Functional Form Robustness:

A Multiverse Approach[1]

Cristobal Young, Cornell University

Sheridan Stewart, Stanford University

May 23, 2020

Working draft: comments welcome. W = 9,948

Abstract

Social scientists face a dual problem of model uncertainty and methodological abundance. There are many different ways to conduct an analysis, but the true model is unknown. This 'uncertainty among abundance' offers spiraling opportunities to discover a statistically significant result. The problem is acute when models with significant results are published, while those with non-significant results go unmentioned. Multiverse analysis addresses this by recognizing 'many worlds' of modeling assumptions, using computational tools to show the full set of plausible estimates. We focus on functional form in common cases where data are dichotomous. Plausible methods include OLS, logit, probit, and various forms of matching. Do empirical results depend on which estimation command researchers choose? Our multiverse estimator takes all plausible combinations of control variables and functional forms, yielding a distribution of estimates that the data can support under alternative assumptions. Our empirical cases examine racial disparity in mortgage lending, the role of education in voting for Donald Trump, and the effect of unemployment on subjective wellbeing. Estimating over 4,300 unique model specifications, we find that OLS, logit, and probit are close substitutes, but matching is much more unstable. Multiverse analysis shows how a point estimate can be either strongly robust or a knife-edge result.

[1] All code and data for this article will be publicly available upon publication. Contact the authors at cristobal.young@cornell.edu or sheridan.a.stewart@stanford.edu. The authors regret that coronavirus prevented this paper from being presented at ASA 2020 and the ASA mid-year methodology meeting at the University of Arizona.

Introduction

Social scientists face a dual problem of model uncertainty and methodological abundance. There are many ways to conduct an analysis. When the “true” model is unknown, it is hard to say which imperfect approximation is best. Empirical estimates require exact modeling assumptions about what controls to include, how to clean, code, and categorize the variables, how to compute standard errors, and which functional form or estimation command to use. Within this modeling space, a “garden of forking paths” offers a multiplying array of chances to find statistically significant results (Gelman and Loken 2014; Borges 1962). Sociological theories are typically vague about how exactly they should be empirically tested (Western 1996). Alternative modeling assumptions can often be plausibly invoked, possibly producing different results. 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; Winship and Western 2016). Empirical estimates are a joint product of both the data and the model, and it is often unclear which is driving the results: the data or the modeling assumptions.

There is wide concern today that many statistically-significant results published in the social science literature are not robust and fail to replicate (Camerer et al 2018; Simmons, Nelson, and Simonsohn 2011; Young 2009). This in part because of a large transparency gap: authors can run endless alternative models to learn about possible results, but readers typically only see a handful of estimates curated for publication. In an age of computational power, we need better ways of revealing what estimates the data can support.

This study is part of a larger program to build a comprehensive multiverse analysis encompassing all the major analytical decisions that feed into an empirical estimate (Muñoz and Young 2018a; Young and Holsteen 2017; Steegan et al 2016; Brodeur et al 2020; Leamer 1983). The central goal is to improve the transparency of research. A multiverse analysis emphasizes that there are multiple universes of plausible, alternative modeling assumptions. One author may strongly defend the assumptions behind their estimates, but that same author might invoke different modeling assumptions under different circumstances – for example, if their role was to be a critic rather than the author. The difference between a single, preferred estimate and a multiverse analysis is the recognition that (many) other reasonable modeling assumptions could be made and defended. A multiverse analysis brings light to these possibilities; it allows one to backtrack along the garden of forking paths, showing the sensitivity of each modeling assumption to reasonable alternative specifications. The approach uses computational power to estimate hundreds or thousands of theoretically-reasonable models, estimating all unique combinations of specified model ingredients. The result is a modeling distribution of estimates, similar to a bootstrap sampling distribution – a bootstrap of the model assumptions.[2]

This paper advances methods for functional form robustness. We focus on functional forms often considered in the analysis of binary variables. Binary outcome variables are common in social science, but there is limited consensus on what functional form researchers should use: logit, probit, or OLS? Likewise, in matching models, where there is a binary treatment variable, researchers face a choice among many approaches to matching, for example, propensity-score and coarsened-exact matching (Morgan and Harding 2006; Iacus, King, and Porro 2012). All of these offer different ways of estimating a treatment effect by conditioning on observables. How much do the results depend on the choice among these possible estimators?

