missing the 1979, 1999, and 2011 elections.
Slovakia	Missing the 2012 election.
Spain	Missing the 1989, 1996, and 2011 elections.
United Kingdom	Missing the 2001 and 2005 elections.

Refinements

For a number of countries the elections recorded are for largely figurehead presidents. This can affect both when elections are recorded and how many years are given until the next election as figurehead presidents often have longer terms than parliaments. In these cases the current **yrcurnt** variable is not a valid measure of *government* election timing.

Some countries are less clear-cut because they have semi-presidential systems. Nonetheless, in a number of these cases the prime minister is the leader of the government and largely sets the domestic policy agenda. These powers are most relevant for studying topics such as public budgeting.

The following refinements should be made to create a

[3]The DPI Codebook can be found at: `http://siteresources.worldbank.org/INTRES/Resources/469232-1107449512766/DPI2012_Codebook2.pdf`. Accessed February 2015.

[4]See `http://www.nsd.uib.no/european_election_database/country/`. Accessed February 2015.

[5]See `https://www.wikipedia.org/`. Accessed September 2014 and February 2015.

more valid indicator for research on how election timing affects policymaking:

Country	Refinements to make to the DPI **yrcurnt** variable for the EU-28
Austria	Use parliamentary rather than (figurehead) presidential elections.
Lithuania	Use parliamentary rather than presidential elections. Lithuania is a semi-presidential system where the president appoints the PM, the legislature's approval is needed. The PM is more responsible for domestic policy.
Romania	Romania is semi-presidential where the president appoints the PM, but the PM must be approved by the parliament. The PM is both the head of government and sets the legislative agenda. Before 2008, presidential and parliamentary elections occurred in the same year. Since then they have diverged, from which point the parliamentary elections should be used.
Slovenia	Use parliamentary rather than (figurehead) presidential elections.

A data set that implements these corrections and refinements for the EU-28 countries can be found on GitHub.[6] The data (including both the original and corrected versions of the variable) can be downloaded directly into R using the `repmis` package (Gandrud 2015) with the following code:

```
yrcurnt_corrected <- repmis::source_data('http://bit.ly/1EM8EVE')
```

This file has also updated election timing data through 2013.

References

Alt, James, David Dreyer Lassen, and Joachim Wehner. 2014. "It Isn't Just About Greece: Domestic Politics, Transparency and Fiscal Gimmickry in Europe." *British Journal of Political Science* 44 (04): 707-16.

Beck, Thorsten, George Clarke, Alberto Groff, Philip Keefer, and Patrick Walsh. 2001. "New Tools in Comparative Political Economy: The Database of Political Institutions." *World Bank Economic Review* 15 (1): 165-76.

Gandrud, Christopher. 2015. *repmis: Miscellaneous Tools for Reproducible Research with R.* `http://cran.r-project.org/web/packages/repmis/`.

Graphical Presentation of Regression Discontinuity Results

Natália S. Bueno
Yale University
natalia.bueno@yale.edu

Guadalupe Tuñón
University of California, Berkeley
guadalupe.tunon@berkeley.edu

During the last decade, an increasing number of political scientists have turned to regression-discontinuity (RD) designs to estimate causal effects. Although the growth of RD designs has stimulated a wide discussion about RD assumptions and estimation strategies, there is no single shared approach to guide empirical applications. One of the major issues in RD designs involves selection of the "window" or "bandwidth" – the values of the running variable that define the set of units included in the RD study group.[1]

This choice is key for RD designs, as results are often sensitive to bandwidth size. Indeed, even those who propose particular methods to choose a given window agree that "irrespective of the manner in which the bandwidth is chosen, one should always investigate the sensitivity of the inferences to this choice. [...] [I]f the results are critically dependent on a particular bandwidth choice, they are clearly less credible than if they are robust to such variation in bandwidths" (Imbens and Lemieux, 2008, p. 633). Moreover, the existence of multiple methods to justify a given choice opens the door to "fishing" – the intentional or unintentional selection of models that yield positive findings (Humphreys, Schanchez de la Sierra, and van der Windt, 2013).

In this note, we propose a simple graphical way of report-

[6] See `https://github.com/christophergandrud/yrcurnt_corrected/tree/master/data`.

[1] Formally, the window is an interval on the running variable, $W_0 = [\underline{r}, \overline{r}]$, containing the cutoff value r_0. The analysis then focuses on all observations with R_i within this interval. Scholars have developed numerous tools to determine the right window for a given application and estimator (Imbens and Lemieux, 2008; Imbens and Kalyanaraman, 2011, Calonico, Cattaneo and Titiunik, 2015).

