Regular article

Sociological Methods & Research
2017, Vol. 46(1) 3-40

DOI: 10.1177/0049124115610347
journals.sagepub.com/home/smr

Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis

Cristobal Young[1] and Katherine Holsteen[2]

Abstract

Model uncertainty is pervasive in social science. A key question is how robust empirical results are to sensible changes in model specification. We present a new approach and applied statistical software for computational multimodel analysis. Our approach proceeds in two steps: First, we estimate the modeling distribution of estimates across all combinations of possible controls as well as specified functional form issues, variable definitions, standard error calculations, and estimation commands. This allows analysts to present their core, preferred estimate in the context of a distribution of plausible estimates. Second, we develop a model influence analysis showing how each model ingredient affects the coefficient of interest. This shows which model assumptions, if any, are critical to obtaining an empirical result. We demonstrate the architecture and interpretation of multimodel analysis using data on the union wage premium, gender dynamics in mortgage lending, and tax flight migration among U.S. states. These illustrate how initial results can be strongly robust to alternative model specifications or remarkably dependent on a knife-edge specification.

[1] Department of Sociology, Stanford University, Stanford, CA, USA
[2] Department of Epidemiology and Clinical Research, Stanford University, Stanford, CA, USA

Corresponding Author:
Cristobal Young, Stanford University, Stanford, CA 94305, USA.
Email: cristobal.young@stanford.edu

Keywords

model uncertainty, model dependence, computational methods, robust, multimodel analysis

Introduction

Model uncertainty is pervasive and inherent in social science. Social theory provides empirically testable ideas but by its nature does not give concrete direction on how the testing should be done (Leamer 1983; Raftery 1995; Western 1996; Young 2009). Indeed, social theory rarely says which control variables should be in the model, how to operationally define the variables, what the functional form should be, or how to specify the standard errors. When the "true" model is unknown, it is hard to say which imperfect approximation is best. As a result, theory can be tested in many different ways and modest differences in methods may have large influence on the results.

Empirical findings are a joint product of both the data and the model (Heckman 2005). Data do not speak for itself, because different methods and models applied to the same set of data often allow different conclusions. Choosing which model to report in a paper is "difficult, fraught with ethical and methodological dilemmas, and not covered in any serious way in classical statistical texts" (Ho et al. 2007:232). A growing challenge in social science is evaluating and demonstrating model robustness: the sensitivity of empirical results to credible changes in model specification (Durlauf, Fu, and Navarro 2012; Glaeser 2008; Young 2009).

We advance a framework for model robustness that can demonstrate robustness across sets of possible controls, variable definitions, standard errors, and functional forms. We estimate all possible combinations of specified model ingredients, report key statistics on the modeling distribution of estimates, and identify the model details that are empirically most influential. We emphasize the natural parallel between uncertainty about the *data* and uncertainty about the *model*. The usual standard errors and confidence intervals reflect uncertainty about the data indicating how much an estimate changes in repeated sampling. Our computational robustness strategy addresses uncertainty about the model—how much an estimate changes in *repeated modeling*.

Our framework builds on existing foundations of model uncertainty and model averaging (Leamer 1983, 2008; Raftery 1995; Sala-i-Martin 1997; Sala-i-Martin, Doppelhofer, and Miller 2004; Western 1996).[1] In contrast to model averaging, however, we allow analysts to retain focus on a core

preferred estimate, while also displaying for readers the distribution of estimates from many other plausible models. Moreover, we present a "model influence" analysis that shows how each element of model specification affects the reported results. This allows authors to clarify and demonstrate which modeling assumptions are essential to their empirical findings and which are not (Durlauf et al. 2012; Kane et al. 2013).

Do the results depend on minor and idiosyncratic aspects of model specification? Is there critical dependence on "convenient modeling assumptions that few would be willing to defend" (King and Zeng 2006:131)? When critically evaluating a research paper, scholars often look *outside* the reported model, thinking of new control variables that might moderate or overturn the results. It is equally important, however, to probe *inside* the model, to unpack the model ingredients, and see which elements are critical to obtaining the current results. This is the role of our model influence analysis.

It has been noted that "the diffusion of technological change in statistics is closely tied to its embodiment in statistical software" (Koenker and Hallock 2001:153). To this end, we introduce a new Stata module that implements our approach and can be flexibly used by other researchers. We illustrate the approach using three applied examples that demonstrate varying degrees of model robustness, drawing on data on the union wage premium, gender dynamics in mortgage lending, and the effect of income taxes on cross-border migration. These illustrate how initial results can be strongly robust to alternative model specifications or remarkably dependent on a knife-edge specification.

Point Estimates as Model Assumption Sets

Empirical results are driven by both the data and the model, but statisticians generally fail to acknowledge the role of model assumptions in their estimates. Consider a researcher with encyclopedic knowledge of statistical techniques and a rich set of empirical observations. In classical statistics, the true causal model is assumed to be known and only one model is ever applied to a sample of data. However, in common practice, the true model is not known and there are many possible variants on one's core analytic strategy. Edward Leamer describes some of the dimensions of model uncertainty:

> Sometimes I take the error terms to be correlated, sometimes uncorrelated; . . . sometimes I include observations from the decade of the fifties, sometimes I exclude them; sometimes the equation is linear and sometimes nonlinear; sometimes I control for variable *z*, sometimes I don't. (1983:37-38)

The potential modeling space is a broad horizon. Statistical analysts select options from a large menu of modeling assumptions making choices about the "best" functional form, set of control variables, operational definitions, and standard error calculations. These are necessary choices: Point estimates cannot be calculated until these modeling decisions are made. Indeed, calculating a point estimate often requires suppressing tangible uncertainty about the model and neglecting many plausible alternative specifications. In this sense, a point estimate represents a package of model assumptions and frequently captures just "one ad-hoc route through the thicket of possible models" (Leamer 1985:308). When just one estimate is reported, these assumptions are effectively elevated to "dogmatic priors" that the data *must* be analyzed *only* with the exactly specified model (Leamer 2008:4). Multimodel analysis is a way of relaxing these assumptions.

Model uncertainty easily leads to a problem of asymmetric information between analysts and readers (Young 2009). In the process of applied research, authors typically run many plausible models but in publication usually report only a small set of curated model specifications. Analysts, therefore, know much more about the sensitivity of their results than do their readers. In this context of asymmetric information, it is hard for readers to know if the reported results are powerfully robust to model specification or are simply an "existence proof" that significant results can be found somewhere in the model space (Ho et al. 2007:233).

There are two conditions under which a single point estimate is sufficient to represent the full distribution of estimates (Young 2009). First, if the true model is known, then all other models are inaccurate and misleading, and should not be reported. This is an untestable assumption that few analysts would assert. Second, if all other relevant models yield the same estimate, then these alternative specifications are redundant to report. This is an empirical question and can be tested by relaxing model assumptions and estimating alternative specifications.

Our perspective is that point estimates imply a set of testable model assumptions with a null hypothesis that other plausible models yield similar estimates. In this sense, there are two separate nulls for a point estimate. First is the classical significance test: Is the estimate different from zero? Second is the robustness test: Is the estimate different from the results of other plausible models?

How broad such a robustness analysis will be is a matter of choice. Narrow robustness reports just a handful of alternative specifications, while wide robustness concedes uncertainty about many details of the model. In field areas where there are high levels of agreement on appropriate methods and

Table 1. Robustness Footnotes in Top Sociology Journals, 2010.

	Total articles	Quantitative articles	Articles with 1+ robustness footnote	Percentage of Articles	Average robustness footnotes per article
Am Soc Review	39	32	26	81	3.0
Am Journal of Soc	35	28	25	89	3.5
Total	74	60	51	85	3.2

Source: Authors' review and coding of all articles published by these journals in 2010. The full data set listing the articles and our coding of them is available on request.

measurement, robustness testing need not be very broad. In areas where there is less certainty about methods, but also high expectations of transparency, robustness analysis should aspire to be as broad as possible.

Model Uncertainty and Multimodel Analysis in Current Practice

Today, there is tacit, widespread acknowledgement of model uncertainty. We often see footnotes about additional, unreported models that are said to support the main findings—an informal and ad hoc approach to multimodel inference. To see how ubiquitous the practice is in sociological research, we tallied the average number of footnotes referring to additional, unreported results in recent editions of two major sociology journals: the *American Journal of Sociology* and *American Sociological Review*. Of the 60 quantitative articles published in 2010, the vast majority—85 percent—contained at least one footnote referencing an unreported analysis purporting to confirm the robustness of the main results (see Table 1). The average paper contained 3.2 robustness footnotes. The text of these notes is fairly standard: "we ran additional models *X*, *Y*, and *Z*, and the results were the same/substantially similar/support our conclusions." Not one of the 164 footnotes we reviewed failed to support the main results. At least in footnotes, authors do not disclose models that qualify, weaken, or contradict their main findings.

Robustness footnotes represent a kind of working compromise between disciplinary demands for robust evidence on one hand (i.e., the tacit acknowledgement of model uncertainty) and the constraints of journal space on the other. In the end, however, this approach to multimodel inference is haphazard and idiosyncratic with limited transparency. These checks offer

reassurance but remain ad hoc and leave open the question of how much effort or critical reflection went into finding the full range of credible estimates. Moreover, they signal little about which model assumptions lend stronger or weaker support for a conclusion.