These examples are indicative of a broader challenge for applied research: *methodological abundance*. The variety and novelty of statistical techniques available today is remarkable. The *Handbook of Econometrics*, for example, runs across 77 chapters and more than 5,000 pages (Heckman and Leamer 2007). New methods aim to improve estimation; however, these methods come with unknown sources of error and bias; they expand the garden of forking paths and increase researcher degrees of freedom to discover false positive results (Glaeser 2008; Muñoz and Young 2018a). Functional form robustness shows how far methodological abundance, in practice, can shape the findings of research.

We demonstrate the multiverse methodology in a series of important sociological applications: racial disparity in mortgage lending, the role of education in voting for Donald Trump, and the effect of job loss on subjective wellbeing. These results suggest that some methodological concerns – such as OLS v. logit – are overheated, while matching estimators in particular should be treated with greater caution.

[2] Combined with an influence analysis, this can also show each model ingredient, on average, affects the coefficient of interest.

1. Into the Multiverse

The diversity of modeling strategies available today is on wide display in science. Two different analysts studying a topic scarcely ever use the same model specifications, as the "many analysts, one data set" project illustrates. That project recruited 61 researchers to study a data set on discrimination by skin tone in soccer (Silberzahn et al. 2018; c.f. Magnus and Morgan 1999). Despite a lively methodological discourse among the participating analysts, the project yielded wide-ranging empirical strategies, diverse results, and a roughly 70-30 split on the empirical conclusion. Moreover, the diversity of results was not readily explainable: neither the researchers' level of skill and training, nor their personal confidence in their own analysis, nor peer ratings of methodological quality explained why some estimates were larger or smaller, significant or non-significant. Equivalent levels of competence and quality generated diverging results. The group concluded that "significant variation in the results of analyses of complex data may be difficult to avoid, even by experts with honest intentions" (ibid: 337).

This gives intuition of the multiverse of methodological practices and assumptions. Authors make modeling choices that set them down different analytical paths. This opens up 'parallel worlds' that start from the same original data set, but increasingly diverge and come to depend more on the choices of the analyst than on the underlying data. While some 'adjacent' worlds may be very similar, the divergence tends to grow with the number of different decisions researchers could reasonably make.

In a multiverse, all possibilities are realized even if we cannot see them. The 'paths not taken' also come into being in parallel worlds. Writer Jorge Luis Borges, in a classic short story, imagined a novel in which all narrative junctures are written out, creating "diverse futures… which themselves also proliferate and fork" into an expanding "net of divergent, convergent, and parallel times" (Borges 1962:13; 16). When a character in that novel faces a crossroad with several alternatives, they choose all of them simultaneously. In cosmology, multiverse theorists argue that all mathematically-possible laws of physics probably exist in parallel universes beyond our current perception (Gribbin 2010). Every possible law of physics is an accident of history; all accidents of history eventually play out somewhere in the multiverse.

When applied to modeling strategies, this means recognizing that one good analysis is simply an example of a bifurcating world of many plausible analyses – all of which could in principle exist at once. In conventional analysis, when analysts face a set of alternatives, they choose one and eliminate the others. In a multiverse analysis, the goal is to realize all plausible modeling strategies. The realistic result of the modeling process is not a point estimate, but rather a distribution of estimates showing what possibilities the data can support, in different worlds of model assumptions.

The problem of the multiverse is not that it exists, but rather that it is not transparent. Authors can selectively choose a preferred result from among many plausible estimates. Scientists are susceptible to biases and conflicts of interest in a world of 'publish or perish.' Moreover, researchers often test theories against a zero-effect-null that they do not believe or see as credible, and are willing to change their modeling assumptions when the estimates seem 'wrong.' This presents the problem of motivated reasoning, highlighted in classic books like *The Art of Self-Persuasion* (Boudon 1994). Scholars can convince themselves that favorable model assumptions are clearly superior. Often, objectively incorrect beliefs are sincerely held; such views are wrong but feel convincing, and were generated through a process of painstaking, though motivated, reasoning: the authors 'convinced themselves.'

