Original Articles

Sociological Methodology
1–33

DOI: 10.1177/0081175018777988
http://sm.sagepub.com

WE RAN 9 BILLION REGRESSIONS: ELIMINATING FALSE POSITIVES THROUGH COMPUTATIONAL MODEL ROBUSTNESS

John Muñoz*
Cristobal Young*

Abstract

False positive findings are a growing problem in many research literatures. We argue that excessive false positives often stem from model uncertainty. There are many plausible ways of specifying a regression model, but researchers typically report only a few preferred estimates. This raises the concern that such research reveals only a small fraction of the possible results and may easily lead to nonrobust, false positive conclusions. It is often unclear how much the results are driven by model specification and how much the results would change if a different plausible model were used. Computational model robustness analysis addresses this challenge by estimating all possible models from a theoretically informed model space. We use large-scale random noise simulations to show (1) the problem of excess false positive errors under model uncertainty and (2) that computational robustness analysis can identify and eliminate false positives caused by model uncertainty. We also draw on a series of empirical applications to further illustrate issues of model uncertainty and estimate instability. Computational robustness analysis offers a method for relaxing modeling assumptions and improving the transparency of applied research.

*Stanford University, Stanford, CA, USA

Corresponding Author:
John Muñoz, Sociology, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA.
Email: jdmunoz@stanford.edu

Keywords

model uncertainty, false positive, type I error, computational methods, robust, multimodel analysis

1. INTRODUCTION

One of the central challenges facing researchers today is the problem of model uncertainty. Think of an analysis where the goal is to estimate a true "treatment effect" of a key explanatory variable using a conditioning-on-observables approach. Researchers do not typically know what is the underlying "true model"[1] that generated their data and are never sure which exact model specification is best for an unbiased and efficient estimator of the treatment effect (Leamer 1983; Raftery 1995; Sala-i-Martin 1997; Western 1996; Winship and Western 2016; Young 2009). However, the number of plausible models researchers can and do run is often large and prone to becoming a "garden of forking paths" (Gelman and Loken 2014). Any particular model specification represents a complex bundle of modeling assumptions representing analytic choices among the set of appropriate controls, alternative variable definitions, standard error calculations, as well as possible estimation commands and functional form issues (Young and Holsteen 2017). The growth of computational power in recent times allows analysts to select a preferred model after testing the results of hundreds or thousands of model variants. This leaves researchers to explore a wide model space and select a preferred model among many plausible candidates. Arbitrary refinements to model specification can create false positive errors—parameter estimates that are statistically significant even when there is no real relationship in the data.

Some model specifications may yield significant estimates by leveraging idiosyncratic aspects of the data—capitalizing on chance associations rather than real effects. Researchers have incentives to find statistically significant effects (Brodeur et al. 2016; Gerber et al. 2010; Glaeser 2008), and through a process of motivated reasoning, they may be prone to see superior methodology in models that generate significant results (Epley and Gilovich 2016). This may explain why there appears to be an alarming overabundance of false positive, nonreplicable results in many social science literatures (Chabris et al. 2012; Open Science Collaboration 2015; Prinz, Schlange, and Asadullah 2011; Sala-i-Martin, Doppelhofer, and Miller 2004).

Empirical results are a joint product of both the data and the modeling assumptions. There is no assumption-free way of conducting empirical analysis (Heckman 2005). To draw conclusions based on a single empirical estimate is to tacitly assume that other plausible specifications either yield the same result or are incorrect and misleading. It is common to see footnotes in research articles mentioning alternative specifications that (inevitably) support the main conclusions. However, this footnote approach to model uncertainty offers weak and *ad hoc* evidence of the robustness of the analysis. Social science can and should have more rigorous standards for examining critical assumptions.

Multimodel analysis and computational robustness testing can help correct the problems of inflated significance and nonrobust results (Sala-i-Martin et al. 2004; Young 2009). The approach relies on computational power to systematically estimate an entire (theoretically informed) model space, defined by all possible combinations of specified model ingredients: possible control variables, alternative definitions, estimation commands, functional forms, and standard error calculations (Young and Hosteen 2017). The output of computational robustness analysis is a distribution of estimates showing how the parameter estimates could change if a different plausible model specification were used. An author's preferred estimate could then be interpreted in light of the larger modeling distribution of estimates.

We use large-scale simulations with random noise data to show how model uncertainty can generate inflated significance levels and false positive results. False positives obtain from seemingly innocuous model refinement strategies, such as dropping insignificant variables (Freedman 1983; O'Brien 2017; Raftery 1995). When there are no true relationships in the data but many variables to choose from, statistically significant results can be readily found.

However, when false positive results are subjected to computational robustness testing, many of them are easily rejected as nonrobust. False positives often hinge on a "knife edge" specification and lose their significance in almost any alternative plausible model. Model robustness analysis helps minimize false positive errors by showing the instability of the results across other plausible model specifications. Not all false positives are overturned, but after robustness testing, the false positive rate is generally at or below the 5 percent rate expected in conventional statistical tests. In our simulations, robustness analysis corrects the excess false positives associated with model uncertainty.

We also draw on empirical applications to illustrate the challenges of model uncertainty and robustness in practice. We examine a small-*N* exploratory study with high levels of model uncertainty (Jung et al. 2014a, which originally showed that "female" hurricanes are more deadly). We also examine a confirmatory study of how job training programs affect earnings using a large sample size and drawing on considerable prior research (Dehejia and Wahba 1999; LaLonde 1986). For the confirmatory study of job training programs, we contrast two separate data sets with identical covariates: (1) field experiment data that address problems such as selection bias through random assignment and (2) observational data, which depend on covariate adjustment to rule out spurious relationships and may well be more sensitive to model specification. In these cases, model robustness analysis sheds important light on the reliability of the results. These applications help show how estimate instability identifies likely false positives and establishes more empirically defensible conclusions.

In the course of this project—testing the model robustness of all false positives appearing in our simulations—we estimated over 9 billion regressions. That this is even possible serves as an important reminder: The current practice of reporting one or two preferred estimates in a research paper is out of touch with modern computing power. Social science needs better ways of reporting the multitude of regression parameter estimates that authors are able to calculate in the course of applied research.

1.1. *False Positives and Asymmetric Information*

Our starting point is a statistical analysis that is oriented toward understanding a treatment effect. For example, what is the effect of attending private school, rather than public school, on learning and educational attainment? Many applied research questions in sociology take this form, in which the goal is to understand how a key variable of interest (e.g., private school attendance) affects a specific outcome (learning). Often researchers adopt a conditioning-on-observables strategy to estimate a treatment effect: Regression adjustment is used to control for confounding influences (Heckman 2005; Morgan and Winship 2007). Control variables are introduced in the regression model with the goal of obtaining an unbiased and efficient estimator of the true treatment effect.

The simple, linear statistical model takes the following form: The outcome, y_i, is given by

$$y_i = \beta_1 x_{1i} + \{\beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki}\} + \varepsilon_i, \quad (1)$$

in which x_1 is the treatment variable (or variable of interest), and the treatment effect is given by β_1. The challenge of statistical modeling is to find a set of control variables such that the conditional mean of the error term is zero—namely, $E[\varepsilon_i | x_{ki} = 0]$.[2] Standard proofs that ordinary least squares (OLS) provides the best linear unbiased estimator of β_1 (the true treatment effect) invoke the assumption that the "true model" for equation (1) is known and applied. In practice, researchers do not know what the true model is, and they are limited to working with models that are simply plausible or preferred on some reasonable grounds. It is also typically the case that many different versions of equation (1) could be seen as reasonable, especially when we consider the views of critics and scholarly readers.