The uniformly reassuring tone of robustness footnotes stands in contrast to results from replication, repeated study, and meta-analysis. In areas of intensive research, where there are multiple studies on the same question, the estimates across studies tend to vary greatly, and by much more than their standard errors would suggest. In meta-analysis, this is known as "excess variation"—differences in results across studies that cannot be accounted for by sampling uncertainty. Excess variation is "the most common finding among the hundreds of meta-analyses conducted on economics subjects. . . . . The observed variation . . . [across studies] is always much greater than what one should expect from random sampling error alone" (Stanley and Doucouliagos 2012:80). Most of the differences between studies are not due to having different samples but rather having different models.

Given this, it is perhaps not surprising that the robustness of much published literature is open to question. There are several field areas where cutting-edge research has been subjected to careful replication with deeply disappointing conclusions. This includes research into the causes of cancer (Begley and Ellis 2012; Prinz, Schlange, and Asadullah 2011), genetics research on intelligence (Chabris et al. 2012), and the determinants of economic growth across countries (Sala-i-Martin et al. 2004). These are not marginal research lines but rather at their peak represented some of the most exciting research in their fields, produced by leading scholars and published in the top journals. In each of these areas, large portions of "exciting" and even "path breaking" research have turned out to be nonrobust, false-positive findings.

In psychology and behavioral genetics, a large accumulated literature has found evidence for genetic determinants of general intelligence, identifying at least 13 specific genes linked to IQ (Payton 2009). However, in comprehensive replication, applying the same core model to multiple large-scale data sets, a major interdisciplinary research team found that virtually all of these associations appear to be false positives (Chabris et al. 2012). Across 32 replication tests, only one gene yielded barely nominal significance. This is roughly the expected rate of significant findings when there are no true associations in the data.

Medical research has been an area with especially detailed replication efforts. Private-sector biotech laboratories look to the published literature for primary science findings that could be developed and scaled-up into new medicines and treatments. However, industry laboratories that try to replicate

published biomedical research often find the results are not robust and are unable to reproduce the findings. The biotech giant Amgen reported on 10 years of efforts to replicate 53 "landmark" studies that pointed to new cancer treatments. With its team of 100 scientists, only 11 percent of these studies could be replicated (Begley and Ellis 2012).[2] As an Amgen vice president noted, "on speaking with many investigators in academia and industry, we found widespread recognition" of the lack of robustness in primary medical research (Begley and Ellis 2012:532).

In macroeconomics, the literature on economic growth likewise appears thick with nonrobust results. In a set of now classic robustness studies, Sala-i-Martin (1997) and Sala-i-Martin et al. (2004) revisited 67 "known" determinants of national economic growth—variables that had been previously shown to have a significant effect on gross domestic product. Testing their robustness against sets of possible controls, only 18 growth determinants (roughly 25 percent) showed consistent, nontrivial effects; 46 of the variables were consistently weak and nonsignificant; some were significant in only 1 out of 1,000 regression models. There is now a widespread doubt of whether anything at all was learned from the extensive literature on cross-country economic growth (Durlauf, Johnson, and Temple 2005; Ciccone and Jarociński 2010).

All of this fits distressingly well with arguments in medicine (Ioannidis 2005) and psychology (Simmons, Nelson, Simonsohn 2011) that most published research findings are false positives, and that most empirical breakthroughs are actually dead ends.

There is, in summary, great need for robustness analyses that make research results more compelling and less prone to nonrobust, false-positive results. Such robustness analyses should aim to be developmental, transparent, and informative. As we will show, our framework advances each of these goals. First, it is developmental: It encourages analysts to consider a greater range of models than they otherwise would. Second, it is transparent: It reveals to readers a greater range of models than can be shown in conventional tables. Third, it is informative: It shows which model ingredients have greater or lesser influence on the reported results, so that analysts and readers alike know which assumptions (if any) are driving the results.

Moreover, our framework aims to have minimal costs of adoption and is designed as a complement, rather than replacement, to the current practices of applied sociological researchers. Our goal is for researchers to first conduct their analyses as they have always done and then adopt the multimodel computational robustness framework as an additional step to expand on their findings, support the credibility of their analysis, and to show confidence in their results.

Conceptual Foundations

Our approach to model robustness proceeds in two steps and has two key objectives:

(1) Show the extent to which empirical conclusions are driven by the data rather than the modeling assumptions. How many modeling assumptions can be relaxed without overturning the conclusions? This step is focused on computing the modeling distribution.
(2) If robustness testing finds conflicting results, which elements of the model specification are critical assumptions required to sustain a particular conclusion? In contrast, which modeling assumptions are noninfluential and do not affect the conclusions? This step conducts the model influence analysis.

We begin with step 1, calculating the distribution of estimates from a model space. We detail the logic and methods of the approach and illustrate the analysis using two empirical applications that show differing levels of model robustness. After building familiarity with the core approach, we proceed to the influence analysis of step 2: The decomposition of the modeling distribution, showing what elements of the model have greatest influence on the conclusions. In the final step, we combine these in a broad analysis of functional form robustness and model influence.

Degrees of Freedom: Defining the Model Space

A key step in robustness analysis is defining the model space—the set of plausible models that analysts are willing to consider. Our approach is to take a set of plausible model ingredients and populate the model space with all possible combinations of those ingredients. Each model ingredient has at least one alternative (e.g., logit vs. probit), which can be taken in combination with all other model elements (sets of controls, different outcome variables, etc.). We begin by focusing on the model space as defined by control variables (Leamer 2008; Raftery 1995; Sala-i-Martin 1997). With some additional complexity, this will be extended to alternative forms of the outcome variable, different forms of the variable of interest, different standard error calculations, and different possible estimation commands.

Control variables are a central strategy for causal identification in observational research (Heckman 2005). Yet, control variables are a common source of uncertainty and ambivalence. Rarely do the controls represent the exact processes of fine-grained theoretical expectations. As a result, adding

or dropping control variables is routine practice in ad hoc robustness testing. This is not without reason. When the "true model" is not actually known, control variables can have unpredictable consequences. Adding additional control variables to a model is often expected to reduce bias and lead to better results. However, this intuition holds only under highly stylized circumstances when the true model is completed by the additional control variable(s). When the model is wrong—when there are remaining unobserved variables—controlling for *some but not all* variables can increase bias just as well as reduce it (Clarke 2005, 2009). A misspecified model with 10 controls is not naturally better or less biased than a misspecified model with only five of those controls. Extra controls can leverage correlations with other omitted variables, amplifying omitted variable bias (Clarke 2005). Controls can also leverage backward causal linkages with the outcome, producing reverse causation or selection bias (Elwert and Winship 2014). Without knowing the full set of multiple correlations among all the measured and unmeasured variables in the true model, adding an additional control variable to an incomplete model can just as easily amplify as diminish omitted variable bias (Clarke 2005, 2009; Pearl 2011). Recognizing when a given control variable can lead the analysis astray is difficult and calls for "prudent substantive judgment, and well-founded prior knowledge" (Elwert and Winship 2014:49). Reporting results from many different combinations of controls relaxes the need for an author's judgment to be exactly correct and highlights situations when more judgment is needed.

For a robustness analysis, most control variables probably deserve some skepticism.[3] We should be skeptical of results that critically depend on a very specific constellation of control variables—especially when some of the controls lack strong a priori intuition or are themselves not statistically significant. Allowing all possible combinations of controls, in essence, generates random disruptions to an author's preferred specification. No exact specification in this modeling space is given particular or unique substantive justification. But we allow the possibility that a competent researcher could, with motivation, develop ad hoc but *plausible* reasons for favoring any one of the specifications. Moreover, running all combinations of controls allow one to observe which controls are critical to the analysis and thus deserving of additional scrutiny and judgment.

The downside cost of sustaining skepticism about model specification is computational demands. When using all possible combinations of controls, the modeling space increases exponentially. When there are p possible control variables, there are 2^p unique combinations of those variables. For three controls, there are $2^3 = 8$ possible combinations. With 17 possible control

variables, there are $2^{17} = 131{,}072$ unique possible models.[4] This is tractable, but eventually, too much uncertainty leads to more models than are computationally feasible.

From the Sampling Distribution to the Modeling Distribution

Classical statistics is focused on the quantification of uncertainty, in the form of standard errors and confidence intervals, but this is limited to uncertainty about the data stemming from random sampling. We expand on the concept of the sampling distribution to incorporate uncertainty about the model.

Consider a baseline regression model, $Y_i = \alpha + \beta X_i + \varepsilon_i$, in which after collecting a sample of data we compute an estimate b of the unknown parameter β. This single estimate b is not definitive, but based partly on random chance, since it derives from a random sample.

In classical statistics, there are thought to be K possible samples that could have been drawn $\{S_1, \ldots, S_K\}$, each of which yields a unique regression coefficient $\{b_1, \ldots, b_K\}$. In repeated sampling, we would draw many samples and compute many estimates which make up a *sampling distribution*. For clarity, the mean of the estimates is denoted as $\bar{b}$ and the standard deviation is $\sigma_S = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(b_k - \bar{b})^2}$. This sampling standard error, σ_S, indicates how much an estimate is expected to change if we draw a new sample. Actual repeated sampling is rarely undertaken, but parametric formulas and/or bootstrapping approximate this standard error, and are used to decide if an estimate b is statistically significant.

However, the sampling distribution critically assumes that the true model is known. What happens when we admit uncertainty about one or more aspects of model specification—when we are no longer confident about how to model the true "data generation process"? The key change is that there will be more than K estimates so that the sampling distribution alone does not convey the distribution of possible estimates.