[2] This choice is posterior to defining the estimand and choosing an estimator for the causal effect of treatment. While the plots we propose focus

ing RD results that shows the sensitivity of estimates to a wide range of possible bandwidths.[2] By forcing researchers to present results for an extensive set of possible choices, the use of these plots reduces the opportunities for fishing, complementing existing good practices in the way in which RD results are presented. Some empirical applications of RD designs have presented their results using plots that are in some ways similar to the ones we propose. However, this note explores the virtues of the plots more systematically (e.g., in connection with balance tests and the discussion of the RD estimator) and provides code so that scholars can adapt them to their own applications. The following paragraphs describe how RD results are usually reported in two top political science journals and the ways in which the graphs we propose can improve on current practices. An R function to construct these plots for a wide set of applications is publicly available on the online Appendix.

Reporting RD Analyses: Current Practice

How do political scientists report the results of regression-discontinuity designs? We reviewed all papers using RDDs published in the *American Political Science Review* and *American Journal of Political Science*. We surveyed these papers and coded (1) their choice of estimators, (2) whether they present any type of balance test and, (3) if they do, the window(s) chosen for this.

Out of a total of twelve RD papers published in these journals, five report results using a single estimator.[3] Four articles present results for a single window – usually the full sample. The remaining papers present results using multiple windows, but the number and selection of windows are neither systematic nor extensive.[4] Seven papers present some type of balance test, but while researchers often report their main results using a handful of windows, they do not report balance tests for different windows to the same extent.[5]

A Graphical Alternative: An Example

We use electoral data from the U.S. House of Representatives from 1942 to 2008 collected by Caughey and Sekhon (2011) and a regression discontinuity design examining incumbency advantage to illustrate how a researcher can use plots to present RD results in a transparent and systematic way. This application has two advantages for illustrating the use of these plots. First, close-race electoral regression-discontinuity designs are one of the main applications of this type of design in political science – we are thus presenting the plot in one of the most used RDD setups, although researchers can use this type of graph in all sorts of RD designs. Second, the use of close-race RDDs to measure the effect of incumbency advantage has sparked a vigorous debate about the assumptions and validity of these designs in particular settings. Using this application allows us to show an additional advantage of the plots we propose: tests of balance and other types of placebo tests.

Figure 1 plots the estimates of local average treatment effects as a function of the running variable, here vote margin. For example, the first solid back circle represents the average difference in vote share between an incumbent party that won by 0.45% or less and an incumbent party that lost by 0.45% or less is about 11 percentage points. It also reports the average difference in vote share between incumbent parties that won and lost by sequential increases of 0.2% in vote margin between 0.45% to 9.85%. Figure 1 has an additional feature: the solid gray line represents the results using an additional estimator that enables us to compare the effects estimated from two different estimators, across different windows, in the same plot. In this case, we present the estimated effects of party incumbency on vote share using a local linear regression with a triangular kernel, across different bandwidths. Researchers could use different estimators, such as a polynomial regression model.[6] Note, however, that the black circles and the solid gray line represent different quantities. The black circles represent estimates of the average causal effect for the RD study group N. The difference of means estimator $(Y^T - Y^C)$ – the difference between average vote share for an incumbent party minus the average vote share for a non-incumbent party – is unbiased for the average causal effect (τ_{ACE}), represented in equation (1).[7] The gray line presents estimates of the average causal effect precisely at the point of discontinuity (c). We fit a local linear regression with a triangular kernel, within the RDD bandwith, to estimate this limit parameter (τ_{lim}), represented in equation (2).[8]

$$\tau_{ACE} = \frac{1}{N}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)]. \quad (1)$$

on presenting the sensitivity of results to bandwidth choice, we also show how they can be used to explore the sensitivity of results to these other choices. For discussions of estimands in RD design see Dunning (2012) and Calonico et al. (2015).

[3]Polynomial regression is the most popular model: nine papers use a type of polynomial regression, five employ a local linear regression, and three use a difference of means (via OLS). Only one presents results using all three estimators.

[4]Gerber and Hopkins (2011), Ferwerda and Miller (2014), and Eggers et al. (2015) are exceptions–they show the robustness of the main result using plots similar to the one we suggest here.

[5]All of the papers that use local linear regressions also use a type of standard procedure to choose the "optimal" bandwidth – either Imbens and Lemieux (2008) or Imbens and Kalyanaraman (2011).

[6]We present an example of this plot using an fourth-degree polynomial regression in Figure A.1 available at the online Appendix.

[7]See Dunning (2012) and Bueno, Dunning, and Tuñón (2014) for a discussion and proofs.