A false positive error occurs when researchers report a statistically significant treatment effect when no such effect actually exists (e.g., reporting a significant effect of private school attendance in a world where private school *per se* has no real effect on learning). Statistical conclusions come with a risk of being wrong. By using a *p* value of .05 for significance tests, researchers accept that they will falsely declare a null relationship to be significant 5 percent of the time purely by chance.[3] There has been a growing "crisis of science" in recent years of an *excess* number of false positives in the scientific literature (Ioannidis 2005; Open Science Collaboration 2015). Many of the findings in fields as broad as biomedicine, macroeconomics, and behavioral genetics appear riven with false positives when research is subject to careful replication testing (Begley and Ellis 2012; Chabris et al. 2012; Sala-i-Martin et al. 2004). For example, in biomedicine, private industry labs have reported that 65 percent to 89 percent of "landmark" publications are based on nonreplicable false positive findings (Begley and Ellis 2012; Prinz et al. 2011).

Something is driving a deep wedge between a statistical test that on paper has a 5 percent rate of false positives and bodies of published research with rates of false positives dramatically higher. Model uncertainty is central to the problem of excess false positives. If the "true model" were known, authors would have to run only one regression. In

practice, model uncertainty typically leads authors to run many models, in a process of repeated model refinement. Each tested (but unreported) model in the refinement process can increase the risk of stumbling on a false positive result.

The problem emerges most forcefully when only one or two preferred models are reported. What began as model uncertainty becomes a problem of asymmetric information between author and reader: Authors know much more about the sensitivity of the results than readers (Young 2009), and they may well have strong arguments supporting their preferred model. However, researchers must openly acknowledge the risk of confirmation bias and the pressure to publish significant findings (Reason 1995; see reviews in Franco, Malhotra, and Simonovits 2014; Nickerson 1998). In a process of "motivated reasoning," researchers may not be self-consciously aware of a bias in their thinking—they are actively reasoning their way to a compelling decision that others could accept (Epley and Gilovich 2016; Kunda 1990). However, with a different motivation—such as serving as critic rather than author—they could develop good reasons to favor very different model specifications.

One of the deepest challenges to scientific research is that statistical significance is partly under the control of the analyst. Readers do not know how much leeway analysts have in selecting different results to report. Preferred models are identified *after* knowing what estimates they produce. The reported results may be unrepresentative of other plausible models.

1.2. *Model Robustness*

To assess the stability of results, we draw on the computational model robustness framework as formulated (and implemented in Stata) by Young and Holsteen (2017). The model robustness approach can be used across many of the common estimation frameworks used in the social sciences, including OLS, maximum likelihood estimation, panel data estimators, and others.[4] When findings are generated through a process of successive model refinement—such as dropping nonsignificant variables—one way to check the credibility of the results is to look for estimate instability. As Raftery (1995:113) cautions, "the standard approach of selecting a single model and basing inference on it underestimates uncertainty . . . because it ignores uncertainty about model form" (see also Winship and Western 2016).

When significant results are discovered by dropping nonsignificant control variables, the results will tend to fall in and out of significance with trivial changes to model specification. In contrast, results that are due to the underlying relationships in a data set tend to be not very sensitive to arbitrary changes in the model. Testing multiple models is not problematic *per se* so long as there is transparency about how the estimates vary across those models.

Model robustness analysis tests the stability of an estimate across all unique combinations of plausible model ingredients within a theoretically informed model space. Researchers may distinguish between "necessary" model ingredients, such as certain variables that must be in the model, and possible/plausible model ingredients, such as controls that might or might not be used. Suppose we are interested in the effect of x on y (estimated by β_1). We are confident that x_2 must be included in the model as a necessary control but are open to questioning whether x_3 and x_4 belong in the true model (and credible arguments could be made either way). This uncertainty is represented in the following set of four possible models:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \tag{2}$$

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \tag{3}$$

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_4 x_{4i} + \varepsilon_i \tag{4}$$

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i. \tag{5}$$

These four equations represent different reasonable ways of specifying the model given the uncertainty, and they give four plausible estimates of β_1. As the number of plausible model ingredients increases, the model space increases exponentially: With J plausible control variables, there are 2^J unique combinations of those variables. With two cases of uncertainty (regarding two plausible controls) in the previous example, there are $2^2 = 4$ unique models. With 12 possible controls, there are $2^{12} = 4{,}096$ unique models, and with 20 possible controls, there are more than 1 million unique models. In each unique specification, the estimated coefficient β_1 may be subtly different depending on which controls are included or not, which is not transparent when reporting a few point estimates. To what degree do these different regression specifications impact our conclusion regarding the significance of the effect of x_1 on y?

Once all regressions within the model space are estimated, the result is a "modeling distribution" of estimated coefficients for the explanatory variable of interest. This modeling distribution is analogous to the conventional sampling distribution of classical statistics. The sampling distribution shows how an estimate varies in repeated sampling, addressing uncertainty about the sample of data (e.g., Efron and Tibshirani 1993). The modeling distribution shows how the estimate varies in repeated modeling, addressing uncertainty about the model. Taken together, these address the two fundamental sources of uncertainty: how an estimate might change if we took a new sample and how it might change if we used a different model.

A useful robustness statistic is the *total* standard error, which is a measure of both how much the estimate varies across the sampling distribution *and* the modeling distribution: the combined sampling and modeling standard errors. This is used to calculate the "robustness ratio": the preferred estimate of β_1 divided by the total standard error. As a simple rule of thumb, a variable is considered to have a robust statistical relationship with *y* if the robustness ratio is above a critical value of 2 (by analogy to the *t* statistic; see Raftery 1995).

1.3. *Contrast with Machine Learning*

While this robustness analysis relies on computational power to estimate many models, there are two important differences between robustness analysis and machine learning. First, machine learning is generally about model selection in high-dimensional space (where conventional approaches to variable selection are impractical—e.g., when there are thousands of potential predictor variables). The computational robustness approach, in contrast, does not use an algorithm to select the "best" model. The robustness framework expects authors to use theory, judgment, and previous research to develop their preferred model—as is standard practice in applied sociological research.[5] The contribution of robustness analysis is to reveal what alternative parameter estimates can be found in other plausible models. Do other models readily overturn a statistical conclusion, or do they offer repeated support? The ideal is to build a robustness framework that parallels what a task force of diverse social scientists would find when analyzing the same data with many different views of what is the better model specification (Young forthcoming). The fundamental goal is to show readers what is the range of

possible estimates and reduce the asymmetry of information between analyst and reader—to distinguish between "knife edge" estimates and results supported under diverse, credible model assumptions.

Second, machine learning is generally part of a predictive modeling framework that is not attempting to estimate a specific treatment effect (β_1). Rather, predictive modeling aims simply to predict the outcome (y), without interest in which variables do the predicting or whether the coefficients represent real treatment effects. Techniques such as the LASSO (least absolute shrinkage and selection operator) choose a parsimonious set of predictor variables that maximize model fit. The LASSO is roughly equivalent to selecting variables on the basis of a conservative t test and dropping variables that are not strongly correlated with y. This approach does not aim to select the variables that would be most important in a treatment effects framework (Belloni, Chernozhukov, and Hansen 2014). A key role of control variables for treatment effects analysis is to reduce omitted variable bias in the estimated effect (β_1). The LASSO does not accomplish this because it does not take into account a possible control variable's correlation with the treatment variable (x_1). As the well-known omitted variable bias (OVB) formula tells us, bias from an omitted variable (say, x_3) is driven by the correlation (x_1, x_3) just as much as it is by the correlation (y, x_3) (Greene 2012). Maximizing model fit is the wrong criteria for selecting control variables in a treatment effects analysis because it reflects only a control variable's correlation with y. Young and Holsteen (2017) use the computational robustness framework to show how much a control or other model ingredient empirically changes the estimated treatment effect (β_1). These influence scores would be better criteria for model selection than fit statistics such as the R^2 or the Bayesian Information Criterion (BIC) when the goal is to understand treatment effects. A LASSO estimator will generally drop a possible control x_3 when it has a low correlation with y but a high correlation with the treatment variable x_1. Yet such variables are often the most influential factors in applied analysis (Young and Holsteen 2017; see also O'Brien 2017; Winship and Western 2016).