When there is a range of possible methodological techniques that could reasonably be applied, this set of models provides not a point estimate but a modeling distribution of many possible estimates. The modeling distribution can be understood as analogous—and complementary—to the sampling distribution. Consider a set of plausible models $\{M_1, \ldots, M_J\}$ that might be applied to the data, each of which will yield its own unique estimate $\{b_1, \ldots, b_J\}$. In repeated modeling, we apply many different models to the data, and the resulting set of estimates forms the *modeling*

distribution. The average of these estimates is denoted as $\bar{b}$ and the standard deviation of the estimates is $\sigma_M = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(b_j - \bar{b})^2}$. We refer to σ_M as the *modeling standard error*. This shows how much the estimate is expected to change if we draw a new randomly selected model (from the defined list of J models).

To fully measure the overall uncertainty in our estimates, conceptually we take each possible sample $\{S_1, \ldots, S_K\}$, and for each sample estimate all plausible models $\{M_1, \ldots, M_J\}$, yielding $K \times J$ estimates b_{kj}. Then, we take the mean of these estimates ($\bar{\bar{b}}$), and compute the total standard error as

$$\sigma_T = \sqrt{\frac{1}{KJ}\sum_{k=1}^{K}\sum_{J=1}^{J}(b_{kj} - \bar{\bar{b}})^2}. \quad (1)$$

This expression for σ_T encompasses all the possible sources of variation in our estimates and includes all reasons why different researchers arrive at different conclusions: They either used a different sample, a different model, or both.

The analogy between sampling and modeling standard errors is imperfect. Under the usual ordinary least squares (OLS) or maximum likelihood assumptions, sampling standard errors are better understood than modeling standard errors.[5] Our approach to the modeling distribution is more similar to estimating the sampling distribution for nonlinear models when there is no analytical solution for the standard errors (Efron 1981; Efron and Tibshirani 1993).

The goal of the combined modeling and sampling standard error (σ_T) is to provide a more compelling gauge of what repeated research is likely to find—especially when repeated research involves different authors who may invoke different modeling assumptions. Rather than basing conclusions solely on sampling uncertainty, this incorporates model uncertainty as well. Combining an author's preferred estimate $b_{preferred}$ with the total standard error gives what we term the "robustness ratio": $= \frac{b_{preferred}}{\sigma_T}$. This is constructed as analogous to the t-statistic, but it is worth noting again that the underlying statistical properties of the ratio are not known, and will depend on the specified model space. We recommend the conventional critical values to guide interpretation (e.g., a robustness ratio of two or greater suggests robustness, by analogy to the t-statistic), but this is a coarse interpretation. To augment this, we use simple graphs of the distribution of estimates across models (i.e., the modeling distribution) for a visual inspection that is often very informative.

Other core summary statistics from the modeling distribution include the *sign stability* (the percentage of estimates that have the same sign) and the

significance rate (the percentage of models that report a statistically significant coefficient). Adapting Raftery's rule of thumb for multimodel inference (1995:146), we suggest that a significance rate of 50 percent sets a lower bound for "weak" robustness (i.e., at least 50 percent of the plausible models have a significant result). Likewise, when 95 percent of the plausible models have significant estimates, this indicates "strong" robustness.

Application 1: The Union Wage Premium

Before proceeding to more detailed aspects of model robustness, we illustrate the basic approach—robustness to the choice of controls—using a data set included in Stata, the 1988 wave of the National Longitudinal Survey of Women. We estimate the effect of union membership on wages (i.e., the union wage premium) controlling for 10 other variables that may be correlated with hourly wages (and union membership; (see Table 2). The coefficient on union, 11.1, means that union members earn about 11 percent more than nonunion members. This is on the low side of conventional estimates, which center around a 15 percent premium (Hirsch 2004).

Next, we report the robustness of this finding to the choice of control variables in the model. Does this finding hinge on sets of control variables, or do the findings hold regardless of what assumptions are made over the control variables? Table 3 shows that there are 1,024 unique combinations of the control variables. Running each of these models and storing all of the estimates, we graph the modeling distribution in Figure 1. The result appears strongly robust. The estimated coefficient on union membership is positive and significant in every possible combination of the control variables: both the sign stability and the significance rate are 100 percent. With this list of possible controls, and using OLS, it is not possible to find an opposite signed or even nonsignificant estimate. Figure 1 shows the modeling distribution as a density graph of all the estimates calculated; the vertical line marks the 11 percent wage premium estimate from Table 2. Estimates as low as 9 percent and as high as over 20 percent are possible in the model space.

As shown in Table 3, the average estimate across all of these models is 14.0. This simply represents the average coefficient across all models and is not necessarily the most theoretically defensible. The average sampling standard error is 2.4, and the modeling standard error is 2.5—uncertainty about the estimate derives equally from the data and from the model. The combined total (sampling and modeling) standard error is 3.5.[6] The robustness ratio—the mean estimate divided by the total standard error—is 4.05. By the standard of a *t*-test, this would be considered a strongly robust

Table 2. Determinants of Log Hourly Wage.

	Model: OLS	
Union member	11.1***	(2.2)
Usual hours worked	0.3**	(0.1)
Age	−0.6	(0.3)
Education (grade completed)	6.3***	(0.6)
College graduate	4.6	(3.6)
Married	1.1	(2.0)
Lives in south	−12.2***	(2.0)
Lives in metro area	22.4***	(2.3)
Lives in central city	−3.7	(2.3)
Total work experience	3.2***	(0.3)
Job tenure (years)	0.9***	(0.2)
Constant	56.5***	(15.0)
Observations	1,865	
Adjusted R-squared	0.408	

Note: Standard errors in parentheses. The outcome variable, log of hourly wages, has been scaled by 100 so that coefficients can be interpreted as percent changes in wages for a unit change in the predictor. OLS = ordinary least squares.
*$p < .05$. **$p < .01$. ***$p < .001$.

result, which agrees with the 100 percent sign stability and significance rates. Our conclusion is that, within the scope of these model ingredients, the positive union wage premium is a clear and strongly robust result. This suggests that the decline of unionization in America may well have contributed to middle-class wage stagnation—and not just for male workers (Rosenfeld 2014).

Presenting just one model (or a few) is a small slice of what is plausibly and sensibly reportable. The full modeling distribution (given this set of controls) gives a compelling demonstration of that fact.

Application 2: Mortgage Lending by Gender

Next, we draw on an influential study of discrimination in mortgage lending conducted by the Federal Reserve Bank of Boston (Munnell et al. 1996). What factors lead banks to approve an individual's mortgage application? The initial study focused on race, showing compelling evidence of discrimination against black applicants. In this application, we focus on the effect of an applicant's gender. We regress the mortgage application acceptance rate on a dummy for female as well as other variables capturing the demographic

Table 3. Model Robustness of the Union Wage Premium.

Linear regression			
Variable of interest: union			
Outcome variable: wage		Number of observations	1,865
Possible control terms: 10		Mean R^2	0.26
Number of models: 1,024		Multicollinearity	0.06
Model robustness statistics:		Significance testing	
Mean (b)	14.00	Sign stability	100%
Sampling SE	2.37	Significance rate	100%
Modeling SE	2.51		
Total SE	3.46	Positive	100%
		Positive and sig	100%
Robustness ratio:	4.05	Negative	0%
		Negative and sig	0%

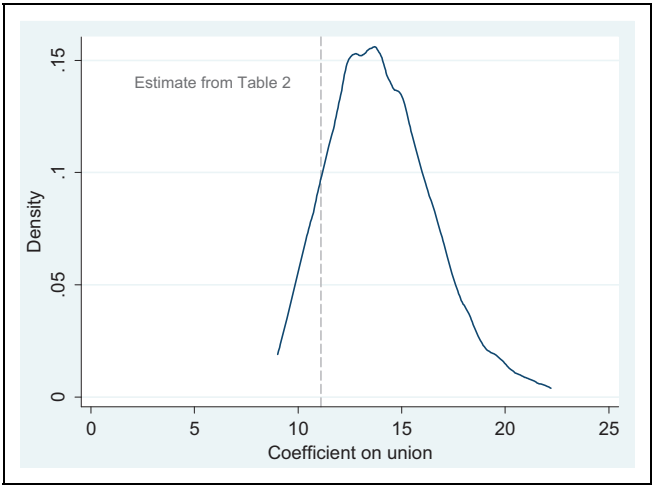

Figure 1. Modeling distribution of union wage premium.
Note: Kernel density graph of estimates from 1,024 models. Vertical line indicates the preferred estimate of an 11 percent union wage premium as reported in Table 2.

and financial characteristics of applicants. The results (Table 4) interestingly show that women are 3.7 percent more likely to be approved for a mortgage, suggesting banks favor female applicants—perhaps because women are seen as more prudent and responsible with household finances.

Table 4. Determinants of Mortgage Application Acceptance.

	Model: OLS	
Female	3.7*	(1.6)
Black	−11.4***	(1.8)
Housing expense ratio	5.8	(10.5)
Self-employed	5.6**	(1.8)
Married	4.6***	(1.3)
Bad credit history	−25.2***	(2.3)
Payment-income ratio	−50.2***	(9.3)
Loan-to-value ratio	11.9***	(3.4)
Denied mortgage insurance	−71.2***	(4.2)
Constant	113.8***	(3.4)
N	2,355	
Adjusted R-squared	0.226	

Note: Standard errors in parentheses. The outcome variable, mortgage acceptance (1 = accepted, 0 = denied), has been scaled by 100 so that coefficients can be interpreted as percent changes in the acceptance rate for a unit change in the predictor.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 5. Model Robustness of the Gender Effect on Mortgage Lending.