[8]We use the terms "window" and "bandwidth" interchangeably, since both denote the values of the running variable (r) that define the set of units included in the RD study group. However, in local linear regression with kernel smoothers, bandwidth refers to the width of the kernel.

$$\tau_{lim} = \lim_{r \Downarrow c}[\overline{Y}_i(1)|R_i = r] - \lim_{r \Uparrow c}[\overline{Y}_i(0)|R_i = r] \qquad (2)$$

The key part of the plotting function is the section that produces the estimate for each window. For this, we first pre-specify functions for each estimator. For example, for the difference of means we have:[9]

```
#Difference of mean (using OLS with robust standard errors)
dom <- function(rescaled, treat, outcome){
  model <- lm(outcome~treat)
  est <- NA
  est[1] <- model$coefficients[2]
  est[2] <- sqrt(diag(vcovHC(model,type="HC3"))[2])
  return(est)
}
```

We then include a loop which takes a window value and subsets the data to keep only the observations within that window.

```
   ests <- matrix(NA,length(windows),3) #object to store loop output
   for (i in 1:length(windows)){
       # select data
       temp <- as.data.frame(data[abs_running<=windows[i],])
```

We take this subset of the data to calculate the estimates of interest for that window, here the difference of means. The plot requires that we calculate both the point estimates and the confidence intervals. In these figures, confidence intervals are calculated using a normal approximation and unequal variances are assumed for standard errors in the treatment and control groups. If the researcher wanted to include an additional estimator, the calculation of the estimate for a particular window would also be included in the loop.[10]

```
   ests[i,1:2] <- with(temp, dom(rescaled=rescaled, treat=treat,
                               outcome=outcome))

   if (ci=="95%") CI <- cbind(ests[,1]+1.96*ests[,2],ests[,1]-1.96*
        ests[,2])
   if (ci=="90%") CI <- cbind(ests[,1]+1.64*ests[,2],ests[,1]-1.64*
        ests[,2])
```

As expected, the confidence intervals in Figure 1 become increasingly smaller for results associated with larger vote margins because the number of observations is larger. This increase in the number of observations associated with larger windows can also be reported in the plot, which we do in the upper axis of Figure 1. To include the number of observations as a function of window size, we order the observed values of the outcome of interest according to their value for the runing variable and allow the user to set the different number of observations that she would want to show in the plot. We calculate the number of observations at the specified values of the running variable – then, we add the number of observations to an upper axis in the plot.

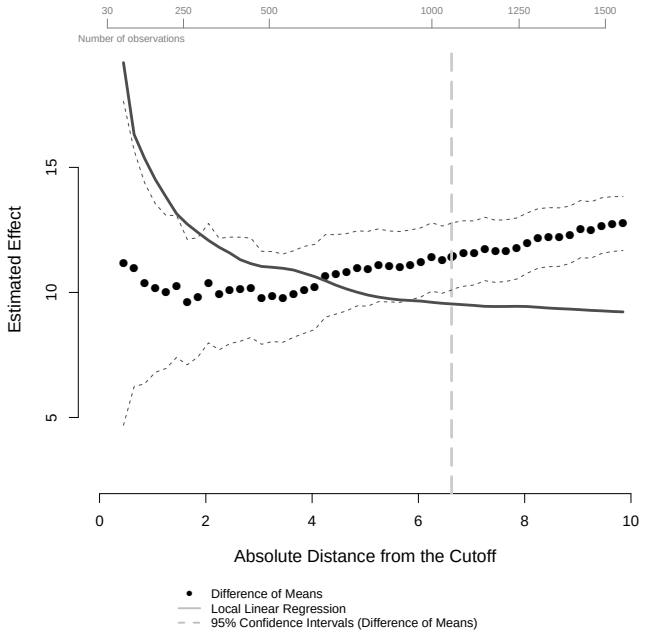
```
   # as an argument in the function, the user defines nr_obs_lab,
   # the labels for the number of observations she would like to
   # include in the plot

   # ordering the vata by the values for the running variable
   data <- as.data.frame(data[order(abs_running),])

   if (nr_obs==T) {
    # binding the labels with the corresponding value for the running
         variable
    nr_obs_lab <- cbind(nr_obs_lab, data$abs_running[nr_obs_lab])
  }

  # Finally, we include an additional axis in the plot
   axis(3, at=c(nr_obs_lab[,2]), labels=c(nr_obs_lab[,1]), cex=.6,
          col="grey50", lwd = 0.5, padj=1, line=1, cex.axis=.7,
          col.axis="grey50")
   mtext("Number␣of␣observations", side=3, col="grey50", cex=.7, adj
        =0)
```

Figure 1: Mean vote share difference between winners and losers by Democratic margin of victory in previous election, U.S. House of Representatives from 1942 to 2008.



Note: Dashed gray line at the optimal bandwidth estimated by the method of Imbens and Kalyanaraman (2011). Difference of means is the average difference in vote share for an incumbent party minus the average vote share for a non-incumbent party. The local linear regression regression uses a triangular kernel.

Figure 2 follows the same logic described for Figure 1 but reports balance tests: each plot shows the effects of incumbency on a pre-treatment covariate as a function of the run-

[9]Note that we compute the difference of means by regressing the outcome variable on a dummy for treatment assignment, with robust standard errors allowing for unequal variances, which is algebraically equivalent to the *t-test* with unequal variances.

