Simulating Representation: The Devil's in the Detail

forthcoming, Research & Politics

Martin Gilens

This note is a response to Omar Bashir's 2015 paper "Testing Inferences about American Politics: A Review of the "Oligarchy" Result," *Research & Politics, 2*(4).

Although he addresses a number of aspects of our work on influence over government policy (Gilens and Page, 2014), the most novel contribution of Omar Bashir's paper is the simulation he reports on pp.3-5. Bashir writes: "I employ a simulation to investigate whether the authors' linear regression of a dichotomous dependent variable on highly correlated independent variables can generate extreme but incorrect results." (p.3) The particular result in question is the tiny coefficient we report for the influence of average citizens on policy outcomes: b=.03 (se=.08) compared with b=.76 (se=.08) for economic elites.

Bashir's simulation is problematic on a number of counts and the central conclusions that he draws from his simulation are not supported. First, Bashir's claim that a strong correlation between predictors in our model violates a statistical assumption of our estimation procedure is mistaken. Second, the simulated data that Bashir constructs do not match our actual data in the ways that he claims. Finally, the key result that Bashir reports from his analysis of his simulated data derives not from any unreliability in our estimation procedure (as Bashir claims), but from errors in the construction of his simulated data.

Correlated predictors

A high correlation between independent variables in a multiple regression increases the uncertainty around the coefficient estimates and results in larger standard errors than would otherwise be the case. But it does not bias the coefficients or the standard errors, nor does it "violate an assumption of both linear and logistic regression" as Bashir claims (p.3). As Achen (1982, p.82) writes:

> Beginning students of methodology occasionally worry that their independent variables are correlated—the so-called multicollinearity problem. But multicollinearity violates no regression assumptions. Unbiased, consistent estimates will occur, and their standard errors will be correctly estimated. The only effect of multicollinearity is to make it hard to get coefficient estimates with small standard errors.

The standard errors we reported in our paper (.08 for both economic elites and average citizens) would have been smaller if the preferences of middle- and high-income American were less strongly correlated. But they are clearly small enough, even so, to

easily distinguish the tiny impact of average citizens from the large impact of the well to-do.[1]

To ensure that the (asymptotically distribution free) analytic standard errors we reported in Gilens and Page 2014 table 3 model 4 were not distorted by our use of a dichotomous dependent variable, and to address potential concerns that our strongly correlated predictors make our coefficient estimates unstable, we used the AMOS structural equation modeling program to calculate bootstrap standard errors based on 2,000 random draws. The bootstrap standard error matched our reported analytic standard errors nearly exactly.

In short, there is no a priori reason to doubt the standard errors we reported or to think that the high correlation between preferences of average citizens and economic elites biased our results. This undercuts the motivation for Bashir's simulation. But more importantly, the simulation itself is flawed in ways that undermine the conclusions that Bashir draws.

[1] Although correlated true scores among the independent variables do not bias regression coefficients, correlated measurement error in the predictors do. This is the reason we use a structural equation model in Gilens and Page (2014). In Gilens (2012) I conducted analyses to measure the magnitude and nature of the random and correlated error in my data. I report these along with the results of alternative approaches to dealing with correlated error (see especially the appendix to chapter 3). Among the findings reported there are the very similar levels of measurement error for low, middle, and high income respondents (p.88) and the very similar levels of correlated error across pairs of income groups (p.256).

Bashir's simulation

As mentioned above, Bashir conducts a simulation "…to investigate whether the authors' linear regression of a dichotomous dependent variable on highly correlated independent variables can generate extreme but incorrect results." Of course any empirical analysis based on a sample of data can generate extreme results due to sampling error; the question is how frequently an extreme or misleading results is expected to occur.

To answer this question, Bashir creates 2,500 simulated datasets based on a multivariate random distribution constructed to match the characteristics of the actual data used in our study as closely as possible, but with the key difference that Bashir chooses a larger "true" coefficient for average citizens (.41 rather than .03) to use in generating the simulated outcome variable. The logic of the simulation, then, is to see how likely it is that one might get results similar to those we report, if in fact the true impact of average citizens was fairly substantial (i.e., with a coefficient of .41 in the multivariate model we estimate).

As he notes in his article, one would expect the mean result of the analyses based on these 2,500 simulated datasets to reproduce the "true" coefficients used to generate the data. The key finding from Bashir's simulation is that a substantial proportion of the simulated datasets (over 20%) produce a coefficient for average citizens that is far from the "true" coefficient of .41 and close to the "near zero" coefficient we report based on our actual data. Bashir concludes from this that even if the world were such that average citizens had much more influence than we claim, results similar to ours could be expected in over 20% of samples.

However, there are two significant errors in Bashir's simulation that account this

result. First, the way he constructed the simulated outcome variable changed the "true" coefficients away from those that he intended to base his replication on. Second, in generating the simulated datasets, he filtered the results such that only about 5% of the simulated datasets were retained (i.e., he actually generated about 50,000 simulated datasets and then discarded about 47,500 of them).