Linear regression			
Variable of interest: female			
Outcome variable: acceptance		Number of observations	2,355
Possible control terms: 8		Mean R^2	0.13
Number of models: 256		Multicollinearity	0.19
Model robustness statistics:		Significance testing:	
Mean estimate	2.29	Sign stability	88%
Sampling SE	1.61	Significance rate	25%
Modeling SE	1.60		
Total SE	2.27	Positive	88%
		Positive and sig	25%
Robustness ratio:	1.01	Negative	12%
		Negative and sig	0%

However, when we relax the assumption that any one of these control variables must be in the model—allowing us to consider all possible combinations of the controls—there is much uncertainty about the estimate. Table 5 reports the model robustness results. Across the 256 possible combinations of controls, the effect of gender is typically positive but only 25

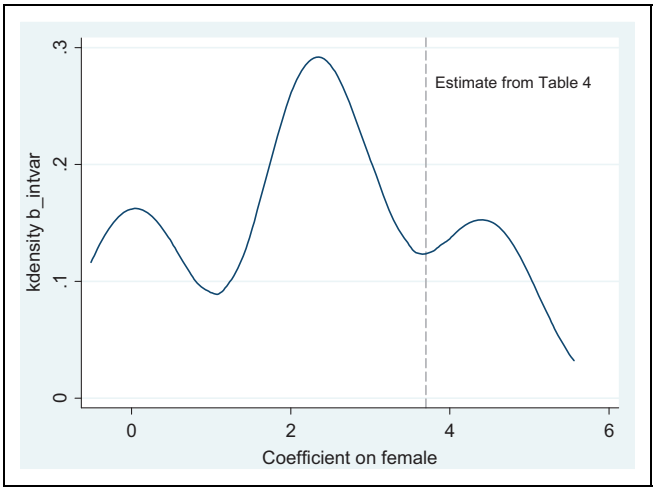

Figure 2. Modeling distribution of the gender effect on mortgage lending.
Note: Kernel density graph of estimates from 256 models. See Table 5 for more information about the distribution. The vertical line shows the preferred estimate from Table 4 (3.7 percent higher acceptance rate for women).

percent of the estimates are statistically significant. And 12 percent of the estimates have the opposite sign (though none of those estimates are significant).[7]

The mean estimate from all models is 2.29 and the average sampling standard error is 1.61—indicating that the mean estimate is not statistically significant. In addition, the modeling standard error is 1.60—the estimates vary across models just as much as would be expected from drawing new samples. The total standard error—incorporating both sampling and modeling variance—is 2.27, roughly the same size as the estimate itself, yielding a robustness ratio of 1.01.[8]

Figure 2 shows the distribution of estimates from all the 256 models with a vertical line showing the "preferred estimate" of 3.7 percent from Table 6. The modeling distribution is multimodal with clusters of estimates around zero, 2.3, and 4.5 percent. It seems hard to draw substantive conclusions from the evidence without knowing more about the modeling distribution. Why do these estimates vary so much? Why is the distribution so non-normal? What combinations of control variables are critical to finding a positive and significant result? These questions lead us to the next stage in our analysis: understanding model influence.

Model Influence: $\Delta\beta$ as the Effect of Interest

Model influence analysis focuses on how the introduction a control variable (or more broadly any model ingredient) changes the coefficient of interest. After calculating all models in the specified model space, influence analysis dissects the determinants of variation across models. Current research practice does a poor job of showing which model assumptions influence the conclusions.

In conventional analysis, it is standard to report the effect of a control variable (Z_i) on the outcome (Y_i). However, if Z_i is truly a control variable, then this coefficient is not directly interesting. The focus should be on how including Z_i *influences* the coefficient of interest.

To anchor this discussion, consider two simple nested models:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \,. \tag{2}$$

$$Y_i = \alpha + \beta^* X_i + \delta Z_i + \varepsilon_i^* \,. \tag{3}$$

We are interested in how changes in X_i affect the outcome, so β is the coefficient of interest. In equation (3), Z_i is a control variable, and its relationship to the outcome, Y_i, is given by δ.[9] When considering control variables, it is conventional to report the δ estimates. But what we most want to know is the *change in* β*:* the difference ($\Delta\beta = \beta^* - \beta$) caused by including the control. We define $\Delta\beta$ as the influence of including Z_i in the model, or simply the *model influence* of Z_i.

Model influence can be directly inferred in the case where there is only one control variable. Indeed, in this specific case, the significance test for $\Delta\beta$ is equal to the usual *t*-test for δ (Clogg, Petkova, and Haritou 1995:1275). However, when there is more than one control variable, the $\Delta\beta$ associated with each control is not observed, and the δ coefficients and their *t*-tests give little guide to which control variables are influential. As a result, it is often unclear what controls (if any) are critical to obtaining a given estimate for β.

The influence of Z_i on the coefficient of interest β is only partly due to the relationship between Z_i and Y_i (i.e., the reported estimate of δ). It is also a function of the correlation between Z_i and X_i, as well as the joint relation of Z_i and X_i with the unknown error term ϵ_i (Clarke 2005, 2009; Pearl 2011). Thus, control variables that have the greatest influence on β may not necessarily have a strong or statistically significant relationship with Y_i, and may look relatively "unimportant" in the main regression.[10] Similarly, control variables that are highly significant in the main regression may have little or no influence on the estimate of interest. The central purpose of including Z_i as a control is not captured in standard regression tables.

To estimate model influence, we draw on established techniques of identifying outlier observations: the Cook's D approach (Andersen 2008; Cook 1977). In a Cook's D analysis, influence scores for each data point are calculated by excluding observations one at a time, and testing how the exclusion of each observation affects the regression estimate. If the exclusion of one specific observation has a "large" effect on the regression coefficient that observation is considered influential and flagged for further inspection and evaluation. We operationalize a similar strategy to calculate an influence score for each control variable (and ultimately, other aspects of model specification). However, rather than simply exclude each variable one at a time, we test all combinations of the controls.

Using results from the full 2^p estimated models, we ask what elements of the model specification are most influential for the results. We formulate an *influence regression* by using the estimated coefficients (for the variable of interest) as the outcome to be explained. The explanatory variables in the influence regression are dummies for the original control variables. For P possible control variables, we create a set of dummy variables $\{D_1, \ldots, D_P\}$ to indicate when each control variable is in the model that generated the estimate. OLS regression then reports the marginal effect of including each variable. For P regressors, there are $J = 2^p$ observations (i.e., coefficient estimates). The influence regression is:

$$b_j = \alpha + \theta_1 D_{1j} + \theta_2 D_{2j} + \ldots + \theta_P D_{Pj} + \varepsilon_j, \tag{4}$$

in which b_j is the regression estimate from the jth model. The influence coefficient θ_1 shows the expected change in the coefficient of interest (b_j) if the control variable corresponding to D_1 is included in the jth model. Each coefficient estimates the conditional mean $\Delta\beta$ effect for each control variable. We offer no explicit definition of a "large" influence; as an intuitive guide, we report the percentage change in the coefficient of interest associated with including each control variable. This, in our view, is the main statistic analysts and readers need to know about the impact of a control variable: How does including each control variable, on average, affect the coefficient of interest?

Returning to the mortgage lending study offers an excellent case in point. Banks appear more likely to approve mortgage applications from women than men but in a robustness analysis that treats all control variables as uncertain, this effect is significant in only 25 percent of models. What model ingredients are driving these findings? Do the larger estimates and more significant results derive from especially compelling model specifications?

Which control variables or model assumptions are critical to a strong conclusion?

Influence Analysis of the Gender Effect in Mortgage Lending

For the mortgage lending analysis, Table 6 shows the influence of control variables on the coefficient of interest (female). The $\Delta\beta$ effect of controls is reported in order of absolute magnitude influence. To aid interpretation, we also report $\Delta\beta$ as a percent change in the estimate from the mean of the modeling distribution (2.29 as in Table 7). Two control variables clearly stand out as most influential: marital status and race. The influence estimate for marriage shows that, all else equal, when controlling for marital status the coefficient on female increases by 2.47, more than doubling the mean estimate across all models. Controlling for race (with the dummy variable "black") also increases the effect size of gender by 1.91, a full 83 percent higher than the mean estimate. The other controls have much less impact on the estimate and have little model influence.

In essence, there are two distinct modeling distributions to consider which are plotted in Figure 3. In one set of models, the controls for race and marital status are always *excluded* but all other controls are allowed in the model space (which gives 128 models). Under these assumptions, the estimates of the gender effect are tightly centered around zero, with an almost even split between positive (52 percent) and negative (48 percent) estimates, none of which are statistically significant. Here, there is no evidence at all for a gender effect. In contrast, the second distribution is defined by the opposite assumption: race and marital status must be in the model, but all combinations of the other controls are possible. Under these assumptions, the estimates cluster around a 4.5 percent higher mortgage acceptance rate for women. Both the significance rate and the sign stability are 100 percent—complete robustness. In order to draw robust conclusions from these data, one must make a substantive judgment about two key modeling assumptions: the inclusion of race and marital status. None of the other model ingredients affect the basic conclusion. These two model assumptions determine the results.

The influence analysis does not tell us which assumptions are correct but simply which ones are critical to the findings. We would simply point out that these influential controls (race and marital status) are variables that scholars of gender and inequality would, a priori, consider important rather than arbitrary to include in the model (though financial economists might

Table 6. Model Influence Results for Gender Effect on Mortgage Lending.