[10]Our plot, and the accompanying documented R code, is flexible to incorporating different ways of estimating standard errors and constructing confidence intervals. For a detailed discussion of standard errors and confidence intervals in RD designs, see Calonico et al. (2015).

[11]In these plots, we chose to omit the axis with the number of observations because even though there are different rates of missing observations for covariates, the number of observations for the windows we were mostly interested in did not vary substantially from those in Figure 1.

ning variable.[11] These plots allow for an extensive examination of the sensitivity of balance to different windows, reporting the magnitude of the difference between treatment and control and its confidence interval. For a given identifying assumption (such as continuity of potential outcomes or as-if random assignment near the threshold), Figures 1 and 2 help the reader to evaluate whether or not – or for which window – these assumptions are plausible.

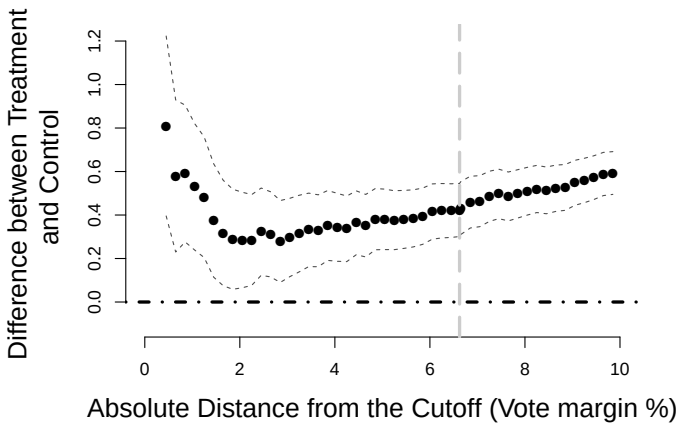
For instance, panel (a) in Figure 2, reports the difference in previous Democratic victories between the incumbent and non-incumbent party and shows a large imbalance, strikingly larger for smaller windows – a point made by Caughey and Sekhon (2011). Panel (b) in Figure 2 shows the difference in voter turnout between treatment and control districts. For the entire set of windows covered by the plot the difference is never statistically different from zero, suggesting that the groups are balanced in terms of this covariate. Note that the size of the difference between treatment and control is much smaller in panel (b) than in panel (a) – also, relatively to the size of the effect of incumbency, the imbalance in panel (a) is substantial.[12]

For ease of presentation, here we present plots for only two pre-treatment covariates. However, analysts should be encouraged to present plots for all pre-treatment covariates at their disposal.[13] Some plots may be more informative than others, for instance, because some pre-treatment covariates are expected to have a stronger prognostic relationship to the outcome. However, presentation of balance plots for the full set of available pre-treatment covariates may reduce opportunities for intentional or unintentional fishing.

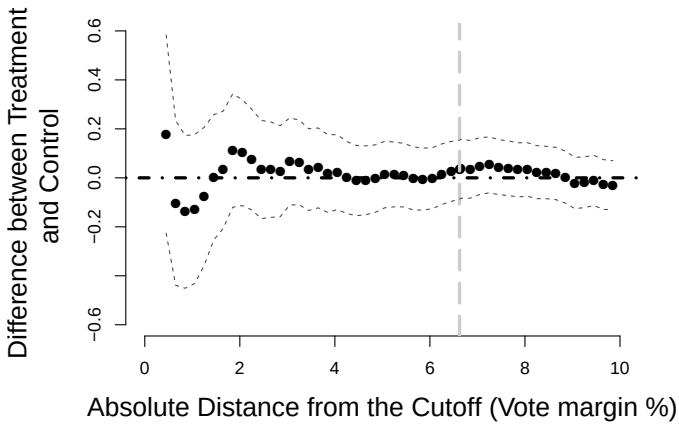
Concluding Remarks

We believe that these simple plots are a useful complement to the standard way in which scholars report results and balance tests from regression-discontinuity designs. They provide a simple visual comparison of how different estimators perform as a function of different windows, communicating the robustness of the results. Moreover, using these plots both for analysis of effects and placebo tests enables an informal visual inspection of how important confounders may be, relative to the size of the effect – this is particularly informative when researchers use pre-treatment values of the outcome variable as a placebo test. However, researchers may also compare the size of treatment effects relative to other placebo tests by using standardized effect sizes across different windows, so that the scale of all plots is comparable. In summary, these plots improve the communication of findings from regression-discontinuity designs by showing readers the results from an extensive set of bandwidths, thus reducing researchers' latitude in presentation of their main findings and increasing the transparency of RD design applications.

Figure 2: Tests for balance: Standardized difference of means of pre-treatment covariates by Democratic margin of victory (95% confidence intervals), U.S. House of Representatives, from 1942 to 2008.



(a) Democratic Win in $t-1$



(b) Voter Turnout %

Note: Dashed gray line at the optimal bandwidth estimated by the method of Imbens and Kalyanaraman (2011). Difference of means is the average difference in vote share for an incumbent party minus the average vote share for a non-incumbent party.

