Protect Yourself from p-Hacking: 7 Things to Do to Avoid Committing Scientific Malpractice

Frederick Solt
frederick-solt@uiowa.edu

Yue Hu
yue-hu-1@uiowa.edu

Kevan Hudson
kevan-hudson@uiowa.edu

Jungmin Song
jungmin-song@uiowa.edu

Dong 'Erico' Yu
dong-yu@uiowa.edu

November 16, 2015

Abstract

Replication crisis

LaCour scandal

p-Hacking

malpractice not always so blatent or intentional: confirmation bias, garden of forking paths (see http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf)

Introduce **?**, perhaps noting the press attention it has received (e.g., http://www.psmag.com/health-and-behavior/five-studies-bernie-sanders-says-the-rich-are

1 Ensure Reproducibility

Reproducibility as bare minimum for replication; DA-RT APSA guidelines

script all work

packrat and checkpoint packages in R; version command in Stata

quote **?** replication materials

Table 1 and 2 cannot be reproduced exactly

Table 3 cannot be reproduced at all: more parameters than observations

2 Work in Public

Researching transparency is always an important criterion for scientific researches, including qualitative and quantitative political scientist studies (**??**). One way to match this standard is to open every data managing and estimating step in public. It may sound time consuming, unsafe, and unnecessary, especially in the views of researchers who already provide "replication file" ((e.g., **?**)). But, as shown in the previous section, the actual problem is the offered "replication file" does not always work ((e.g., again, **?**)). In this case, other researchers can go directly to the publicly accessed documents and files which the researcher used to conduct the analyses to check what was wrong in the replication file.

Moreover, the contemporary computer and internet technologies also make working in public very easy and safely. Taking GitHub as an example,[1] what

[1]There are alternative agents, such as SVN and Gitlab, offering similar services, although there might be slight difference in operations and compatible software, and most of them are free to register and use.

requires the researcher to do is simply to spend 10 second to build a repository and push the new committed file (viz. the updated file) after every important change in the analysis, as we did for this project. At the same time, the study per se is also safe as long as the paper and data is not accessed, which is actually also not required before the study is published.

Another way to do publicly study is to preregister the research. It is in the same line as pushing staged analyses onto GitHub, but under a more specific supervision of the academia of political science. The preregistration asks researchers to publish their research plan prior to conducting the analyses. The purpose of this is mainly to avoid the result-oriented researches in which the researcher manipulate the data or change the theory and hypotheses based on the empirical results they have.[2]

3 Examine All Available Data

examine as much relevant evidence as possible

discuss Figure **??**

discuss Figure **??**

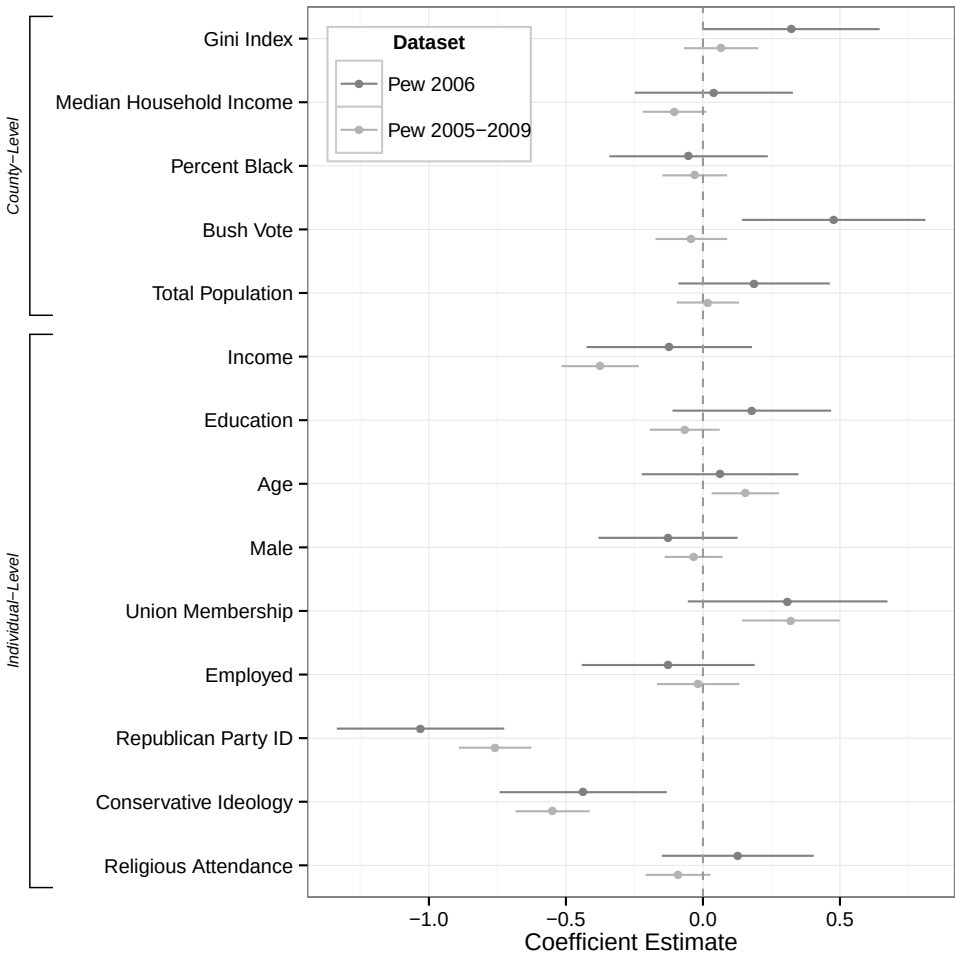
An important step in the process of conducting sound research is striving to utilize all available data. Researchers should seek to conduct their analysis on the entirety of data that they have at their disposal. Doing so provides a number of benefits: it increases the likelihood that the sample utilized captures the true distribution of the underlying population, and it affords greater leverage in testing the implications of one's hypothesis. While the issue of selection bias cannot be avoided simply by including all available data, limiting analysis to a particular dataset, particularly when alternatives are available, may cast doubt on the inferences drawn from that analysis. By including all relevant data researchers are better able to observe the implications of their theory, thus providing greater support for the hypotheses they advance. If research is limited to a particular source of data, the findings may be called into question. The limited data may provide evidence of a relationship that is not present in a larger, more representative sample. We can draw an example of the dangers of not including all relevant data from **?**. In an attempt to test some underlying assumptions of their theory the authors rely on 2006 Pew Research Center dataset due to its "unique set