2. TESTING COMPUTATIONAL MODEL ROBUSTNESS

Our first goal is to reduce the number of false positive errors. We build on a classic simulation study to better understand the problem of false

positives when analysts engage in model refinement (Freedman 1983). Using noise-on-noise regressions, we create stylized procedures to simulate researcher decisions. The first step is to create a set of variables and observations using a random number generator. The resulting data set contains no true causal associations.

In a first-stage regression, all explanatory variables are regressed on an outcome variable. Next, following Freedman (1983), the least significant variables ($p > .25$) are excluded, and a second-stage regression is estimated. This reflects uncertainty about which variables should be included in a model. We extend this simulation and use it to test the utility of model robustness analysis. We show how the rate of false positives varies systematically by the sample size of the data and the degree of model uncertainty (in this case, the number of candidate variables that could be included in the analysis). We then demonstrate how well model robustness analysis can overturn these false positives. A large portion of false positives are highly fragile in computational robustness testing.

Our final objective is to evaluate estimate (in)stability using model robustness analysis on real data. We use three different empirical data sets, including a small-*N* data set that was analyzed with vague theory and two large-*N* data sets used in a confirmatory study with strong prior theory—a field experiment and a cross-sectional analysis.

2.1. *Simulation Data and Methods*

Our basic simulation strategy is to construct a noise-on-noise regression analysis. We start by generating a random noise data set. We create 51 variables and 100 observations drawn from a standard normal distribution: a random-noise outcome, Y, and 50 random-noise explanatory variables $X_1, \ldots, X_{50}$. By construction, none of the X variables have any true relationship with y. Any statistically significant associations in the data are false positives, which we expect to see in 5 percent of our statistical tests due to our chosen significance level of .05.

In the first-stage regression analysis, all the $X_1, \ldots, X_{50}$ variables are included in an initial regression. From these results, we screen out the least significant variables and retain only those variables that were significant at the .25 level for the next stage. Results from one iteration help to clarify the procedure. We report the full results from a typical first-stage regression in Table A1 in the Appendix;[6] with 50 covariates,

Table 1. Regression Results from Second-stage Model, Predicting the Random Noise Outcome, *Y*

	Second-stage Regression	
Variable	Estimate	Standard Error
X_1	.46*	(.10)
X_2	.51*	(.10)
X_3	–.22*	(.09)
X_4	–.17	(.09)
X_5	.23*	(.10)
X_6	–.17	(.09)
X_7	–.13	(.10)
X_8	–.23*	(.09)
X_9	.24*	(.08)
X_{10}	.18*	(.09)
X_{11}	–.14	(.11)
X_{12}	–.18*	(.09)
X_{13}	–.13	(.10)
X_{14}	–.14	(.09)
X_{15}	–.15	(.09)
X_{16}	–.19*	(.09)
X_{17}	–.16	(.10)
X_{18}	–.16	(.10)
Constant	–.04	(.09)
Observations	100	
R^2	.419	

Note: Simulated data with no true associations. First-stage regression used the full 50 explanatory variables; the second stage (reported here) retained all explanatory variables significant at the .25 level.

*$p < .05$.

the table is unwieldly and uninteresting. This first-stage analysis looks, appropriately, like a noise-on-noise regression. There are 50 explanatory variables, and only 3 of those variables are statistically significant, meaning 6 percent of variables were significant in the first stage (roughly as expected using a 5 percent significance threshold).[7]

Along with these three significant variables, another 15 variables were significant at the "exploratory" .25 level. We retain these for the second-stage regression, while the other 32 "less significant" variables are screened out. This simulates a stylized scenario in which researchers drop variables that seem less relevant.

The second-stage refined model regression results shown in Table 1 look dramatically different from the first stage. There are now nine

variables that achieve statistical significance. This includes the initial three variables significant in the first-stage regression as well as six new variables, corresponding to an 18 percent (9/50) "false positive rate."[8] Based on *t* tests from the second-stage regression model, this false positive rate is alarmingly high.

In this simple example, model refinement achieved two critical and troubling results. First, it inflated the significance of many coefficients so that nine variables, rather than the initial three, were significant. Second, it dramatically reduced the *apparent* number of candidate variables that were ever considered in the analysis. As a result, the first pass looked like a noise-on-noise regression, but Table 1 looks like a serious analysis with a strong set of relevant variables. If we gave these variables substantive names (e.g., *income, religion*, or *political party affiliation*), the table would not look out of place in a major sociology journal, even though it is based on random noise.

How well can model robustness analysis reduce or eliminate the false positive errors? We test the model robustness[9] of all variables that are significant at the 5 percent level in the second-stage regression. In Table 1, 9 variables out of the 18 explanatory variables were significant. We treat each of these 9 variables as a variable of interest and subject each one of them to robustness testing. For each variable of interest, there are 17 other variables that serve as possible controls, yielding a model space with $2^{17} = 131{,}072$ unique regression models for each variable tested. In total, this involved running nearly 1.2 million regressions for this single iteration of our simulation test.[10]

To illustrate, Table 2 shows the output from running model robustness analysis on variable X_{10}. While X_{10} was significant in the second-stage regression, it is significant in only 5 percent of the thousands of small variations on that model. The robustness ratio is .89, well below the "rule of thumb" critical value of 2. Thus, accounting for how the estimate changes across models leads us to conclude X_{10} is not robust and that the statistically significant coefficient found in the second-stage regression model in Table 2 a likely false positive.

In Table 3, we report the robustness of all the variables from the second-stage ("final") regression (in Table 1). For the 18 explanatory variables retained in the analysis, Table 3 reports (1) the first-stage estimates, (2) the second-stage results, and (3) the robustness measures of significance rate and robustness ratio. We find that seven out of the nine variables significant in the second-stage regression were nonrobust

Table 2. Model Robustness of X_{10}

Linear Regression			
Variable of interest	X_{10}		
Outcome variable	y1	Number of observations	100
Possible control terms	17	Mean R^2	.16
Number of models	131,072	Multicollinearity	.24
Model Robustness Statistics		**Significance Testing**	**Percent**
Mean (b)	.0928	Sign stability	100
Sampling SE	.0926	Significance rate	5
Modeling SE	.0464		
Total SE	.1036	Positive	100
		Positive and significance	5
		Negative	0
Robustness ratio	.8966	Negative and significance	0

(including all six that were newly significant in the second-stage model but not the first regression). These variables had robustness ratios well below 2, and they were significant in less than 50 percent—and often less than 5 percent—of the model space. Both X_5 and X_{20} appear robust as they have robustness ratios greater than 2 and are significant in more than 90 percent of the models. For the data set as a whole, this gives a robustness rate of 4 percent (as 2 out of 50 variables were deemed robustly related to Y). Robustness analysis did not eliminate all false positives, but it did eliminate the false positive errors generated as a result of model refinement in the second-stage regression. This illustrates how robustness analysis can combat the problem of false positives.