The first of these two problems results from the fact that Bashir first constructed a continuous outcome variable based on his chosen "true" coefficients (e.g., .41 for average citizens) and *only then* dichotomized that variable to mirror our observed dichotomous outcome. The unintended consequence of this procedure is that the "true" coefficients are all reduced in size from their intended magnitude before the data are put to use.[2]

After dichotomizing the outcome variable, Bashir then filters the results, retaining only those datasets in which the bivariate coefficients for each of the three predictors fall within a narrow range of the observed bivariate coefficients we report from our actual data.[3] Thus the retained datasets have, by Bashir's construction, a substantially larger multivariate coefficient than the actual data but a very similar bivariate coefficient. This filtering procedure further reduces the size of the multivariate coefficient for average citizens—the key estimand of interest.

As a result of these two procedures, Bashir's otherwise accurate claim that "One would expect the estimated coefficients produced by subsequent regression to be close to the true coefficients chosen to seed each iteration," no longer holds—not due to any bias or unreliability in the estimation procedure, but due to the way the simulated data were generated.

By first dichotomizing and then filtering the simulated datasets, Bashir retains a highly unusual subset of the 50,000 simulated datasets: only those datasets with unusually weak associations between average citizens' preferences and policy outcomes. Specifically, the estimated coefficient for average citizens in the 2,500 datasets Bashir retains has a mean of only about .11 (in contrast with the intended "true" coefficient of .41).

It is this low mean of the estimated coefficient for average citizens —not its variance—that accounts for the high proportion of simulated coefficients that are "near zero." Rerunning Bashir's simulation using the R script he posted online, but without dichotomizing Y and without filtering the simulated datasets, produces estimated coefficients for all the predictors that are virtually identical to the "true" chosen values used to seed the data. (This remains true whether one uses our observed value of .03 for the average citizens' coefficient or Bashir's alternative .41 value; see panels B and C in the table below). When Bashir's simulation is run without dichotomizing Y and without filtering the simulated datasets, the proportion of the estimated coefficients for average citizens that are "near zero" is reduced from over 20% to under 1% (panel B of the table). In other words, if we applied our statistical

---

[2] Specifically, the process of dichotomizing the outcome variable reduces the coefficients for all three predictors--from .41 to.30 (for average citizens), from .76 to.57 (for economic elites), and from .56 to.42 for (interest groups). These numbers were produced by rerunning Bashir's R script without his filter, thereby revealing the impact of dichotomizing the outcome variable.

[3] Specifically, the range Bashir uses is within .05 for the bivariate coefficients for average citizens and economic elites, and within .25 for interest groups.

procedure to Bashir's alternative world in which average citizens have substantial influence over policy (at .41 rather than .03) we would very rarely (< 1%) get results similar to those that we found with our actual data.

Based on his simulation, Bashir claims that "The statistical approach employed in the study's central test seems too unreliable to gauge how much influence median-income citizens enjoy…" (p.6). Although it is impossible to tell from the results Bashir reports in his article, there is in fact no evidence of unreliability in the simulation he conducts. His finding that a large portion of the estimated coefficients are "near zero" results not from high variance (i.e., unreliability) but from the unexpectedly low estimated coefficient (which is not reported in Bashir's article). As explained above, the low coefficient results not from any failure of our estimation procedure to reproducing the "true" coefficients, but from the way the datasets were constructed and filtered.

Bashir reports a variety of other results from this simulation, such as the difference in the coefficients for average citizens and economic elites. These results are similarly affected by the problems described above.

We hope that other scholars will expand and improve upon our work (the data used in Gilens and Page (2014) and related publications are available from Gilens' website). While every empirical study has its strengths and limitations—many of which are discussed in our various publications—we are confident that these specific objections raised by Omar Bashir are misplaced.

		"true" coefficients used to seed the simulated data	estimated coefficients from simulation results	percent of simulated datasets retained	percent of estimated coefficients for average citizens <=.05
(A) Bashir's original simulation	average citizens economic elites interest groups	.41 .76 .56	.11 .69 .44	5.3%	18.5% [a]
(B) Same as (A) but without dichotomizing the outcome variable or filtering the datasets	average citizens economic elites interest groups	.41 .76 .56	.41 .76 .56	100%	0.7%
(C) Same as (B) but with average citizens "true" coefficient set to .03 to match findings in Gilens & Page (2014)	average citizens economic elites interest groups	.03 .76 .56	.03 .76 .56	100%	57.6%

[a] Bashir reports that "over 20%" of the datasets from his simulation produced coefficients for average citizens of <= .05. I found 18.5% when I reran his R script, but every set of 2500 retained simulations will produce slightly different results given the randomized data generating process.

References

Achen, Christopher H., 1982. *Interpreting and Using Regression*, Sage Publications, Beverly Hills, Calif.

Gilens, Martin. 2012. *Affluence and Influence: Economic Inequality and Political Power in America*. Princeton NJ and New York, NY: Princeton University Press and the Russell Sage Foundation.

Gilens, Martin, and Benjamin I. Page. 2014. " Testing Theories of American Politics: Elites, Interest Groups, and Average Citizens." *Perspectives on Politics* 12 (3).