	Effect of variable inclusion	Percentage change from mean estimate
Married	2.47	107.8%
Black	1.91	83.3%
Self-employed	−0.30	−13.3%
Loan-to-value ratio	−0.25	−10.7%
Bad credit history	−0.23	−10.1%
Housing expense ratio	0.19	8.4%
Payment-income ratio	−0.18	−8.1%
Denied mortgage insurance	−0.03	−1.1%
Constant	0.50	
R-squared	0.98	

Note: Based on 256 estimates reported in Table 5.

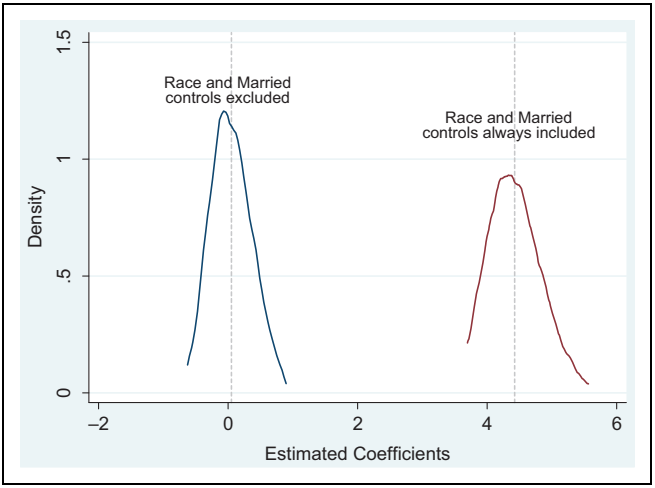

Figure 3. Modeling distributions for the gender effect under different assumptions.

overlook them). Indeed, further unpacking indicates that *among single applicants*, banks favor women over men and especially favor black women over black men. These patterns are part of why the marriage and race variables are so critical to model robustness.[11]

Research articles often seek to tell a "perfect story" with an unblemished set of supportive evidence. Yet, acknowledging ambiguity in empirical

results can lead to deeper thinking and greater insight into the social process at work. In a framework that emphasizes model robustness, we need greater tolerance for conflicting results and more willingness to reveal the factors that are critical to a given finding. Model influence analysis takes us well beyond the robustness results or a simple model averaging approach: We can see which assumptions matter, evaluate their merits, and explore their implications. The fact that model robustness is contingent on key controls illuminates greater insight and greater appreciation of subtleties in gender dynamics in the mortgage lending market.

One final observation highlights the critical difference between the significance of a control variable and its model influence. The variable most significant in the main regression (from Table 6) is having been denied mortgage insurance by a third-party insurer. When banks see such an applicant, they almost never approve a mortgage application. However, this variable is also the *least influential* control. Similarly, a bad credit history has a striking effect on lending decisions, reducing the approval rate by 25 percent. Yet, credit history has very little model influence and has no real bearing on the conclusions about the gender effect. Moreover, the variables that are critically influential (race and marital status) had modest coefficients in the main regression and did not stand out as key determinants of mortgage lending. Influential variables may be nonsignificant, and significant variables may well be noninfluential. Insight into which control variables are critical to the analysis is not visible in a conventional regression table. This is a transparent flaw in conventional regression tables that can be readily corrected with multimodel influence analysis.

Functional Form Robustness

The question of model robustness extends well beyond the choice of control variables. How robust are empirical results to different functional forms such as different estimation commands and variable constructions? Often, there are many credible ways of conceptualizing and measuring core concepts such as ''inequality'' (Leigh 2007; van Raalte and Caswell 2013), ''globalization'' (Brady, Beckfield, and Seeleib-Kaiser 2005), or ''social capital'' (Lochner, Kawachi, and Kennedy 1999). Capturing uncertainty about the measurement of outcome variables, Wildeman and Turney (2014) test the effect of parental incarceration on 21 different measures of children's behavioral problems. In a study of how globalization affects the welfare state, Brady et al. (2005) note that ''the measurement of globalization is contested and that the literature has yet to converge on a single measure'' (928);

embracing this uncertainty, they test 17 different measures of globalization (including trade openness, foreign direct investment, migration, and the like). Moreover, for any particular variable, there can be many alternative functional form specifications. Educational attainment, for example, has been tested across 13 different functional forms—ranging from linear years of schooling, to sets of dummies for degree completion, to splines, and combinations thereof—each of which map on to unique hypotheses of how education affects mortality (Montez, Hummer, and Hayward 2012). Finally, these combinations of variables can be connected together in different link functions and estimation commands. Brand and Halaby (2006) show the similarity of estimates from OLS and matching for the effect of elite college attendance on seven career outcome variables. In a study of teen childbearing, researchers emphasize the variation across estimates when using OLS, propensity score matching, parametric, and semiparametric maximum likelihood models, which helped to clarify why past studies had shown such mixed results (Kane et al. 2013).

The existing literature on multimodel inference has been primarily focused on choice of controls, with little focus on functional form robustness (e.g., Ciccone and Jarociński 2010; Leamer 2008; Raftery 1995; Sala-i-Martin et al. 2004). Functional form robustness is less combinatorially tractable and requires much more specific input from applied researchers, requiring the specification of alternatives for each model ingredient. However, our approach provides a machinery to layer functional form robustness over top the core control variable robustness. This allows us to examine the intersection of every control set with every specified functional form.

One critical detail to note is that functional forms typically offer strict alternatives not lists of possible combinations. Instead of combinations, the approach is one of "either/or" alternatives. When choosing among three control variables, all possible combinations of the three can be estimated. However, when choosing among three link functions—OLS log linear, Poisson, and negative binomial—the methods cannot be used in combination. One could use *either* OLS, *or* Poisson, *or* negative binomial but combinations thereof are not possible. The same is true for variable definitions and other aspects of functional form. Consider multiple operational definitions of a variable (X_i and X_i'), such as inequality measured either as the Gini index (X_i) or the share of income held by the top 1 percent (X_i'). The functional form robustness analysis tests the stability of results across the alternative measurements (X_i or X_i') but excludes models that include *both versions* of the variable. Models including both terms would give the effect of X_i (Gini index) holding constant X_i' (the top 1 percent share). This is quite different from a robustness

analysis and would typically neutralize the analysis by partialling out most of the variation in X_i. This distinction between strict alternatives and combinations is a simple but important element in implementing functional form robustness. We detail our algorithm in the Online Appendix S1.

In the following empirical application, the coefficients across functional form specifications are in comparable units. These models all return b_J estimates that have the same meaning. However, this will not always be the case. For example, when comparing across linear probability, logit, and probit models, the coefficients all express different quantities. In Online Appendix S2, we extend the robustness and influence analyses to settings where the resulting coefficients are not directly comparable, focusing on the signs and significance tests across different functional forms.[12]

Application 3: Tax-induced Migration

In our final application, we bring together functional form robustness and influence analysis in a study of tax-induced migration across U.S. states. Do higher income tax rates cause taxpayers to "vote with their feet" and migrate to states with lower taxes (Kleven, Landais, and Saez 2013; Young and Varner 2011)? For this analysis, we construct an aggregate 51×51 state-to-state migration matrix using data from the 2008 to 2012 American Community Survey (ACS). We also use comparable migration data from administrative tax returns provided by the Internal Revenue Service (IRS) over the years 1999 to 2011 (Gross 2003). To analyze these data, we draw on a gravity model of migration (Conway and Rork 2012; Herting, Grusky, and Rompaey 1997; Santos Silva and Tenreyo 2006). The number of migrants (Mig_{ij}) from state i (origin) to state j (destination) is a function of the size of the base populations in each state (Pop_i and Pop_j), the distance between the states ($Distance_{ij}$), and a variable indicating if the states $\{i, j\}$ have a shared border ($Contiguity_{ij}$). These are the core elements that define the basic laws of gravity for interstate migration (e.g., Santos Silva and Tenreyro 2006). To this core model, we add the difference in income tax rates between each state pair ($Tax\_Difference_{ij}$) as our variable of interest (the tax effect). Finally, we specify this as a log-linear model, taking logs of the right-hand side count variables and estimating with Poisson:

$$\begin{aligned} Mig_{ij} = \exp(\alpha + \beta_1 \log Pop_i + \beta_2 \log Pop_j + \beta_3 \log Distance_{ij} \\ + \beta_4 Contiguity_{ij} + \beta_5 Tax\_Difference_{ij}) + \varepsilon_{ij} \end{aligned} \quad (5)$$

The coefficients from the log-linear model give the semi-elasticity of migration counts with respect to the tax rate—the percent change in migration flows for each percentage point difference in the tax rates.

In Table 7, we show our main analysis. Model 1 includes just the base populations of the origin and destination states and the income tax differences between them. When the income tax rate in the origin state is higher, there tends to be more migration from the origin state to other (lower tax) destinations. Migration flows are 1.4 percent higher for each percentage point difference in income tax, but the estimate is not statistically significant. Model 2 adds in controls for contiguity, distance, the sales and property tax rates, state income, and a measure of natural amenities (topographical/landscape variability). The tax effect is now larger and statistically significant. For each one-point difference in the tax rate, migration flows are 2.4 percent higher. Finally, in model 3, when using an IRS migration data with *the same set* of controls, we find a similar significant effect. This gives seemingly compelling evidence that high income taxes cause migration to lower tax states.