References

Bueno, Natália S., Thad Dunning, and Guadalupe Tuñón. 2014. "Design-Based Analysis of Regression Discontinuities: Evidence From Experimental Benchmarks. *Paper Presented at the 2014 APSA Annual Meeting.*

[12] See Figure A.2 in the online Appendix for the standardized effect of incumbency on vote share.

[13] An extensive set of balance plots for pre-treatment covariates in Caughey and Sekhon (2011) can be found in Figure A.3 of the online Appendix.

Calonico, Sebastian, Matias D. Cattaneo and Rocio Titiunik. 2015. "Robust Nonparametric Confidence Intervals for Regression-discontinuity Designs." *Econometrica* 86 (2): 2295-326.

Caughey, Devin and Jasjeet S. Sekhon. 2011. "Elections and the Regression Discontinuity Design: Lessons from Close US House Races, 1942-2008." *Political Analysis* 19 (4): 385-408.

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: a Design-Based Approach.* New York: Cambridge University Press.

Eggers, Andrew C., Olle Folke, Anthony Fowler, Jens Hainmueller, Andrew B. Hall and James M. Snyder. 2015. "On the Validity of the Regression Discontinuity Design For Estimating Electoral Effects: New Evidence From Over 40,000 Close Races." *American Journal of Political Science* 59 (1): 259-74.

Ferwerda, Jeremy and Nicholas L. Miller. 2014. "Political Devolution and Resistance to Foreign Rule: A Natural Experiment." *American Political Science Review* 108 (3): 642-60.

Gerber, Elisabeth R. and Daniel J. Hopkins. 2011. "When Mayors Matter: Estimating the Impact of Mayoral Partisanship on City Policy." *American Journal of Political Science* 55 (2): 326-39.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis* 21 (1): 1-20.

Imbens, Guido and Karthik Kalyanaraman. 2011. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *The Review of Economic Studies* 79 (3):933-59.

Imbens, Guido W. and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142 (2): 615-35.

Why Can't We Just Make Up Instrumental Variables?

Thomas B. Pepinsky
Cornell University
pepinsky@cornell.edu

Introduction

Valid instrumental variables are hard to find. The key identification assumption that an instrumental variable Z be correlated with an endogenous variable X, but not causally related to Y, is difficult to justify in most empirical applications. Even worse, the latter of these requirements must be defended theoretically, it is untestable empirically. Recent concerns about "overfishing" of instrumental variables (Morck and Yeung 2011) underscore the severity of the problem. Every new argument that an instrumental variable is valid in one research context undermines the validity of that instrumental variable in other research contexts.

Given such challenges, it might be useful to consider more creative sources of instrumental variables that do not depend on any theoretical assumptions about how Z affects X or Y. Given an endogenous variable X, in a regression $Y = \beta X + u$, why not simply generate random variables until we discover one that is, by happenstance, correlated with X, and then use that variable Z_{random} as an instrument? Because Z_{random} is generated by the researcher, we can be absolutely certain that there is no theoretical channel through which it causes Y. And because we have deliberately selected Z_{random} so that it is highly correlated with X, weak instrument problems should not exist. The result should be an instrument that is both valid and relevant. Right?

If it were only so easy. In Figure 1 below I show what would happen if we used random instruments, generated in the manner just described, for identification purposes. The data generating process, with notation following Sovey and Green (2011), is

$$Y = \beta X + \lambda Q + u \quad (1)$$

$$X = \gamma Z + \delta Q + e \quad (2)$$

X is the endogenous variable, Q is an exogenous control variable, and Z is an instrumental variable. Endogenity arises from correlation between u and e, which are distributed multivariate standard normal with correlation $\rho = .5$. For the simulations, $\beta = 5$, with $\lambda = \gamma = \delta = 1$.

Figure 1 shows the distribution of estimates $\hat{b}$ from 100 simulations of the data-generating process above. In each simulation, I generate each variable in Equations 1 and 2, and then repeatedly draw random variables until I happen upon one that is correlated with the randomly generated X at the 99.9% confidence level. I use that variable Z_{random} as an instrument to estimate $\hat{b}_{IV(Fake)}$, which I compare to the naive OLS estimates $\hat{b}_{OLS}$ and estimates using the real instrument Z from Equation 2, $\hat{b}_{IV}$.

Clearly, OLS estimates are biased. But more importantly, the "fake" IV estimates are biased and inefficient, while "true" IV estimates using Z are accurate. What has happened to the fake IVs?