[2]More discussions about preregistration were in the "Symposium on Research Registration", of Winter 2013 issue of *Political Analysis*

of questions tapping perceptions of economic hierarchy and inequality and respondents perception of their own position within such a hierarchy" (**?**). As presented by the authors, this dataset provides the only source of information on responses to the question whether or not an American thinks of America as being divided into haves and have-nots, and whether they think of themselves as being haves or have-nots. Employing this dataset the authors find further support for their theory, in situations of higher inequality respondents are more likely to believe that America is divided in such a way, and the poor are more likely to identify themselves as have-nots. In reality, these questions are not unique to the 2006 dataset, but are instead present in each of the surveys the authors used in their earlier analysis. Perhaps it is by coincidence that, as shown in Figure **??**, the coefficient of interest only achieves statistical significance when using the 2006 data. As illustrated, no other dataset produces a statistically significant coefficient for Gini according to the authors' model. This provides a clear illustration of the importance of including all relevant data; failure to do so can lead to biased results. The authors' use of this severely truncated data has implications beyond the coefficient of interest as well. While Figure (INSERT FIGURE NUMBER) clearly demonstrates that a more careful inclusion of all data produce results that run counter to the findings of Newman et al, including all available data drastically changes the entre model, not merely the coefficient for Gini. Figure **??** provides estimates of the coefficients from Newman et al's Table 2 (p. 336) with a sample that includes data from the 2005, 2006, 2007, and 2009 surveys they use earlier in their article. When all relevant data is included, the results are drastically different. Not only does the primary variable of interest (Gini coefficient) lose statistical significance, but others do as well. Having voted for Bush is no longer a statistically significant predictor of believing America is divided into the haves and have-nots, but income becomes strongly negative and significantly associated with the same belief. Additionally, union membership gains statistical significance, indicating that belonging to a union increases the likelihood an individual perceives that have/have-not division. Ultimately, Figure (INSERT NUMBER) provides graphical representation of the dangers of not including all available data. By limiting their analysis to the sole dataset that produced a statistically significant coefficient for inequality, the authors have disguised the true relationship in order to support their theory. A properly crafted analysis reveals findings that are far less surprising; the wealthy are less likely to see America as divided into the haves and have-nots, while union members are more likely