For these reasons, experimental science has long demanded double-blind procedures to minimize the chance that (un)conscious biases taint the data collection, analysis, and results. When conducting a trial of a new medical treatment, researchers are naturally hopeful that the treatment will work. These hopes can bias the study's results in multiple ways. Ideally, neither the researchers nor the participants know who is in the treatment or control groups until the study is complete (Shultz and Grimes 2002). Blinding procedures have allowed scholars to admit having non-objective personal views, while still producing objective work: the methods of science insulate the evidence from the personal views of the scientist. When blinding is not viable or not practiced, the alternative is transparency. If procedures cannot assure objectivity about the analysis, then we need transparency about what analyses *could have been* selected: we need to see into the multiverse.

1.1 The Multiverse of Controls

We approach statistical analysis as a framework in which a researcher is concerned with identifying the effect of a treatment on some outcome. The goal is not prediction (i.e., improving "model fit"), but rather minimizing bias in the parameter estimate, producing a more accurate estimate of the relationship between an explanatory variable and the outcome (Muñoz and

Young 2018b). In other words, we focus on $\hat{b}$ questions, rather than $\hat{y}$ questions (Molina and Garip 2019).

A basic multiverse analysis considers the set of possible control variables.

(1) $Y = f(x, Z, Q)$

In eq. 1, Y is the outcome, x is the variable of interest, Z is a vector of control variables deemed essential, and Q is a vector of *n* plausible controls $[q_1, q_2, \ldots\ q_n]$, each of which may or may not be included in the model. The vector Z allows authors to impose dogmatic assumptions about necessary controls to ensure the model space is credible. Q represents model uncertainty: different analysts may make different assumptions about which elements belong in the model.

The purpose of robustness analysis is to understand how the results change when including any element of Q in the analysis. It is important to note that in equation 1, *f* is a fixed choice of functional form made by the analyst. To make this concrete, consider a simple linear model with a variable of interest (x), one necessary control (z), and two plausible control variables, q_1, q_2. This simple multiverse is represented in the following set of four possible models:

(2) $y_i = \beta_1 x_i + \beta_2 z_i + \varepsilon_i$
(3) $y_i = \beta_1 x_i + \beta_2 z_i + \beta_3 q_{1i} + \varepsilon_i$
(4) $y_i = \beta_1 x_i + \beta_2 z_i + \beta_4 q_{2i} + \varepsilon_i$
(5) $y_i = \beta_1 x_i + \beta_2 z_i + \beta_3 q_{1i} + \beta_4 q_{2i} + \varepsilon_i$

A model could include q_1, or q_2, both, or neither. These four equations represent different reasonable ways of specifying the model given the uncertainty, and offer four plausible estimates of β_1. As the number of plausible model ingredients increases, the model space increases exponentially: with *n* plausible control variables, there are 2^n unique combinations of those variables. With two cases of uncertainty (regarding two plausible controls) in the example above, there are $2^2 = 4$ unique models. With 10 possible controls, there are $2^{10} = 1{,}024$ unique models, and with 20 possible controls there are over one million unique models.

A multiverse analysis can also be applied to different ways of cleaning, coding, and categorizing the data (Leahey et al. 2003). In one study, researchers revisited the coding decisions of a publication, using other coding rules that the original authors had themselves used in different articles (Steegan et al. 2016). For example, the semi-continuous variable called "fertility" was split into categories (bins) in different ways in each of five different studies – showing five evidently-valid coding strategies for that variable alone. The original results were not robust to the alternative data coding decisions. Though not the focus of this study, these elements of the analysis are no less important than variable selection itself, and alternate versions of variables based on different coding rules can be readily incorporated into a computational analysis (Young and Holsteen 2017).

A general rule of thumb for multiverse analysis is that if, in one world, an author would argue for one analytical decision, there is an alternate world where that same author could argue against it. Similarly, as Winship and Western note, "even if we are individually confident about a particular specification, there are almost certainly others who will argue that we have gotten it wrong" (2016:646). While we allow that there exist some Z controls which are beyond dispute and always belong in the model, we think most controls come with non-trivial degrees of uncertainty and deserve examination in a multiverse analysis.