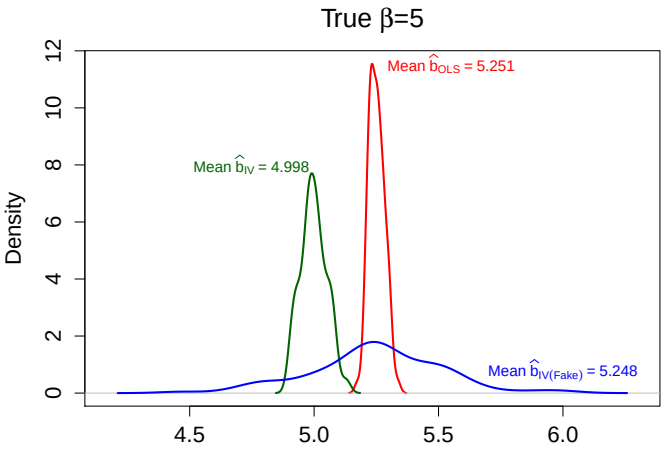


Figure 1: Random Instruments

How We Talk about Instruments

Understanding why this procedure fails to generate valid instruments requires a more precise understanding of the identification assumptions in instrumental variables regression than we normally find. The problem can be seen immediately in the introductory paragraph of this essay. My language followed common conventions in the ways that applied researchers *talk* about instrumental variables rather than the precise technical assumptions found in econometrics textbooks. I described the two requirements for instrumental variable as follows: an instrumental variable is some variable Z that is

1. correlated with an endogenous variable X
2. not causally related to Y

These two statements are close, but not exact. The precise requirements (see e.g. Wooldridge 2002, 83-84) are that Z is

1. partially correlated with X, conditional on Q.
2. conditionally independent of u, meaning that $\text{Cov}[Z, u] = 0$.

Precision matters here. Assumption 1, usually referred to as the "relevance" of an instrument, emphasizes that X and Z must be partially correlated *conditional on* the exogenous variables Q. The procedure described above simply looked for an unconditional correlation between X and Z, which may or may not suffice.

This is, however, a minor problem. We could follow Stock and Watson's (2003, 350) advice, and search for variables Z_{random} that satisfy the rule of thumb that the F statistic for a test that $\gamma = 0$ in equation (2) is greater than 10. The more fundamental assumption that prevents any type of randomized instrument from serving as a valid instrument for X is the second assumption of "validity," or that $\text{Cov}[Z, u] = 0$.

This matters because u is an unobservable quantity that reflects a *theoretical* claim about the data generating process that produces Y. Consider what happens when we generate a Z_{random} that is partially correlated with X. X is a function of e, which is in turn correlated with u—it is this correlation between u and e that motivates the instrumental variables approach to begin with. This chain of association $Z_{random} \rightarrow X \rightarrow e \rightarrow u$ makes it very likely that any Z_{random} will also be correlated with u, violating the assumption of validity. That chain of correlations from Z to u is what explains the biased and inefficient estimates for $\hat{b}_{IV(Fake)}$ in Figure 1.

Notice that not every Z_{random} will be correlated with u, but we can never know if any particular random instrument is. Consider would it would take to verify the validity of a random instrument. If we happened to observe the error term u, then we could discard the majority of candidates for Z_{random} that happened to also be correlated with u, and select only valid random variables that happened to be partially correlated with X and uncorrelated with u. Figure 2 below shows two examples of binary Z_{random} that are highly correlated with X, but in the top case, Z_{random} is also highly correlated with u, whereas in the bottom case, Z_{random} and u are unrelated.

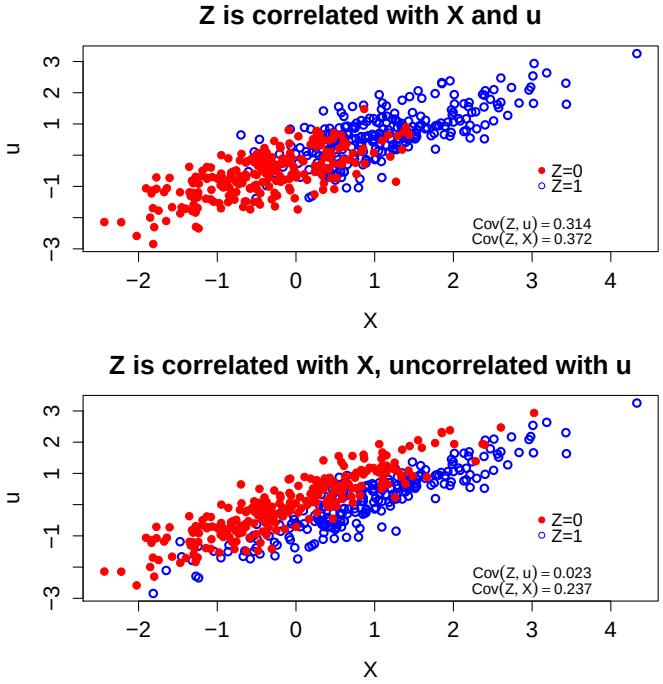


Figure 2: Comparing Z, X, and u

Again, if it were possible to observe u, then it would be possible to choose only those instrumental variables Z that were uncorrelated with u. But the error term u is unobservable *by definition*. So it is actually true that random instruments could work if we could check their correlation with u, but since we cannot, the point is irrelevant in practice. Without access to u, instrument validity is always a theoretical claim.