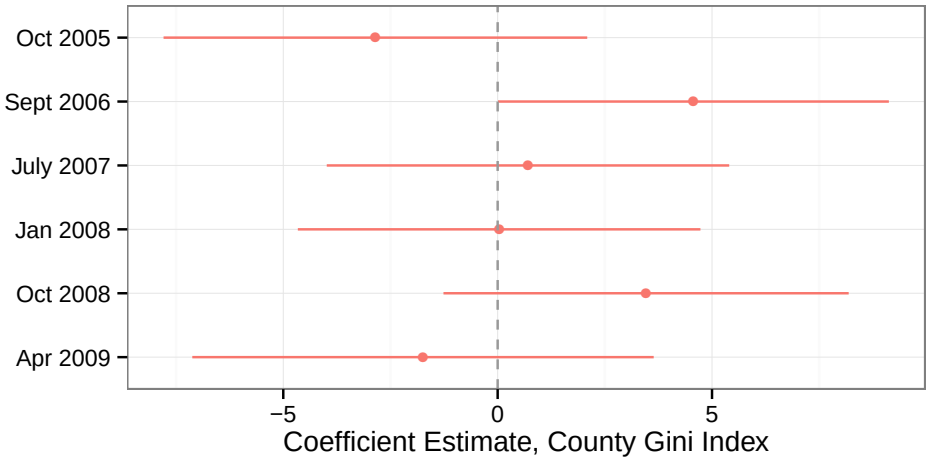
to do so. These findings counter the primary argument advanced by Newman et al, and lend strong evidence to the claim that "we should be willing to take whatever information we can acquire so long as it helps us learn about the veracity of our theory," while illustrating the pitfall of picking and choosing data that confirm our theory, while ignoring data that does not (**?**).

Figure 1: Local Inequality and the Perception of America as Divided into 'Haves' and 'Have-Nots': Results Using All Available Data



Notes: Results from replications of the model presented in Table 2 of **?** on the 2006 Pew survey analyzed in that article and on pooled data from the six Pew surveys that included the same item and were conducted in the time period the article examines. The statistically significant result for county income inequality in the 2006 survey presented in that article is not evident when all of the available data are examined.

Figure 2: Local Inequality and the Perception of America as Divided into 'Haves' and 'Have-Nots': Results Using Each Available Dataset



Notes: Results for county income inequality from replications of the model presented in Table 2 of **?** on data from each of six available surveys conducted in the in the time period examined in that article. Of the six surveys, the only one that yields a statistically significant result is the 2006 survey presented in the article.

4 Use Consistent Measures

text

5 Wrangle Data with Care

need to be really careful: double-check! Also need to be transparent.

merging: data on Bush share of vote don't match

As shown in Table **??**, a quick glance at the first handful of counties analyzed in Table 1 and Table 2 (**?**) reveals that something went wrong when information on the Bush share of the vote in the 2004 election was merged into the datasets: across all counties examined in both tables, fewer than 10% have matching data, even when rounded to two decimal places. (The data analyzed in Table 1 correspond that available from other sources, so it appears that it is the Table 2 dataset that is problematic.)

Table 1: Mismatched Data on Bush Vote, from Replication Data

County	Table 1 Data	Table 2 Data
Baldwin, AL	0.76	0.79
Calhoun, AL	0.66	0.66
Chambers, AL	0.58	0.56
Cherokee, AL	0.65	0.65
Choctaw, AL	0.54	0.51
Clarke, AL	0.59	0.57

Notes: **?** replication data on the share of the vote won by Bush in the 2004 presidential election. The first six counties, when listed alphabetically by state and county, are shown; they reveal that the data employed in Table 2 only occasionally matches that employed in Table 1, even when rounded to two digits. Overall, these data match for fewer than 10% of all counties.

coding and recoding: NJL's five point party id scale collapses leaners and weak partisans (not weak and strong partisans, and not leaners and 'true' independents). Should really use the full seven point scale; no reason to throw away that information (or to deviate from common practice)

unemployment is mismeasured in 2005, 2007, and 2009 in Table 1 due to missing employ2 variable—all of those who are not working (students, retired, etc.) are coded as unemployed

6 Multiply Impute Missing Data

Summary Paragraph

In political science study, a common way to contaminate the data and undermine the validity is missing data growing non-response in Pew survey (Curtain, et.al, 2005;);

Reason: poor survey design(literature needed); private information (income;); sensitive question—influence of social desiablity (underreported abortion of black: Jagannathan, 2001; nonrespond and overreponse in female president question; Streb, 2008; nonresponse in religious: Hadway et.al, 1993); survey forms (Chang RDDv.s. Internet; 2009; Amazon-Turk, Berin-

sky, 2012)

To deal with missing data: first identify missing mechanisms: (1) missing completely at random (MCAR): the probability of missingness is the same for all unit; that is, each survey respondent choose not to answer the question based on on coin flip (2)missing at random (MAR), and nonignorable or (King et. al, 2001): the probability whether the survey question is answered may depend on the other factor which are observable in the data: for example, an independents are more tend to decline to answer partisan identification question. (3)nonignorable (NI): because of the unobserved value of the missing response: high income people tend to conceal their real income, and other variables in the data cannot predict which respondent have high income.