What this fails to show, however, is the extreme model dependence in this conclusion. Models 2 and 3 are knife-edge specifications, carefully selected to report statistically significant results, and remarkably unrepresentative of the overall modeling distribution. Both models are highly sensitive to adding or deleting insignificant controls, and this set of controls is the only combination among many thousands that yields a significant result in both the ACS and IRS data.

We embrace a wide robustness analysis that relaxes assumptions about possible controls, possible data sources for migration, and alternative estimation commands. There are two controls that we see as absolutely critical to the gravity model: base populations of the origin and destination states. Combinatorially including or excluding these variables produces models that we regard as nonsense, so we impose the assumption that they must be in all models. However, we leave as debatable the controls for distance, contiguity, other tax rates, economic performance of the states, and a rich set of natural amenities which have been previously shown to influence migration (McGranahan 1999). All possible combinations of these controls give 4,096 models. Moreover, we test these models across the two alternative data sets for migration and population (ACS and IRS), and across three different estimation strategies (Poisson, negative binomial, and OLS log-linear). For each data set, there are three possible estimation commands, and for each (data set $\times$ estimation command), there are 4,096 possible sets of controls. This robustness analysis, therefore, runs 24,576 plausible models.

Table 7. Determinants of Cross-state Migration Poisson Models.

	Model 1 ACS		Model 2 ACS		Model 3 IRS	
Income tax difference	1.38	(1.53)	2.42*	(1.23)	3.00*	(1.33)
Population—origin	0.79***	(0.04)	0.83***	(0.03)	0.82***	(0.03)
Population—destination	0.73***	(0.03)	0.81***	(0.03)	0.80***	(0.03)
Log distance			−0.32***	(0.04)	−0.30***	(0.03)
Contiguity			1.09***	(0.07)	1.10***	(0.07)
Sales tax difference			0.02	(0.01)	0.02	(0.01)
Property tax difference			0.02	(0.05)	0.07	(0.05)
Avg. income			0.01**	(0.00)	0.01	(0.00)
Natural amenities (landscape)			−0.00	(0.00)	−0.00	(0.00)
Constant	−16.22***	(0.86)	−18.64***	(0.85)	−18.30***	(0.87)
N	2,015		2,015		2,015	
Pseudo R^2	0.525		0.788		0.780	

Note: Robust standard errors in parentheses. ACS = American Community Survey data; IRS = Internal Revenue Service data.
*$p < .05$. **$p < .01$. ***$p < .001$.

As shown in Table 8, the tax coefficient is statistically significant in only 1.5 percent of all models. The mean estimate is almost exactly zero, and estimates are evenly split between positive tax flight estimates (48.9 percent) and wrong-signed negative estimates (51.1 percent). Among the few statistically significant results, the great majority are wrong signed: estimates with negative signs indicate migration toward *higher tax* states. Only 0.2 percent of estimates are significantly positive compared to 1.3 percent that are significant and wrong signed. The robustness ratio—the mean estimate divided by the total standard error—is 0.01. The modeling distribution is relatively normal: There are no critically important modeling decisions that generate bimodality in the estimates. As shown in Figure 4, the significant estimates reported in Table 7 above are extreme outliers in the modeling distribution.

In this case, when the robustness analysis is so overwhelmingly nonsupportive, the influence analysis has less to work with. However, there are a few informative points. Compared to Poisson, the negative binomial and OLS log-linear models give less positive estimates. Estimates from the models using IRS rather than ACS data are more positive. This suggests that the most supportive evidence will come from using Poisson with the IRS data (reported as model 3 above), and the least supportive evidence will come from using OLS log-linear models with ACS data. Yet, even when we narrow our robustness

Table 8. Model Robustness of Tax Migration.

Variable of interest: income tax rate		Number of models	24,576
Outcome variable: migration		Number of observations	2,015
Possible control terms: 17		Mean R^2	.479
Model robustness statistics:		Significance testing	
Mean (b)	0.01	Sign stability	51.9%
Sampling SE	1.10	Significance rate	1.5%
Modeling SE	0.83		
Total SE	1.38	Positive	48.9%
		Positive and sig	0.2%
Robustness ratio:	0.01	Negative	51.1%
		Negative and sig	1.3%

Note: SE = standard error.

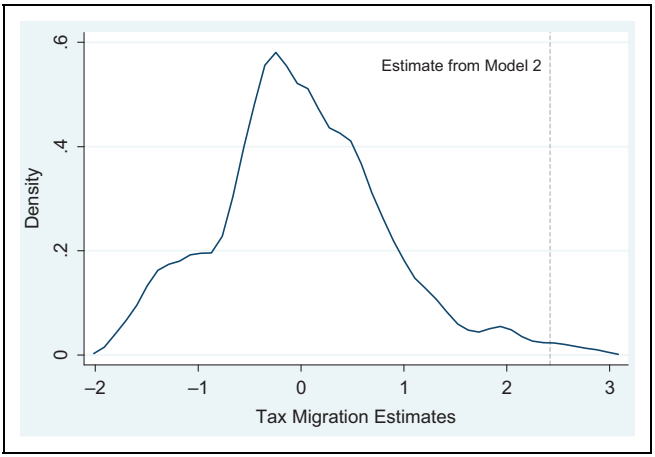

Figure 4. Modeling distribution of tax migration estimates.
Note: Kernel density graph of estimates from 24,576 models.

testing to the most supportive estimator (Poisson) and data set (IRS), there is weak support: while the sign stability is 100 percent, the income tax effect is significant in only 1 percent of those models.[13] By control variables, the sales tax rate, average income, and the property tax rate have the most positive influence—generating more positive estimates of tax flight when these controls are included. (Note, however, that none of these controls were significant

Table 9. Influence Analysis of Tax Migration Estimates.

	Marginal Effect of Specification	Percentage change from null model
Estimation command		
Poisson	*Reference category*	
Negative Binomial	−0.73	−80.2%
OLS log-linear	−1.31	−144.2%
Data source		
Am. Community Survey	*Reference category*	
IRS tax returns	0.645	70.9%
Control Variables		
Temperature diff (winter to summer)	−0.435	−47.8%
Winter temperature	−0.411	−45.1%
Sales tax rate	0.340	37.4%
Summer humidity	−0.225	−24.7%
Winter sun	−0.149	−16.4%
Unemployment rate	−0.125	−13.8%
Avg. income	0.106	11.6%
Contiguity	−0.092	−10.1%
Log distance (between states)	−0.057	−6.3%
Property tax rate	0.049	5.4%
Coastal/water access	−0.044	−4.8%
Topographical/landscape variation	−0.040	−4.4%
Constant	0.910	
R-squared	0.785	

Note: The number of models is given by $(2^{12}) \times 2 \times 3 = 24{,}576$. This 12 control variables, two data sets, and three estimation commands. The null model in the influence regression is the reference category data and estimation command, with no control variables, given by the constant in the influence regression.

in model 3.) All other controls push the tax migration estimate toward a zero or wrong-signed result, and virtually *must* be excluded to support the hypothesis.

In these results, we see another case where the most significant control has among the least model influence. In the main regression models 2 and 3, distance between the states is a powerful predictor of migration flows, showing *t*-statistics greater than 10. Yet, including distance in the model has almost no influence on the tax migration estimate (−6.3 percent in Table 9).

While it is possible to support the tax flight hypothesis with a few knife-edge model specifications, there is remarkably little support even in a more narrow and supportive robustness analysis. This shows how extreme the difference can be between a curated selection of regression results (Table 7) and a rigorous robustness analysis (Table 8). While one offers an existence proof that a significant

result can be found, the weight of the evidence from many credible models gives scant support to the tax migration hypothesis. It remains technically possible that the one-in-a-thousand specifications of Table 7 present the best, most theoretically compelling estimates. If so, authors would need to carefully explain to readers why such painstakingly exact model assumptions are required, and why virtually *any* departure from model 2 or 3 fails to support the conclusions.

Discussion: Comparison with Model Averaging

Finally, we consider our multimodel analysis in light of the existing approaches of model averaging. Our work builds directly on foundations laid in research on model averaging (Berk, Brown, and Zhao 2009; Efron 2014; Hoeting et al. 1999; Raftery 1995), which allows researchers to succinctly summarize the results of many alternative models. Model averaging, as in our approach, requires specifying a model space and estimating all of those models. However, model averaging remains focused on the goal of presenting a definitive, "one best estimate." Focusing on the modeling distribution—emphasizing the spread of estimates and the specific model assumptions that produce them—advances the project of multimodel analysis to be more compelling, transparent, and intuitive than model averaging alone.

Model averaging has had limited take up in applied social science, we think in large part because applied researchers believe in the approach of developing and reporting a substantive, preferred estimate. Rather than attempting to replace an author's preferred estimate with "something better," we embrace the preferred estimate and use it as the starting point for understanding model robustness. Our framework expects users to engage in their own process for developing their preferred model. Then, we ask two questions: (1) how many model assumptions can be relaxed without overturning the conclusion from that estimate? and (2) which model assumptions are most critical to the results?

Model averaging glosses over the question of influence and prematurely closes the conversation about critical modeling assumptions. If a model-averaged estimate is close to zero, there is little pathway for a conversation about the merits of different modeling choices. Likewise, if a model-averaged estimate is large in magnitude, the conclusion appears robust even if some important sets of models report conflicting results. Our influence analysis shines light on which aspects of model specification should be treated as uncontroversial and which model ingredients deserve more careful attention and detailed justification.