1.2 The Multiverse of Functional Forms

We now extend model uncertainty to the area of functional form. A simple way to think about many possible functional forms is with the generalized linear model, which provides an umbrella concept that treats the link function as a variable (McCullagh and Nelder 1989). This specifies a family of models, allowing for non-linear transformations of a linear combination of explanatory variables. The link function could be OLS, logit, probit, ordered logit or probit, multinomial logit, Poisson, negative binomial, and more. From a statistical programming perspective, link functions can be thought of as estimation commands.

In more formal terms, the choice of functional form generalizes equation 1 as,

$$(6) \quad Y = f_m(x, Z, Q)$$

The shift we propose here is to make the function *f* a variable with subscript *m*, indicating a vector of models $[f_1, f_2, \dots, f_M]$. For example, the link function *f* is OLS if *m* =1, logit if *m* = 2, propensity score matching if *m* = 3, and so on. Our goal is not to estimate which link function is ‘best,’ but rather to allow a range of possibilities – each paired to different views about which functional form would provide the least biased estimate of β_1. We also require that if any link function changes the scale of the β_1 coefficients (relative to OLS), those estimates must be converted back into marginal effects (Mize, Doan and Long 2019).

With *n* uncertain control variables and *m* plausible link functions, there are $2^n \times m = J$ unique models. Retrieving the estimated effect of x, $\hat{\beta}_1$, from each of these models provides a modeling distribution that is directly analogous to the sampling distribution. The mean of the modeling distribution is $\overline{\hat{\beta}_1} = \frac{1}{J}\sum_{j=1}^{J}\hat{\beta}_{1j}$, and the modeling variance is $V = \frac{1}{J}\sum_{j=1}^{J}(\hat{\beta}_{1j} - \overline{\hat{\beta}_1})$. The square root of V is the modeling standard error. This is computed the same way as sampling standard errors in a bootstrap process, and can be thought of as a standard error that comes from bootstrapping the model, rather than the sample (Efron and Tibshirani 1993). As with sampling standard errors, large observed modeling standard errors cast doubt on the reliability of a point estimate. To the extent that the specified multiverse captures rival views of how best to conduct a given applied analysis, the resulting modeling distribution will show the range of estimates that may be found in a skeptical replication or an “adversarial collaboration” with other researchers (Bateman et al. 2005).

Alternative functional forms can affect the modeling distribution in two different ways, changing either the mean or the variance of the distribution. Statisticians evaluate estimators by focusing on the mean and the variance of their *sampling* distributions - behavior in repeated sampling. We are likewise interested in the mean and the variance of the *modeling* distributions – behavior in repeated modeling. Different functional forms could shift the *mean* of the distribution, as if by adding or reducing systematic bias. In this case, for any given set of plausible controls, one functional form tends to have larger estimates than the other. This systematic influence would invite deliberation about which is the best functional form assumption, even if that is hard to fully resolve (Bettey et al. 2019). Secondly, alternative functional forms could change the *variance* of the modeling distribution – having larger or smaller modeling standard errors. In this case, the influence of functional form is seen mostly in the tails of the distribution: one functional form gives more extreme estimates – potentially both larger *and* smaller – than the other. Such an idiosyncratic influence is more perplexing, seeming to be a purely negative property of a functional form, contributing to the model dependence of findings (King and Nielsen 2019). In any event, the mean and the variance of modeling distributions are key features for understanding functional form robustness.

2. Binary Outcomes and Treatments: Logit, Matching, and OLS

Modeling choices are often interconnected with the structure of the data. In the simplest empirical analysis, outcome and treatment variables – or, the left-hand and the right-hand side – are both continuous. There is no special statistical model for such cases; the common default is to use OLS. However, when the outcome variable is binary, sociologists commonly adopt a logit model. Other fields, such as economics, traditionally favored probit in such cases (Angrist and Pischke 2009:102-7). Regardless, binary outcomes are often thought to indicate a non-linear functional form, for a variety of reasons (Long 1997). Similarly, when the treatment variable is binary, matching models have emerged as a prominent method for estimating the effect of a binary variable (Iacus, King, and Porro 2012; Morgan and Harding 2006).