It is helpful to relate this conclusion about randomly generating instruments for observational data to randomized assignments to treatment offers in an experimental design with noncompliance. In such designs, assignment to treatment, or "intent to treat," is an instrument for actual treatment. Here, randomization does ensure that the binary Z is an instrument for treatment status D, but only with appropriate assumptions about "compliers," "never-takers," "always takers," and "defiers" (see Angrist, Imbens and Rubin 1996). Gelman and Hill (2007) give the example of an encouragement design in which parents are randomly given inducements to expose their children to Sesame Street to test its effect on educational outcomes. If some highly motivated parents who would never expose their children to Sesame Street regardless of receiving the inducement or not (never-takers) use these inducements to "purchase other types of educational materials" (Gelman and Hill 2007, 218) that change their children's educational outcomes, then the randomized inducement no longer estimates the average treatment effect among the compliers. The point is that randomizing treatment still requires the assumption that $\text{Cov}[Z, u] = 0$, which in turn requires a model of Y, for randomization to identify causal effects.

Precision, Insight, and Theory

How we as political scientists talk and write about instrumental variables almost certainly reflects how we think about them. In this way, casual, imprecise language can stand in the way of good research practice. I have used the case of random instruments to illustrate that we need to understand what would make an instrument valid in order to understand why we cannot just make them up. But there are other benefits as well to focusing explicitly on precise assumptions, with import outside of the fanciful world of thought experiments. To begin with, understanding the precise assumption $\text{Cov}[Z, u] = 0$ helps to reveal why all modern treatments emphasize that the assumption that Z is a valid instrument is not empirically testable, but must be theoretically justified. It is not because we cannot think of all possible theories that link Z to Y. Rather, it is because because the assumption of validity depends on the relationship between the unobservable error term u and any candidate instrument Z.

Precision in talking and writing about instrumental variables will produce more convincing research designs. Rather than focusing on the relationship between Y and Z in justifying instrumental variables as source of identification, it may be more useful to focus instead on the empirical model of Y. The most convincing identification assumptions are those that rest on a theory of the data-generating process for Y to justify why $\text{Cov}[Z, u] = 0$. Recalling that u is not actually a variable, but might be better understood as a constructed model quantity, helps to emphasize that the theory that justifies the model of Y is also the theory that determines the properties of u. Indeed, without a model of the Y it is hard to imagine how u would relate to any particular instrument. A good rule of thumb, then, is *no theory, no instrumental variables*.

Thinking explicitly about u rather than Y and the model that it implies also helps to make sense of of what are sometimes called "conditional" instrumental variables. These are instruments whose validity depends on inclusion of an exogenous control variable (in the above example, Q) in the instrument variables setup. Conditional instruments are well understood theoretically: Brito and Pearl (2002) provide a framework for what they call "generalized" instrumental variables using directed acyclic graphs, and Chalak and White (2011) do the same in the treatment effects framework for what they call "extended" instrumental variables. Yet the imprecise short-hand description of the exclusion restriction as "Z is unrelated to Y except through X" obscures these types of instruments. The observation that $u = Y - \beta X - \lambda Q$ immediately clarifies how control variables Q can help with the identification of β even if X is still endogenous when controlling for Q. Once again, a theory of the data generating process for Y is instrumental for an argument about whether Q can help to achieve identification.

Still another reason to pay close attention to the expression $\text{Cov}[Z, u]$ is that it sheds light on how over identification tests work, and also on their limits. I have repeatedly stipulated that identification assumptions are not empirically testable, but it of course true that in an overidentified model with more instruments than endogenous variables, it is possible to reject the null hypothesis that all instruments are valid using a Hausman test. This test works by regressing the instruments $\mathbf{Z}$ and exogenous variable Q on $\hat{u}$, the two-stage least squares residuals when using $\mathbf{Z}$ as instruments for X. With sample size N, the product $NR^2_{\hat{u}}$ is asymptotically distributed χ^2_{K-1}, where $K \geq 2$ is the number of excluded instruments. Assuming homoskedasticity—there are alternative tests that are robust to heteroskedasticity—this provides a test statistic against the null that $\text{Cov}[\mathbf{Z}', u] = 0$. It is true that a large test statistic rejects the null of overidentification, but small values for $NR^2_{\hat{u}}$ could have many sources: small N, imprecise estimates, and others that have nothing to do with the "true" correlation between $\mathbf{Z}$ and u. And even rejecting the null need not reject the validity of the instruments in the presence of treatment effect heterogeneity (see Angrist and Pischke 2009, 146).