2.2. *Full-scale Simulations*

We conducted 5,000 iterations of the aforementioned simulation for each of 17 unique conditions for data size (from 75 to 2,000 observations) and degree of model uncertainty (candidate variables from 20 to 90). Each condition involved testing roughly 550 million regressions. To expedite computation, we used parallel processing in which computation is distributed across thousands of processors. This allowed us to run more than 9 billion regressions in the space of several months rather than almost a decade on a single desktop computer.[11]

Table 3. Comparison of Significance and Robustness Statistics

Variable	First-stage Regression[a]	Second-stage Regression	Robustness	
			Significance Rate (%)	Robustness Ratio
X_1	.54* (.16)	.46* (.10)	93	2.36[R]
X_2	.51* (.14)	.51* (.10)	100	3.53[R]
X_3	–.30* (.13)	–.22* (.09)	2	–.64
X_4	–.21 (.13)	–.17 (.09)	<1	–.59
X_5	.23 (.14)	.23* (.10)	21	1.43
X_6	–.21 (.13)	–.17 (.09)	<1	–.82
X_7	–.25 (.15)	–.13 (.10)	2	–1.22
X_8	–.22 (.13)	–.23* (.09)	1	–1.16
X_9	.22 (.13)	.24* (.08)	28	1.70
X_{10}	.20 (.12)	.18* (.09)	5	.90
X_{11}	–.23 (.15)	–.14 (.11)	<1	–.74
X_{12}	–.19 (.13)	–.18* (.09)	43	–1.92
X_{13}	–.20 (.13)	–.13 (.10)	<1	–.37
X_{14}	–.19 (.13)	–.14 (.09)	<1	–.41
X_{15}	–.16 (.13)	–.15 (.09)	<1	–.55
X_{16}	–.17 (.14)	–.19* (.09)	<1	–1.09
X_{17}	–.17 (.15)	–.16 (.10)	<1	–.60
X_{18}	–.17 (.14)	–.16 (.10)	<1	–.56
Constant	–.05 (.13)	–.04 (.09)		
Observations	100	100		
R^2	.544	.419		

Note: Standard errors in parentheses. [R] = robust.

[a]Estimates from full regression with the full set of $X_1 - X_{50}$ variables. Only variables significant at .25 level are reported for comparison to the second-stage regression and robustness analysis.

*$p < .05$.

3. SIMULATION RESULTS

How consistently can model robustness analysis eliminate the false positives that arise after arbitrary model refinement? We first establish how many false positives systematically arise under such model refinements. We begin by simulating 5,000 iterations of model refinement on a random data set with 50 variables and 100 observations. In the first-stage regression—before refinement—the average percentage of significant variables is indeed 5 percent (see boldfaced row from Table 4). After dropping the least significant variables from this first-stage regression, however, we find that the false positive rate in the second-stage regression is 11.3 percent. This shows substantial significance inflation.

Table 4. False Positive and Model Robustness Rates by Sample Size

Number of Candidate Variables	Sample Size	Stage 1 Significance Rate (%)	Stage 2 Significance Rate (%)	Robustness Rate (%)	Total Number of Regressions Run
50	75	5.0	11.9	3.6	952,813,303
50	**100**	**5.0**	**11.3**	**4.4**	**850,106,422**
50	150	5.0	9.4	4.7	706,011,976
50	200	5.0	8.2	4.7	551,003,368
50	500	5.0	6.1	4.7	360,167,632
50	750	5.1	5.8	4.8	319,617,774
50	1,000	5.1	5.6	4.8	318,600,298
50	2,000	5.0	5.2	4.8	304,287,434

Note: Five thousand iterations. Bold row represents the starting point conditions of our simulations: 50 variables and 100 observations.

In model robustness analysis, however, many of these false positives appear nonrobust, meaning that their significance is highly dependent on the exact model specification from the second-stage regression. Compared with the 11.3 percent false positive variables in the second-stage regression, only 4.4 percent of the variables are found robust after accounting for how the estimate changes systematically across the model space. Thus, more than half of the false positive variables are not robust to the set of possible controls—a 6.9 percentage point elimination of false positives. Model robustness analysis effectively eliminated the excess false positive errors that arose due to (rather arbitrary) model refinement.

3.1. *False Positives and Model Robustness across Sample Size*

How do the false positive and robustness rates depend on sample size? Table 4 presents the simulation results where we start by generating data sets with 75 observations and increase up to a large data set with 2,000 observations while the number of candidate variables is held constant at 50. Figure 1 gives a visual presentation of the results. The dashed line represents the false positive rate using standard significance tests from the second-stage regression. The solid line plots the percentage of variables found to be robust to alternate model specifications. The shaded area between these two lines corresponds to the quantity of nonrobust false positive errors (false positives that were eliminated by model robustness analysis).

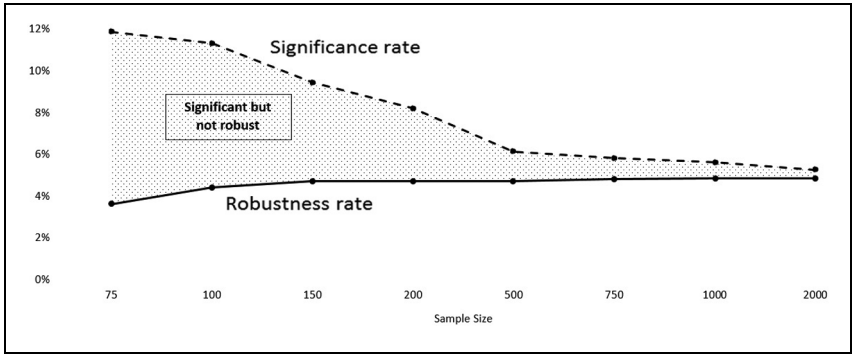

Figure 1. Significance and model robustness rates in Stage 2 across sample size. *Note:* Five thousand iterations per data point.

The rate of false positives for small sample data sets is nearly 12 percent, more than double our expectation of a 5 percent false positive rate by chance alone. Under our simulations, small data sets are especially prone to exhibit inflated significance after model refinement. In small-*N* data sets, the robustness rate hovers around 5 percent or below. With 75 observations, only about 3.6 percent of variables are robust. Compared with the 12 percent of variables that are false positives, this translates into an elimination of about 8.3 percentage points of false positive errors—a substantial reduction.

As sample size increases, we see the rate of false positives—the dashed line—steadily drops toward 5 percent. Once sample size reaches 500 or more observations, the rate of false positives ranges from 6.1 percent to a low of 5.2 percent with 2,000 observations. Larger data sets thus mostly avoid inflated significance errors associated with screening out nonsignificant variables for a second regression. In a sense, large data sets are less "twitchy" and less likely to generate false positives. In these large data sets, the robustness rate remains around 5 percent. Even though researchers with big data are at a lower risk of discovering false positives, using model robustness analysis likewise helps ensure a 5 percent error rate or below.

3.2. *False Positives and Model Robustness across Degree of Model Uncertainty*

Next, we examine the rates of false positives and model robustness across the degree of model uncertainty. The number of candidate

Table 5. False Positive and Model Robustness Rates by Degree of Model Uncertainty

Number of Candidate Variables	Sample Size	Stage 1 Significance Rate (%)	Stage 2 Significance Rate (%)	Robustness Rate (%)	Total Number of Regressions Run
20	100	5.0	7.6	6.4	960,807
30	100	5.0	8.9	5.3	18,693,932
40	100	5.1	10.2	4.7	272,695,876
50	**100**	**5.0**	**11.3**	**4.4**	**850,106,422**
60	100	5.0	11.8	3.8	1,192,472,153
70	100	5.1	11.8	3.1	1,177,587,028
75	100	5.1	11.7	2.8	965,930,118
80	100	5.0	10.9	2.4	752,395,175
90	100	5.2	9.4	1.6	305,500,163

Note: Five thousand iterations. Bold row represents the starting point conditions of our simulations: 50 variables and 100 observations.

variables is a measure of the extent of model uncertainty: More candidate variables mean there are a greater number of possible model specifications (i.e., variables that could be included as a control or not). For a fixed sample size of 100, we consider model uncertainty rising from 20 to 90 candidate variables.

The results of the simulations are reported in Table 5. Figure 2 graphically depicts the second-stage false positive rate (dashed line) and robustness rate (solid line) across the number of potential control variables. The area between the curves reflects the percentage of variables that are significant but not robust.

At low levels of model uncertainty, there is less severe significance inflation. With 20 possible controls, 7.6 percent of variables turn out to be significant (compared with the expected 5 percent). The number of significant variables peaks at roughly 12 percent when there are between 60 and 75 candidate variables. At very high numbers of candidate variables, the significance rate declines toward 9 percent. At this point, however, the number of variables approaches the number of observations.