Model averaging approaches typically weight the estimates either by model fit or by Bayesian priors. We focus on the raw (unweighted) distribution of

estimates, as a way of revealing, not what is the best estimate, but rather what estimates can be obtained from the data. Weighting the estimates inevitably requires additional assumptions. A simple reason for having a high model fit is the unfortunate inclusion of an endogenous variable that is jointly determined with the outcome (Elwert and Winship 2014; Pearl 2011; Sala-i-Martin 1997:180). Weighting by model fit assumes that none of the control variables are endogenous. In other words, metrics of model fit require, for their validity, exogeneity assumptions that we wish to treat as open to question. Similarly, weighting by Bayesian priors (representing an author's beliefs about model validity) privileges a given set of model assumptions. However, the purpose of robustness analysis is to demonstrate how results may change under *different* beliefs about the correct model. Weighting the estimates is a valid way of incorporating an author's own (private) uncertainty about model specification, but the approach is not especially transparent. Weighted estimates do not allow alternative views of key model assumptions and do not address the asymmetry of information between analyst and reader. When focusing on the modeling distribution, the raw unweighted estimates require the fewest assumptions for understanding how model specification affects the results.

Finally, since the foundational work of Raftery (1995) and Sala-i-Martin (1997), model averaging has long been limited to sets of control variables (c.f. Efron 2014; Ho et al. 2007). We make concrete advances in computationally developing the model space, allowing alternative functional forms, standard error calculations, variable definitions, and estimation commands to be part of the model space. The development here has been to augment the "all combinations" algorithm (which does not apply to functional form robustness) to incorporate strict either/or alternatives for functional form possibilities. This allows the combinatorial intersection of control variables with functional forms. One limitation is that the coefficient of interest can sometimes be dramatically rescaled in functional form transformations, making the average estimate uninterpretable. However, in such cases the signs and significance tests are still directly comparable as we explain in Online Appendix S2. This addresses a central limitation in existing approaches to multimodel analysis and pushes the computational robustness project into a new territory of addressing uncertainty about functional form.

Conclusion

Empirical research is often described as "data analysis." This is something of a misnomer, since what is being analyzed is how model assumptions combine with data to produce estimates. While the data are often external to the

researcher, the model assumptions are not. It is often unclear how much results are given by the data and how much they are given by the model (Glaeser 2008; Leamer 1983; Young 2009). "The modeling assumptions," as Durlauf, Fu, and Navarro state, "can control the findings of an empirical exercise" (2013:120). Relaxing these modeling assumptions makes results more empirical, less model dependent, and focuses attention on the model ingredients that are critical to the results.

Uncertainty about model specification is no less fundamental than uncertainty about sample data. We emphasize the conceptual analogy between the *sampling* distribution and the *modeling* distribution. While the sampling distribution shows whether a point estimate is statistically significant (i.e., different from zero), the modeling distribution shows whether it is different from those of other plausible models. Together, these address the two fundamental sources of uncertainty about parameter estimates.

A point estimate, we argue, represents a bundle of exact model assumptions. Relaxing these assumptions about the choice of controls, functional forms, estimation commands, or variable definitions allows many plausible models and yields a modeling distribution of estimates. How many model assumptions can be relaxed without overturning an empirical conclusion? What is the range and distribution of plausible estimates from alternative models? Which model assumptions are most important?

The current norm in top journals of reporting a handful of ad hoc robustness checks is weakly informative and lags behind the reality of modern computational power. Our framework and statistical software provide a flexible tool to demonstrate the robustness of an estimate across a large set of plausible models, enabling more efficient and rigorous robustness testing, and inviting greater transparency in statistical research.

In our empirical applications, we have shown that multimodel analysis can turn out strongly robust to the choice of controls (as in the union wage premium) or reveal extreme model dependence where the conclusions are sustained in less than one in a 100 alternative models (as with tax migration). Somewhere in between lies limited or mixed robustness, in which one or two critical modeling judgments must be made in order to draw conclusions from the data (as for gender effects in mortgage lending).

Model robustness is fundamentally about model transparency, with the goal of reducing the problem of asymmetric information between analyst and reader (Young 2009). If an author's preferred result is an extreme estimate, readers should know this, and it is incumbent on the author to explain why the preferred estimate is superior to those from other readily available models. This advances both the underlying goals of science and readers'

understanding of the research. Often preferred results *are* robust across alternative models, and in such cases our framework provides a simple and compelling way to convey this to readers. And even when results critically depend on one or two model ingredients, this can yield new insight into the social process in question, deepening the empirical findings.

Multimodel analysis also allows researchers to unbundle their model specifications and observe the influence of each model ingredient. In conventional regression tables, the influence of model ingredients is either opaque or completely unknown. Typically, analysts report the effect that a control variable (Z_i) has on the outcome (Y_i). In practice, what should be reported is the effect that a control variable has on the *conclusions* (i.e., how including Z_i *influences* the coefficient of interest). We describe this influence effect as $\Delta\beta$: the change in the coefficient of interest associated with each model ingredient. We show repeatedly that the statistical significance of control variables gives limited indication of their influence on the conclusions: in our empirical applications, the most significant controls often have little or no influence on the coefficient of interest and often it is the nonsignificant, seemingly unimportant controls that have surprisingly strong influence. Our model influence analysis shows what control variables are critical to the results, and this extends readily to other aspects of model specification.

Robustness analysis helps to simulate the process of repeated study and bring into the analysis what skeptical replicators might find. This, in turn, points to a key reason why authors tend to avoid wide robustness testing: allowing many disturbances to an author's preferred specification creates strong potential that at least some of the models will fail to achieve significance or have the "wrong" sign. In publication, authors prefer to report—and reviewers and readers prefer seeing—a wall of confirming evidence for a hypothesis. In rigorous multimodel analysis, we need greater tolerance for "imperfect stories" and more focus on the weight of the evidence.

Finally, we encourage a tone of modesty in conclusions about the robustness of research results. Causal inference, as Heckman has noted, is provisional in nature because it depends on a priori assumptions that, even if currently accepted, may be called into question in the future (Heckman 2005). Robustness has a similarly provisional nature. In particular, the potential model space is not only large but also open ended—new additions to the model space can always be considered. We aim for robustness analysis that is developmental and compelling but accept that it is never definitively complete. We focus on robustness to concrete methodological concerns, rather than generic robustness to all conceivable alternative modeling strategies.

For example, none of the applications in this article have specifically addressed unobserved heterogeneity—potential bias from unmeasured variables. However, models that address this concern would be a valuable ingredient in a future, even wider robustness analysis. Such models include instrumental variables (IV), Heckman selection estimators, fixed effects models, and difference-in-differences estimators. IV regression, for example, can control for unobserved variables, or even reverse causation, under strong assumptions of instrument exogeneity and relevance (Angrist and Krueger 2001; Heckman 2005). However, IV estimation creates second-order questions of uncertainty about the choice of possible instruments, and uncertainty about how well the instruments meet critical relevance and exogeneity assumptions (e.g., Hahn and Hausman 2003). More complex models, as Glaeser (2008) notes, give researchers more degrees of freedom in technical specification, are less transparent to readers, and allow greater range for analysts to discover and report a nonrobust preferred estimate. Incorporating such models can add great richness to a multimodel analysis, but they simultaneously make robustness testing all the more important.

In the future, we believe model robustness will be at least as important as statistical significance in the evaluation of empirical results and reporting extensive robustness tests will be a strong signal of research quality. In a world with growing computational power and increasingly broad menus of statistical techniques, multimodel analysis can make research results more compelling and less dependent on idiosyncratic assumptions—and in the process, allow the empirical evidence to shine in new ways.

Authors' Note

Software download instructions: For software and replication materials, run the Stata do file install_mrobust.do (available on Young's website). This installs the program, loads in data sets, and runs all the analyses in this article.

Acknowledgment

The authors thank Michelle Jackson, Adam Slez, Gary King, Scott Long, Tomas Jimenez, Aliya Saperstein, Ariela Schachter, Erin Cumberworth, Christof Brandtner, and Patricia Young for helpful feedback and suggestions. John Muñoz provided valuable research assistance. Young especially wishes to thank four cohorts of graduate students in Sociology 382 for their curiosity and questions that have pushed this project to achieve much greater clarity. Send comments to cristobal.young@stanford.edu.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

The online appendices are available at http://journals.sagepub.com/doi/suppl/10.1177/0049124115610347.

Notes

1. Gary King and colleagues (Ho et al. 2007; King and Zeng 2006) have laid out a similar concept of model dependence: how much one's empirical results depend on model specification.
2. At the German pharmaceutical Bayer laboratories, replications often involved three to four scientists working for six to twelve months on a study. The company recently reported that in its efforts at replication, two-thirds of the published findings it studied could not be supported (Prinz et al. 2011).
3. Of course, there is no requirement that all controls be treated as uncertain (e.g., Leamer 2008). There are many cases when there is strong a priori theory that certain controls must be in the model. These kinds of strong assumptions are simple in practice to incorporate, as we will show in the final section.
4. The details of this algorithm and other formulae used in our mrobust.do Stata module are described in Online Appendix S1.
5. For example, under classical assumptions, the sampling standard error σ_S derives from a normal distribution of parameter estimates in repeated sampling. However, the underlying distribution that the modeling standard error σ_M derives from is unknown.
6. To obtain the total standard error, one does not add the sampling and modeling standard errors. Instead, one must compute the square root of the sum of the squared standard errors. With the bootstrapping option, the total standard error is simply the square root of the variance of all the b_{kj} estimates from all models applied to all bootstrap resamples. We find that these two procedures produce very similar estimates of the total standard error.
7. In Online Appendix S2, we also run all combinations of controls with logit and probit models as well as with both default and heteroscedastic-robust standard errors. However, we postpone the discussion of functional form robustness for later sections of this article.