Finally, a precise description of identification assumptions is important for pedagogical purposes. I doubt that many rigorous demonstrations of instrumental variables actually omit some discussion of the technical assumption that $\text{Cov}[Z, u] = 0$, but this is presented in many different ways that, in my own experience, are confusing to students. Even two works from the same authors can differ! Take Angrist and Pischke, from *Mastering 'Metrics*: Z must be "unrelated to the omitted variables that we might like to control for" (Angrist and Pischke 2014, 106). And from *Mostly Harmless Econometrics*: Z must be "uncorrelated with any other determinants of the dependent variable" (Angrist and Pischke 2009, 106). It is easy to see how these two statements may cause confusion, even among advanced students. Encouraging students to focus on the precise assumptions behind instrumental variables will help them to understand what exactly is at stake, both theoretically and empirically, when they encounter instrumental variables in their own work.[1]

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 91 (434): 444-55.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2014. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton: Princeton University Press.

Brito, Carlos and Judea Pearl. 2002. "Generalized Instrumental Variables." In *Uncertainty in Artificial Intelligence*, Proceedings of the Eighteenth Conference. eds. A. Darwiche and N. Friedman. San Francisco.

Chalak, Karim and Halbert White. 2011. "Viewpoint: An Extended Class of Instrumental Variables for the Estimation of Causal Effects. *Canadian Journal of Economics* 44 (1): 1-51.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Morck, Randall and Bernard Yeung. 2011. "Economics, History, and Causation. *Business History Review* 85 (1): 39-63.

Sovey, Allison J. and Donald P. Green. 2011. "Instrumental Variables Estimation in Political Science: A Readers' Guide. *American Journal of Political Science* 55 (1): 188-200.

Stock, James H. and Mark W. Watson. 2003. *Introduction to Econometrics*. New York: Prentice Hall.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: The MIT Press.

Call for Papers: The Political Methodologist Special Issue on Peer Review

Justin Esarey
Rice University
justin@justinesarey.com

The Political Methodologist is calling for papers for a special issue of TPM concerning the process of peer review in political science!

Peer review is something that every political scientist in the field will be subject to and asked to perform, but is a task for which almost no one receives formal training. Although some helpful guidelines exist in the literature,[1] I believe there is still considerable heterogeneity in how people think about the peer review process and that the community would benefit from discussing these views. Moreover, new developments in the discipline raise new questions about the review process (e.g., the degree to which journals and reviewers have a responsibility to ensure replicability and reproducibility).

A wide variety of topics would fit well into this special issue, including (but not exclusive to):

- how one should write a review, including and especially what constitute fair criteria for evaluation, and what criteria are unfair
- what is the reviewer's role in the process: Quality Assurance? Error Checking? Critical Commentary? or what?
- how one should respond to a review when invited to revise and resubmit (or rejected)
- the role that peer review should play in error checking/replication/verification
- the "larger view" of how peer review does or should contribute to (political) science
- the role of editorial discretion in the process, and how editors should regard reviews

[1]Author's note: Thanks to Justin Esarey for helpful feedback on an earlier draft.
[1]See Miller et al. (2013) and Lucey (2013).

Submissions should be between 2000-4000 words (although shorter submissions will also be considered), and should be sent to thepoliticalmethodologist@gmail.com by December 1, 2015. Accepted articles will be featured on our blog, and also in the print edition of *The Political Methodologist*.

If you are interested in contributing to the special issue and would like to talk about prospective contributions before writing/submitting, please feel free to contact me (justin@justinesarey.com).

References

Lucey, Brian. 2013. "Peer Review: How to Get It Right – 10 Tips." *The Guardian* September 27. `http://www.theguardian.com/higher-education-network/blog/2013/sep/27/peer-review-10-tips-research-paper` (July 2, 2015).

Miller, Beth, Jon Pevehouse, Ron Rogowski, Dustin Tingley, and Rick Wilson. 2013. "How To Be a Peer Reviewer: A Guide for Recent and Soon-to-be PhDs." *PS: Political Science & Politics* 46 (1): 120-3.

Rice University
Department of Political Science
Herzstein Hall
6100 Main Street (P.O. Box 1892)
Houston, Texas 77251-1892

The Political Methodologist is the newsletter of the Political Methodology Section of the American Political Science Association. The support of the Department of Political Science at Rice University in helping to defray the editorial and production costs of the newsletter is gratefully acknowledged.

Subscriptions to *TPM* are free for members of the APSA's Methodology Section. Please contact APSA (202-483-2512) if you are interested in joining the section. Dues are $29.00 per year for students and $44.00 per year for other members. They include a free subscription to *Political Analysis*, the quarterly journal of the section.

Submissions to *TPM* are always welcome. Articles may be sent to any of the editors, by e-mail if possible. Alternatively, submissions can be made on diskette as plain ascii files sent to Justin Esarey, 108 Herzstein Hall, 6100 Main Street (P.O. Box 1892), Houston, TX 77251-1892. LaTeX format files are especially encouraged.

TPM was produced using LaTeX.

THE SOCIETY FOR
POLITICAL METHODOLOGY
ad indicia spectate