The robustness rate, in contrast, is uniformly lower than the significance rate and declines steadily as the amount of model uncertainty increases. Model robustness analysis never overturns all false positives, but it is a clearly higher standard than statistical significance. At low levels of model uncertainty, many of the significant variables are found

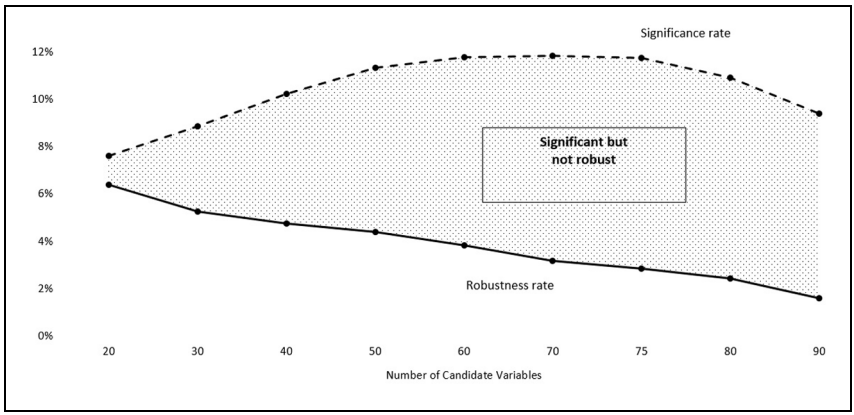

Figure 2. Significance and model robustness rates in Stage 2 across candidate variables.
Note: Five thousand iterations per data point.

to be robust as well—in part because there is little significance inflation occurring when model uncertainty is low. However, as uncertainty increases, model robustness analysis quickly becomes a fruitful strategy to eliminate false positives. At 40 candidate variables and above, more than half of false positives are eliminated in robustness analysis. Studies with greater degrees of uncertainty regarding model specification gain the most from model robustness analysis.

4. EMPIRICAL DATA AND METHODS

How does model robustness analysis perform with real-world data? Empirical analysis is an important addition because the simulations are based on the limiting (and conceptually powerful) case where all of the potential controls are random noise. In applied data, there is potential for bias of almost any magnitude or direction from model misspecification such as that which includes endogenous controls, using intermediate outcomes as controls, or other complex causal pathways (Clarke et al. 2018; Elwert and Winship 2014; Heckman and Navarro-Lozano 2004; Montgomery, Nyhan, and Torres 2016). Our simulations show that even in the absence of such data conditions, the choice of model specification can substantially affect the rate of false positives (i.e., even when model choice should not lead to specific bias). However, we

draw on two applications with observed data to elaborate on the simulation results and extend the findings in suggestive ways.

Unlike what happens in simulations, with applied empirical data, the "true effect" of a variable is not strictly known. Rather than directly quantifying false positives (which in empirical data is always subject to debate), we focus on estimate instability—the possibility of finding (very) different results using other reasonable model specifications.

4.1. *Are Female Hurricanes More Deadly? Testing Robustness in an Exploratory Analysis*

A widely reported study in the journal *Proceedings of the National Academy of Sciences* (*PNAS*) found that hurricanes with feminine-sounding names have higher death tolls than hurricanes with masculine names (Jung et al. 2014a). The authors argue that residents tend to dismiss the destructive potential of storms with feminine names and take fewer precautions against the danger than when storms have masculine names.

This study is exploratory. First, there is no prior empirical research on the topic, so the authors develop a model specification with no past experience as a guide. Second, there is no established theory connecting a storm's death toll with the gender of its name. This blank slate allows researchers to generate hypotheses after analyzing the data, potentially drawing a bull's-eye around the empirical arrow (also known as HARKing—hypothesizing after the results are known). Third, it is a small-*N* study, looking at hurricanes making landfall in the United States but not in any other country (because scaling up would require much greater resources). Finally, there is substantial model uncertainty: There are multiple ways of measuring the same conceptual factors, and there are many explanatory variables that could have been invoked (or not) to analyze the hurricanes and the places they affected. In essence, the number of plausible model specifications for this study *a priori* seems large—especially relative to the number of observations.

For our robustness analysis, we draw on a theoretically informed model space of alternative model ingredients. We incorporate all the published comments and rejoinders that emerged from the scholarly debate that followed the hurricane study and published in later issues of *PNAS* (Bakkensen and Larson 2014; Christensen and Christensen 2014; Jung et al. 2014b, 2014c; Maley 2014; Malter 2014). This includes

different functional forms, treatment of outliers, a range of possible controls, alternative standard error calculations, efforts to address endogeneity, and different estimation commands.[12] Taking all possible combinations of the model ingredients, there are 1,152 unique model specifications.

The first column in Table 6 presents model robustness results from the Jung et al. (2014a) data. Figure 3 presents a graphical view of the modeling distribution. The data are strongly concentrated around an estimate of zero effect of hurricane "femininity." However, the modeling distribution has a positive (rightward) skew, showing that there are some outlier estimates in which "female" hurricanes have higher death tolls. While 64 percent of the estimates have a positive sign (as in Jung et al. 2014a), less than 5 percent of models yield a statistically significant effect.

This analysis does not in itself mean that the hurricane study reported false positive results: That remains debatable (see Jung et al. 2014c). We do not recommend automatically rejecting results that do not meet the threshold for robust. Rather, we believe nonrobust results deserve careful substantive scrutiny to understand why the results depend critically on a particular model specification. In this case, the majority of plausible models shows no significant effect of a storm's "gender." It is not that the data support the author's conclusions—support is driven by choosing a very exact model specification.

4.2. *Empirical Application II: Job Training Programs and Earnings in Large-*N *Analysis*

What about confirmatory research using bigger data? Are the problems of estimate instability and false positives less worrisome when a researcher is armed with large data sets and guidance from previous research? To explore this, we draw on a classic data set on the labor market effects of job training programs (Dehejia and Wahba 1999; LaLonde 1986). The research question is as follows: Do unemployed workers benefit from participating in job training programs? These programs can help by providing temporary work experience and counseling. However, the programs may not help much if the problem is a lack of available jobs rather than the job-readiness of workers.

We use two data sets for testing the effect of program participation on reemployment and earnings.[13] The first is a field experiment with

Table 6. Empirical Robustness Testing

	Female Hurricane Names	Training Program: Field Experiment	Training Program: Cross-sectional Analysis
	Model Robustness Statistics		
Mean (*b*)	.101	1.692	–.815
Sampling *SE*	.176	.636	.598
Modeling *SE*	.180	.076	2.639
Total *SE*	.251	.641	2.705
Robustness Ratio	.400	2.640	–.301
	Significance Testing (%)		
Sign stability	64	100	63
Significance rate	5	100	41
Positive	64	100	63
Positive and significant	5	100	16
Negative	36	0	38
Negative and significant	0	0	25
	Robustness Testing Information		
Outcome Variable	Deaths	Salary	Salary
Number of observations	92	445	16,177
Mean R^2	.17	.04	.35
Number of models	1,152	256	256

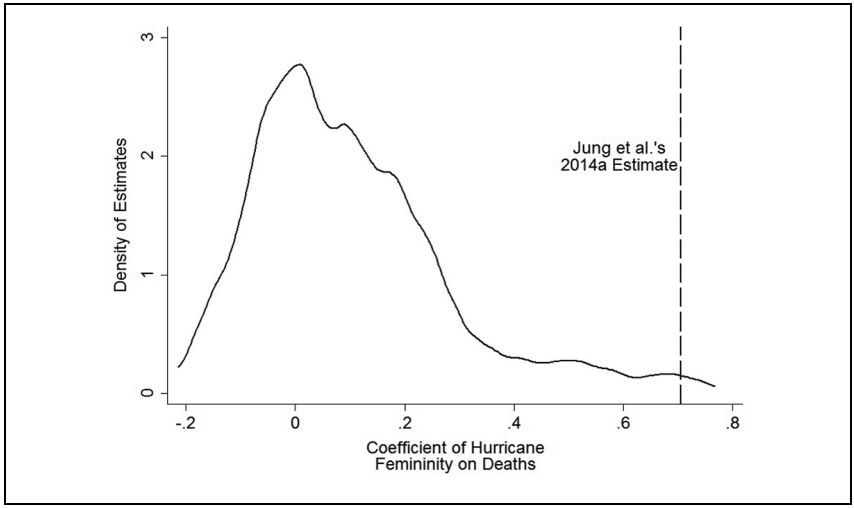

Figure 3. Model robustness results on Jung et al. (2014a) data.
Note: Kernel density graph of estimates from 1,152 models. See Table 6 for more information about the modeling distribution.