8. Recall that the total standard error is the square root of the sum of the squares the sampling and modeling standard errors, so that $\sqrt{1.61^2 + 1.60^2} = 2.27$. The robustness ratio is then simply the ratio of the mean estimate over the total standard error, $2.29/2.27 = 1.01$. Alternatively, one could use the preferred estimate with the total standard error, yielding a robustness ratio of $3.7/2.27 = 1.63$, which still does not appear robust by the conventional critical values for a *t*-type statistic.
9. Equation (3) keeps the notation very simple, but one can think of the δ_{Zi} term in matrix form as $Z_k \delta_k^{'}$ where $\boldsymbol{Z}_k$ is a $k \times 1$ vector of control variables, and $\delta_k^{'}$ is a $1 \times k$ vector of coefficients.
10. In other words, δ may be small as long as it is not strictly zero.
11. These additional results are reported in the Stata do file. Thinking in terms of the omitted variable bias formula, marriage is negatively correlated with female (women applicants are less likely to be married) and positively correlated with mortgage acceptance (married people are more likely to be accepted), suggesting a classic suppressor relationship. Similarly, black applicants are more likely to be female (positive correlation) but less likely to be accepted (negative correlation). The omitted variable bias formula correctly predicts that including these variables makes the estimate for female larger (toward $+\infty$).
12. In future work, we aim to transform parameter estimates from many different functional forms into comparable marginal effects. It is possible to convert, say, logit coefficients into marginal effects that would be comparable with ordinary least squares results but this is not currently feasible in our computational robustness software.
13. These are supplementary results available in the Stata do file.

References

Andersen, Robert. 2008. *Modern Methods for Robust Regression*. Thousand Oaks, CA: Sage.

Angrist, Joshua and Alan Krueger. 2001. "Instrumental Variables and the Search for Identification: from Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15:69-85.

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

Berk, Richard, Lawrence Brown, and Linda Zhao. 2009. "Statistical Inference after Model Selection." *Journal of Quantitative Criminology* 26:217-36.

Brady, David, Jason Beckfield, and Martin Seeleib-Kaiser. 2005. "Economic Globalization and the Welfare State in Affluent Democracies, 1975-2001." *American Sociological Review* 70:921-48.

Brand, Jennie and Charles Halaby. 2006. "Regression and Matching Estimates of the Effects of Elite College Attendance on Educational and Career Achievement." *Social Science Research* 35:749-70.

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.

Ciccone, Antonio and Marek Jarociński. 2010. "Determinants of Economic Growth: Will Data Tell?" *American Economic Journal: Macroeconomics* 2:222-46.

Clarke, Kevin. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22:341-52.

Clarke, Kevin. 2009. "Return of the Phantom Menace: Omitted Variable Bias in Political Research." *Conflict Management and Peace Science* 26:46-66.

Clogg, Clifford, Eva Petkova, and Adamantios Haritou. 1995. "Statistical Methods for Comparing Regression Coefficients between Models." *American Journal of Sociology* 100:1261-93.

Conway, Karen, and Jonathan Rork. 2012. "No Country for Old Men (or Women)—Do State Tax Policies Drive Away the Elderly?" *National Tax Journal* 65:313-56.

Cook, Dennis. 1977. "Detection of Influential Observations in Linear Regression." *Technometrics* 19:15-18.

Durlauf, Steven, Fu Chao, and Salvador Navarro. 2012. "Assumptions Matter: Model Uncertainty and the Deterrent Effect of Capital Punishment." *American Economic Review* 102:487-92.

Durlauf, Steven, Fu Chao, and Salvador Navarro. 2013. "Capital Punishment and Deterrence: Understanding Disparate Results." *Journal of Quantitative Criminology* 29:103-21.

Durlauf, Steven, Paul Johnson, and Jonathan Temple. 2005. "Growth Econometrics." Pp. 555-678 in *Handbook of Economic Growth*, Vol. 1A, edited by Philippe Aghion and Steven Durlauf. Amsterdam, the Netherlands: Elsevier B.V.

Efron, Bradley. 1981. "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods." *Biometrika* 68:589-99.

Efron, Bradley. 2014. "Estimation and Accuracy after Model Selection." *Journal of the American Statistical Association* 109:991-1007.

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

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

Gross, Emily. 2003. *U.S. Population Migration Data: Strengths and Limitations*. Washington, DC: Internal Revenue Service Statistics of Income Division.

Hahn, Jinyong and Jerry Hausman. 2003. "Weak Instruments: Diagnosis and Cures in Empirical Econometrics." *American Economic Review* 93:118-25.

Heckman, James. 2005. "The Scientific Model of Causality." *Sociological Methodology* 35:1-97.

Herting, Jerald, David Grusky, and Stephen Van Rompaey. 1997. "The Social Geography of Interstate Mobility and Persistence." *American Sociological Review* 62: 267-87.

Hirsch, Barry. 2004. "Reconsidering Union Wage Effects: Surveying New Evidence on an Old Topic." *Journal of Labor Research* 25:233-66.

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.

Hoeting, Jennifer, David Madigan, Adrian Raftery, and Chris Volinsky. 1999. "Bayesian Model Averaging: A Tutorial." *Statistical Science* 14:382-417.

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

Kane, Jennifer, Philip Morgan, Kathleen Harris, and David Guilkey. 2013. "The Educational Consequences of Teen Childbearing: Comparisons across Ordinary Least Squares Regression, Propensity Score Matching, Parametric and Semi-parametric Maximum Likelihood Estimation." *Demography* 50: 2129-50.

King, Gary and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14:131-59.

Kleven, Henrik, Camille Landais, and Emmanuel Saez. 2013. "Taxation and International Migration of Superstars: Evidence from the European Football Market." *American Economic Review* 103:1892-924.

Koenker, Roger and Kevin F. Hallock. 2001. "Quantile Regression." *Journal of Economic Perspectives* 15:143-56.

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

Leamer, Edward E. 1985. "Sensitivity Analyses Would Help." *American Economic Review* 75:308-13.

Leamer, Edward E. 2008. "Extreme Bounds Analysis." In *The New Palgrave Dictionary of Economics*, 2nd ed., edited by Steven Durlauf and Lawrence Blume. Palgrave Macmillan. doi:10.1057/9780230226203.0539.

Leigh, Andrew. 2007. "How Closely Do Top Income Shares Track Other Measures of Inequality?" *Economic Journal* 117:F619-33.

Lochner, Kimberly, Ichiro Kawachi, and Bruce Kennedy. 1999. "Social Capital: A Guide to its Measurement." *Health and Place* 5: 259-70.

McGranahan, David A. 1999. "Natural Amenities Drive Rural Population Change." *Agricultural Economic Report* 781. Washington, DC: United States Department of Agriculture.

Montez, Jennifer, Robert Hummer, and Mark Hayward. 2012. "Education Attainment and Adult Mortality in the United States: A Systematic Analysis of Functional Form." *Demography* 49:315-36.

Munnell, Alicia, Geoffery Tootell, Lynne Browne, and James McEneany. 1996. "Mortgage Lending in Boston: Interpreting HMDA Data." *American Economic Review* 86:25-53.

Payton, Antony. 2009. "The Impact of Genetic Research on Our Understanding of Normal Cognitive Ageing: 1995 to 2009." *Neuropsychology Review* 19:451-77.

Pearl, Judea. 2011. "Understanding Bias Amplification." *American Journal of Epidemiology* 174:1223-27.

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. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111-65.

Rosenfeld, Jake. 2014. *What Unions No Longer Do*. Cambridge, MA: Harvard University Press.

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

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.

Santos Silva, J. and Silvana Tenreyro. 2006. "The Log of Gravity." *Review of Economics and Statistics* 88:641-58.

Simmons, Joseph, Leif Nelson, and Uri Simonsohn. 2011. "False-positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22:1359-66.

Stanley, T. D. and Hristos Doucouliagos. 2012. *Meta-regression Analysis in Economics and Business*. New York: Routledge.

van Raalte, Alyson and Hal Caswell. 2013. "Perturbation Analysis of Indices of Lifespan Variability." *Demography* 50:1615-40.

Western, Bruce. 1996. "Vague Theory and Model Uncertainty in Macrosociology." *Sociological Methodology* 26:165-92.

Wildeman, Christopher and Kristin Turney. 2014. "Positive, Negative, or Null? The Effects of Maternal Incarceration on Children's Behavioral Problems." *Demography* 51:1041-68.

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

Young, Cristobal and Charles Varner. 2011. "Millionaire Migration and State Taxation of Top Incomes: Evidence from a Natural Experiment." *National Tax Journal* 64:255-84.

Author Biographies

Cristobal Young is an assistant professor in the department of sociology at Stanford University. He works in the overlapping fields of economic sociology, stratification, and quantitative methodology studying the social processes and public policies that moderate income inequality, ranging from millionaire taxes to unemployment insurance. He has previously published on model uncertainty in the *American Sociological Review*.

Katherine Holsteen is a PhD student in epidemiology and clinical research at Stanford University. She previously completed a BS in mathematical and computational science from Stanford University in 2014 and worked in health policy research at Acumen LLC. She is interested in helping to advance causal inference methods, particularly in application to athletic injury prevention research.