Strategy: to all: 1. listwise deletion or complete-case analysis; + Available case study (use the distribution of other observable variables)

Disadvantages: (1) lead to biased estimates, especially when missing values differ systematically from the completed observed cases; (2) relatively, the standard error can be sensitive due to the original sample size and the deleted case size; (3) in ACS, may lead to omission of a variable hat is necessary to satisfying the assumptions necessary for desired casual inference.

2. Non-response weighting: need to add literatures (for MAR and MCAR)

3. Simple imputation: With high certainty: (1)mean, (2)last value carried forward;

With uncertainty: (3) using information from related observation (for example, in GSS, using reported occupation types and its mean annual salary to infer income; or in SIS, use reported working months to infer income)

Disadvantages: (1) mean, last value, or other singly imputed method bias, especially when the size of missing observation is relatively large; distorting the actual distribution (Gelman, 2006)

(2) Inferred from other related observation: may not to impute all missing data

4. Random Imputation with a single variable or multiple variables

5. Multiple Imputation:

Definition: imputing missing values for each missing case with different imputations to reflect uncertainty levels, and creating a complete data set. (King.et.al, 2001). For example, if assume the missing data of a specific variable is MAR, indicating other observable variables can infer useful information to predict the missing cases, conditional on whichever imputation model adopted.

(1)Multivariate regression:

a. use continuous model to impute missing discrete response.Example: modeling the data as continuous (transfer discrete value into continuous), and imputing continuous values (Gelman and King, 1998)

(2)

Specific to Newman's paper

church attendance—all missing are simply assigned "once or twice a month"

income, a variable of interest, is missing for over 10% of the sample, but values are mysteriously single-imputed (where did these values come from? they aren't meaningful–they fall between categories)

ideology, partyid also single-imputed, it seems

Church attendance: MAR

income: MAR

ideology: MAR

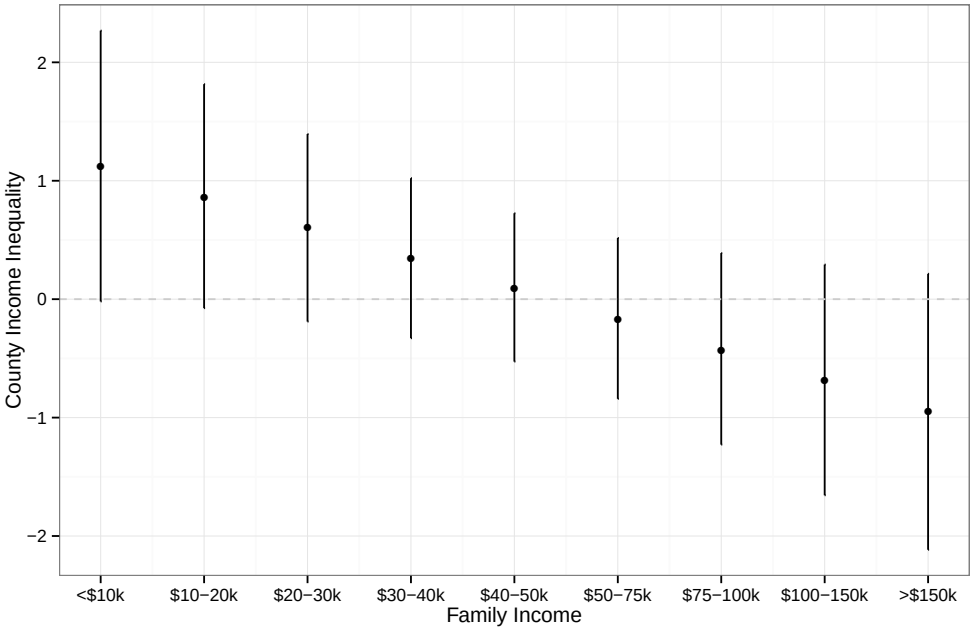
Thus, mulitple imputation is possible and necessary

missing data should be multiply imputed (e.g., **?**)

7 Plot Interaction Terms

It has been well known for over a decade that models containing multiplicative interaction terms require particular care in interpretation (see, e.g., ????).

Figure 3: Logit Coefficients of Local Income Inequality by Respondent Income: Table 1, Model 1, From Replication Data



Notes: The coefficient for county income inequality fails to reach statistical significance for any observed level of respondent family income.