random assignment into job training (n = 445). The second draws on a cross-sectional sample of workers from the Current Population Survey (CPS; n = 16,177). Both data sets contained the same set of treatment and control variables, allowing different models to be identically specified in both data sets. The major distinction then is that the field experiment attempts to control for bias through randomization (although random assignment does not ensure the equivalence of treatment and control groups in any one study—only on average; Deaton and Cartwright forthcoming). The CPS cross-sectional analysis, in contrast, is a conditioning-on-observables strategy that depends more directly on model specification to address concerns about selection bias. We treat the following control variables as plausible model ingredients: past wages and unemployment status, age, race, marital status, and education. Taking all possible combinations of these variables yields 256 unique model specifications, each of which we apply identically to both the experimental and observational data sets.

The second column in Table 6 reports our robustness analysis on the experimental data. The mean estimate is 1.69 (meaning that job training increases wages), and there is remarkably little variation in the results

across model specifications. Both the sign stability and the significance rate are 100 percent. The modeling standard error is only a fraction (12 percent) of the sampling standard error. In other words, with random assignment to job training, model specification has essentially no influence on the results. The modeling distribution is essentially a single spike in estimates, with almost no meaningful variation. The conclusions are given by the data, not by the choice of statistical model.[14]

The final column in Table 6 reports the model robustness analysis on the larger but nonexperimental CPS data. The mean estimate is negative (–.82, meaning that job training reduces earnings). The sign stability is only 63 percent, and only 41 percent of models are statistically significant. Moreover, only 16 percent of the models support the general conclusion of the experimental analysis (a positive and significant effect), and more of the models arrive at the opposite conclusion (25 percent show a negative and significant effect). Overall, the modeling standard error (2.64) is more than four times as large as the sampling standard error (.60).

Figure 4 illustrates how stark the difference is between the two robustness analyses. The analysis using the field experiment data shows almost exactly the same results regardless of which controls are in the model. The cross-sectional analysis, in contrast, allows for a tremendous range of possible estimates, both positive and negative.[15] Across these models, a researcher could easily conclude that the training program significantly raises income, lowers income, or has no impact at all, depending on the preferred model specification.

Several basic conclusions emerge from these empirical analyses. First, if we regard the experimental results as the "true" effect of the job training program, then only a handful of possible models from the observational data report the correct results. It is perhaps encouraging that *any* model with observational data can replicate experimental results. However, it is important to note that the models that replicate experimental findings are not obvious: They include some controls (e.g., past earnings) but exclude others (e.g., age and marital status). It would be difficult *a priori* to know which models would give the correct results. Without benefit of the experimental evidence, it would be hard to draw a robust conclusion from the observational data. This may help explain why the published literature can offer different answers to the same question.

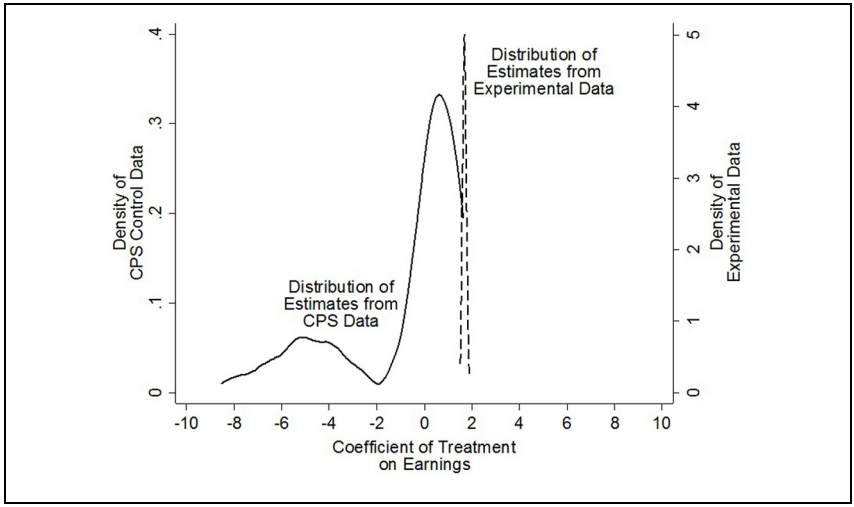

Figure 4. Model robustness results on training program data.
Note: Kernel density graph of estimates from 256 models for each distribution. See Table 6 for more information about the modeling distributions.

Second, the results suggest that some types of analysis allow much wider variation in possible results than others. This is a point that merits further research to understand how consistently different types of analysis show greater or lesser estimate instability (e.g., see Ho et al. 2007).

5. DISCUSSION AND CONCLUSION

Model uncertainty is one of the central challenges for social science researchers in the twenty-first century. Ambiguity and disagreement about how to best specify statistical models are increasingly a main source of tension in empirical research. In classical statistical theory, a single "true model" is assumed to be known prior to seeing the data. In practice, researchers have only a broad intuition about what is the correct model, and there are many plausible ways of testing a hypothesis. The process of formulating a statistical model has been aptly called a "garden of forking paths" (Gelman and Loken 2014). With modern computational power, a tremendous number of reasonable models can be readily tested, but usually only a few are reported (Sala-i-Martin et al. 2004). Indeed, in the current simulation study, we estimated more than 9 billion regressions. The mere fact that this is possible highlights

the reality of computational power today—which is not adequately reflected in the conventional tables of journal articles. The problem is one of asymmetric information: Analysts know (or *can know*) much more about the sensitivity and stability of the results than readers. The problem of model uncertainty is that arbitrary modeling choices can determine the results of empirical analysis. One consequence of this has been growing skepticism and cynicism about published research—a concern that many published papers reveal only a fraction of the possible results and often contain nonrobust, false positive findings (Ioannidis 2005; Leamer 1983). The challenge calls for greater transparency about how modeling choices influence reported findings.

We make several key contributions to the understanding of false positives and unstable estimates in a world of model uncertainty. We demonstrate that uncertainty about the "true model" can easily lead to a high number of false positives. We used simulated random noise data sets with no true statistical relationships so that any significant effects are by definition a false positive (a failure to reject the null of no effect). Under model uncertainty, simple strategies of model refinement (dropping insignificant variables) generate a higher rate of false positives than classical theory assumes. Tests that should have a 5 percent rate of error often have twice as many false positives. However, when false positive findings are generated through model refinement, the estimates tend to be unstable and highly sensitive to arbitrary changes in model specification. Model robustness analysis, by accounting for variation in the estimate across model specifications and rejecting highly unstable estimates, pushes the false positive errors back down to roughly the expected 5 percent rate.

Our simulations also find that false positives are most inflated when (1) sample size is low and (2) model uncertainty is high. In the course of model refinement, small data sets have greater risk of leveraging idiosyncratic data points to produce false positives. Hence, we describe small data sets as "twitchy" and prone to attaching significance to variables in seemingly temperamental ways. Likewise, when model uncertainty is higher—for example, when there are many possible controls to choose from—the risk of a false positive is substantially higher. Model uncertainty gives more "researcher degrees of freedom" to discover a nonrobust significant effect. In such conditions with low sample size and high model uncertainty, model robustness analysis performs the

best, identifying highly unstable estimates and overturning the greatest number of false positives in simulations.

Researchers should be particularly cautious about interpreting the results from exploratory research. By definition, exploratory research has less guiding theory to specify the model and usually does not have the resources to collect large data sets, making such studies doubly prone to false positive results. This is exemplified in the hurricane "gender" study: The estimate is highly unstable across plausible model specifications and rarely achieves significance. The substantive conclusions are derived from a highly specific model selection.

Our simulations with model uncertainty indicate that larger data sets have a lower risk of false positives. This is an important result that agrees with findings from meta-analysis (Doucouliagos and Stanley 2009) and supports the routine use of larger scale data in the social sciences. However, as we show, serious problems of model uncertainty, estimate instability, and nonrobust results can nevertheless remain in large data sets. In our empirical case, model robustness analysis sharply distinguished between observational and experimental methods from their degree of estimate instability. Subjecting empirical data to model robustness analysis can help distinguish between credibly true relationships and suspect findings. Careful attention to model robustness and estimate instability is just as important as statistical significance for identifying important effects. Ultimately, this calls for a normative change in how researchers and readers evaluate statistical models.[16]

We hope that the "pressure to publish" in academia remains tightly coupled with the goal of producing informative, reliable research. But with this pressure to publish significant results and an inherent risk of motivated reasoning, authors can often convince themselves that the most compelling model specifications are the ones that achieve statistical significance. In similar form, critics can often convince themselves that existing research is profoundly flawed and that if the data were in their hands, completely opposite findings can be easily found. The gap between these views is due to model uncertainty—the fact that no one knows the true model—and a lack of transparency about what other reasonable models show. Computational robustness analysis provides a rigorous and transparent method to address the problems of inherent model uncertainty, asymmetric information between analyst and reader, and the overabundance of false positive research findings.

APPENDIX

Table A1. Full Regression Results from First Stage

First-stage Regression	
X_1	.54* (.16)
X_2	.51* (.14)
X_3	–.30* (.13)
X_4	–.21 (.13)
X_5	.23 (.14)
X_6	–.21 (.13)
X_7	–.25 (.15)
X_8	–.22 (.13)
X_9	.22 (.13)
X_{10}	.20 (.12)
X_{11}	–.23 (.15)
X_{12}	–.19 (.13)
X_{13}	–.20 (.13)
X_{14}	–.19 (.13)
X_{15}	–.16 (.13)
X_{16}	–.17 (.14)
X_{17}	–.17 (.15)
X_{18}	–.17 (.14)
X_{19}	–.14 (.13)
X_{20}	–.15 (.15)
X_{21}	–.13 (.13)
X_{22}	.11 (.12)
X_{23}	.09 (.11)
X_{24}	–.10 (.13)
X_{25}	.10 (.12)
X_{26}	.10 (.13)
X_{27}	.10 (.14)
X_{28}	.08 (.13)
X_{29}	.07 (.13)
X_{30}	.07 (.12)
X_{31}	.05 (.12)
X_{32}	.05 (.12)
X_{33}	–.05 (.13)
X_{34}	.05 (.13)
X_{35}	–.05 (.15)
X_{36}	–.05 (.13)
X_{37}	.05 (.15)
X_{38}	–.05 (.14)
X_{39}	.04 (.14)
X_{40}	.03 (.12)
X_{41}	.03 (.12)
X_{42}	.04 (.14)

(continued)

Table A1. (continued)

First-stage Regression	
X_{43}	.03 (.15)
X_{44}	.02 (.14)
X_{45}	.01 (.14)
X_{46}	–.01 (.13)
X_{47}	.01 (.15)
X_{48}	.00 (.12)
X_{49}	.00 (.15)
X_{50}	.00 (.13)
Constant	–.05 (.13)
Observations	100
R^2	.544
df	50

Note: Standard errors in parentheses.
*$p < .05$.

Authors' Note

A replication package, including data and analysis files for the applied section and sample code for our simulations, is available as an online supplement.

Acknowledgments

The authors thank James Chu, Michelle Jackson, and the participants of the Analytic Sociology Workshop and the Inequality Workshop at Stanford University for helpful feedback and suggestions. The authors also thank four cohorts of graduate students in Sociology 382 and students in Sociology 124D for their questions, which helped us distill and clarify this research.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding for this project was provided by the Institute for Research in the Social Sciences (IRiSS) at Stanford University.

Notes

1. Unlike true score theory of measurement that an observed score is the result of a true score plus random error, the true model refers to the true parameterization and specification of the complex relationships of a given phenomenon and is never known or perfectly specified by researchers. The uncertainty regarding how best to approximate this model is what we refer to as *model uncertainty*.
2. There are many ways to complicate this framework with more nuanced model specifications. However, this formulation generalizes to other aspects of model

specification as any specification error can be written mathematically as an omitted variable (Heckman 1979).

3. In contrast, false negative errors occur when variables have a real effect on an outcome but are not found to be significant. While false negatives are conceptually important (Esarey and Danneman 2015), there is little appetite in the academic literature for null findings, and null findings are often difficult to publish (Gerber et al. 2010; Gerber and Malhotra 2008). Researchers, therefore, already have strong incentives to dispel false negative errors.
4. The Stata *mrobust* command is currently compatible with the following estimation commands: *regress, logit, logistic, probit, poisson, nbreg, areg, rreg, xtreg*.
5. We examined every article published in the *American Journal of Sociology* and *American Sociological Review* in 2016 to see whether the article made use of a formal model selection algorithm. Only 4.2 percent of quantitative articles invoked a model selection algorithm. (Additional details are available on request from the authors.) The predominant framework for developing a model specification in contemporary sociology is to draw on theory and existing research.
6. For all tables of the example simulation iteration, variables $X_1, \ldots, X_{50}$ have been renamed for ease of interpretation so that X_1 represents the most significant variable in the first-stage regression, X_{23} represents the twenty-third most significant variable, and so on.
7. On average, there should be 2.5 out of 50 variables, or 5 percent significant at the .05 p level.
8. We track the number of variables significant at that 5 percent level in this second-stage regression and take this value and divide by the total number of candidate variables to calculate a "false positive rate." The false positive indicates what percentage of candidate variables could be found statistically significant even in a noise-on-noise regression. If model refinement does not bias our results, the false positive rate should be equal to our chosen significance rate of a standard significance test (5 percent). A false positive rate above that 5 percent threshold indicates we have bias inflating our significance rates.
9. We tested the robustness using the command *mrobust* in Stata 14.2.
10. Including the first two regressions reported in Table A1 and Table 1, some 1,179,650 total regressions were run (= 131,072 × 9 + 2) in this example of a single iteration.
11. Summing total regressions for each data set structure in "Total Number of Regressions Run" column from Tables 4 and 5 (but counting only the 50 variable, 100 observation data set once) leads to 9.34 billion total regressions.
12. We incorporate the following alternative model ingredients: a different functional form of femininity (main effect rather than interaction with hurricane damages), handling of outliers (excluding hurricanes with more than 100 deaths), addressing potential endogeneity (removing damages as a covariate, adding population in the year of the storm, or adjusting death count for current population), dealing with the change in gender naming of storms (dropping data before 1979), different possible choices of controls (interaction between damages and minimum pressure, U.S. population during the year of the storm, elapsed years), alternative ways of estimating the variance-covariance matrix (standard or robust), and an alternative

type of regression model (negative binomial, or ordinary least squares log-linear regression).

13. Originally analyzed by LaLonde (1986). We use the data made available in Dehejia and Wahba (1999), and we follow Athey and Imbens (2015) for the baseline model specification, which includes prior earnings as covariates.
14. Of course, this does not rule out the possibility that other methods not currently considered might affect the conclusions in a future, more expansive robustness analysis.
15. Only a handful of models with observational data approximated the effect in experimental data, all of which included controls for two prior years of earnings/unemployment status, education, and race but *excluded* age and marital status. It is possible that the variables age and marital status are correlated with other omitted variables so that including them leverages greater overall bias than excluding them (Clarke et al. 2018). For further discussion, see Elwert and Winship (2014) on the problem of conditioning on a collider variable.
16. We thank an anonymous reviewer for making this point.

References

Athey, Susan, and Guido Imbens. 2015. "A Measure of Robustness to Misspecification." *American Economic Review* 105(5):476–80.

Bakkensen, Laura, and William Larson. 2014. "Population Matters When Modeling Hurricane Fatalities." *PNAS* 111(50):E5331–32.

Begley, Glenn, and Lee M. Ellis. 2012. "Drug Development: Raise Standards for Preclinical Cancer Research." *Nature* 483:531–33.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High Dimensional Controls." *Review of Economic Studies* 81:608–50.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8(1): 1–32.

Chabris, Christopher, Benjamin Hebert, Daniel Benjamin, Jonathan Beauchamp, David Cesarini, Matthijs van der Loos, Magnus Johannesson, Patrik Magnusson, Paul Lichtenstein, Craig Atwood, Jeremy Freese, Taissa Hauser, Robert M. Hauser, Nicholas Christakis, and David Laibson. 2012. "Most Genetic Associations with General Intelligence Are Probably False Positives." *Psychological Science* 23: 1314–23.

Christensen, Bjorn, and Soren Christensen. 2014. "Are Female Hurricanes Really Deadlier Than Male Hurricanes?" *PNAS* 111(34):E3497–98.

Clarke, Kevin, Brenton Kenkel, and Miguel Rueda. 2018. "Omitted Variables, Countervailing Effects, and the Possibility of Overadjustment." *Political Science Research and Methods* 6(2):343–54.

Deaton, Angus, and Nancy Cartwright. Forthcoming. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science and Medicine*.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94(448):1053–62.

Doucouliagos, Hristos, and T. D. Stanley, 2009. "Publication Selection Bias in Minimum-wage Research? A Meta-regression Analysis." *British Journal of Industrial Relations* 47(2):406–28.

Efron, Bradley, and Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Elwert, Felix, and Christopher Winship. 2014. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *Annual Review of Sociology* 40: 31–53.

Epley, Nicholas, and Thomas Gilovich. 2016. "The Mechanics of Motivated Reasoning." *Journal of Economic Perspectives* 30(3):133–40.

Esarey, Justin, and Nathan Danneman. 2015. "A Quantitative Method for Substantive Robustness Assessment." *Political Science Research and Methods* 3(1):95–111.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345(6203):1502–05.

Freedman, David A. 1983. "A Note on Screening Regression Equations." *The American Statistician* 37(2):152–55.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science Data-dependent Analysis—A 'Garden of Forking Paths'—Explains Why Many Statistically Significant Comparisons Don't Hold Up." *American Scientist* 102(6): 460.

Gerber, Alan, and Neil Malhotra. 2008. "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods and Research* 37:3–30.

Gerber, Alan, Neil Malhotra, Conor M. Dowling, and David Doherty. 2010. "Publication Bias in Two Political Behavior Literatures." *American Politics Research* 38(4):591–613.

Glaeser, Edward. 2008. "Researcher Incentives and Empirical Methods." Pp. 300–19 in *Foundations of Positive and Normative Economics: A Handbook*, edited by A. Caplin and A. Schotter. Oxford, UK: Oxford University Press.

Greene, William. 2012. *Econometric Analysis*. 7th ed. Upper Saddle River, NJ: Prentice Hall.

Heckman, James. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1):153–61.

Heckman, James. 2005. "The Scientific Model of Causality." Pp. 1–97 in *Sociological Methodology*. Vol. 35, edited by R. M. Stolzenberg. Boston, MA: Blackwell Publishers.

Heckman, James, and Salvador Navarro-Lozano. 2004. "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models." *Review of Economics and Statistics* 86(1):30–57.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15:199–236.

Ioannidis, John. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2:696–701.

Jung, Kiju, Sharon Shavitta, Madhu Viswanathana, and Joseph M. Hilbed. 2014a. "Female Hurricanes Are Deadlier Than Male Hurricanes." *PNAS* 111(24):8782–87.

Jung, Kiju, Sharon Shavitta, Madhu Viswanathana, and Joseph M. Hilbed. 2014b. "Reply to Christensen and Christensen and to Malter: Pitfalls of Erroneous Analyses of Hurricane Names." *PNAS* 111(34):E3499–500.

Jung, Kiju, Sharon Shavitta, Madhu Viswanathana, and Joseph M. Hilbed. 2014c. "Reply to Maley: Yes, Appropriate Modeling of Hurricane Fatality Counts Confirms Female Hurricanes Are Deadlier." *PNAS* 111(37):E3835.

Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108(3):480.

LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76(4):604–20.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73(1):31–43.

Maley, Steve. 2014. "Statistics Show No Evidence of Gender Bias in the Public's Hurricane Preparedness." *PNAS* 111(37):E3834.

Malter, Daniel. 2014. "Female Hurricanes Are Not Deadlier Than Male Hurricanes." *PNAS* 111(34):E3496.

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2016. "How Conditioning on Post-treatment Variables Can Ruin Your Experiment and What to Do about It." Presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.

Morgan, Stephen, and Christopher Winship. 2007. *Counterfactuals and Causal Analysis: Methods and Principles for Social Research*. Cambridge, UK: Cambridge University Press.

Nickerson, Raymond S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2(2):175.

O'Brien, Robert M. 2017. "Dropping Highly Collinear Variables from a Model: Why It Typically Is Not a Good Idea." *Social Science Quarterly* 98(1):360–75.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251):aac4716.

Prinz, Florian, Thomas Schlange, and Khusru Asadullah. 2011. "Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?" *Nature Reviews Drug Discovery* 10:712.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." Pp. 111–64 in *Sociological Methodology*. Vol. 25, edited by P. V. Marsden. Oxford, UK: Blackwell Publishing.

Reason, James. 1995. "Understanding Adverse Events: Human Factors." *Quality in Health Care* 4(2):80–89.

Sala-i-Martin, Xavier. 1997. "I Just Ran Two Million Regressions." *American Economic Review* 87(2):178–83.

Sala-i-Martin, Xavier, Gernot Doppelhofer, and Ronald Miller. 2004. "Determinants of Long-term Growth: A Bayesian Averaging of Classical Estimates Approach." *American Economic Review* 94:813–35.

Western, Bruce. 1996. "Vague Theory and Model Uncertainty in Macrosociology." Pp. 165–92 in *Sociological Methodology*. Vol. 26, edited by A. E. Raftery. Washington, DC: American Sociological Association.

Winship, Christopher, and Bruce Western. 2016. "Multicollinearity and Model Misspecification." *Sociological Science* 3:627–49.

Young, Cristobal. 2009. "Model Uncertainty in Sociological Research: An Application to Religion and Economic Growth." *American Sociological Review* 74(3):380–97.

Young, Cristobal. Forthcoming. "Model Uncertainty and the Crisis in Science." *Socius*.

Young, Cristobal, and Katherine Holsteen. 2017. "Model Uncertainty and Robustness: A Computational Framework for Multi-model Analysis." *Sociological Methods and Research* 46(1):3–40.

Author Biographies

John Muñoz is a doctoral candidate in the Department of Sociology at Stanford University. He previously completed his bachelor's degree in sociology and political science from University of North Carolina at Chapel Hill. His research aims to advance causal inference methods as well as social movement outcomes and participation. His work has previously been published in *Mobilization* and *Sociological Forum*.

Cristobal Young is an associate professor in the Department of Sociology at Cornell University. He studies socioeconomic dynamics that shape the effects of public policies, especially efforts to reduce inequality. He is passionate about evidence-based public policy, and about making the evidence easy to digest. He has previously published on model uncertainty in the *American Sociological Review* and *Sociological Methods and Research*. His first book, *The Myth of Millionaire Tax Flight,* was published in 2017.