The variety of functional forms available for binary outcomes and treatments may be desirable in many ways. A downside, however, is the expanding degrees of freedom it offers researchers, in which a multiplicity of plausible methods provides many additional chances to find and selectively report a statistically significant result. Each functional form, potentially in combination with each control, offers a new opportunity to leverage chance associations in the data.

2.1 Binary Outcomes: Logit, Probit, or OLS?

Different estimators exist for sensible reasons. OLS applied to binary outcomes has well-known problems (Long 1997; Battey et al. 2019). OLS can produce nonsensical predictions for individual observations (i.e., predicted probabilities that are less than zero or greater than one). OLS has a linear functional form, assuming that changes in a treatment variable are monotonic with changes in the probability of the outcome.[3] And applying OLS to a binary outcome increases the risk of heteroscedasticity (Hellevik 2009).

[3] For example, when studying the effect of having children on labor force participation, OLS would assume by definition that the effect of having four children is exactly four times the effect of having one child (Long 1997). Of

Logit and probit models are specifically designed for modeling binary outcomes. In contrast to the OLS equation 2 above, logit and probit are written as

$$(7)\quad \text{Logit:}\qquad \Pr(Y = 1 | \, x, z) = \, (1 + e^{-(\beta_1 x_i \, + \, \beta_2 z_i)})^{-1}$$

$$(8)\quad \text{Probit:}\qquad \Pr(Y = 1 | \, x, z) = \, \Phi(\beta_1 x_i \, + \, \beta_2 z_i)$$

where Φ is the cumulative standard normal distribution function. These models are similar. Both restrict the predicted probabilities to between 0 and 1, and feature non-linear effects of x on y. However, there are also serious shortcomings of these nonlinear models, including problems that can prevent convergence, and challenges in interpreting results across nested models (King and Zeng 2001; Mood 2010).[4] Moreover, sociologists are generally interested in the parameter estimates (i.e., regression coefficients) rather than predicted probabilities for individual observations. In light of various problems with nonlinear models for binary outcomes, Breen, Karlson, and Holm (2018) make several recommendations that boil down to using OLS as the default model.

To understand current standards of research practice, we examined all articles published in the *American Sociological Review* and the *American Journal of Sociology* from fall, 2016, to summer, 2018.[5] Over this time, 103 of the 135 published articles were primarily quantitative (76 percent). Binary outcomes were common, occurring in 39 percent of the quantitative articles. Logistic regression is the favored functional form assumption, as most articles with binary outcomes used logit (58 percent). However, 28 percent of articles with binary outcomes used OLS. There is not a definitive standard of practice for modeling binary outcomes in sociology. Choice between OLS and logit appears to offer a salient degree of freedom to researchers in the analytical process. Is this simply a stylistic preference for linear versus non-linear models? Or would the findings commonly change if different functional form assumptions were applied?

course, analysts can introduce nonlinearities into OLS to address this concern, but logit models are non-linear by default.

[4] Contributing factors include small sample sizes, 'wide' data sets (i.e., a high explanatory variable to observation ratio), flat gradients, multiple local maxima, and collinear explanatory variables (Long, 1997).

[5] For AJS: issues September, 2016, to July, 2018. For ASR: issues August, 2016, to June, 2018.

Table 1. Quantitative Analyses in
ASR and *AJS*, 2016 - 2018

	Quantitative	Binary outcome	Used logit model	Used OLS
N	103	40	23	11
Percent of articles	76%	-	-	-
Percent of quantitative articles	-	39%	22%	11%
Percent of binary outcomes	-	-	58%	28%

Note: Authors' analysis of published articles. For AJS: issues September, 2016, to July, 2018. For ASR: issues August, 2016, to June, 2018.

2.1.1 Scale of Coefficients

A central problem in functional form robustness is that different functional forms often report estimates on different scales (Mize, Doan, and Long 2019). OLS models report estimates on the probability scale as marginal effects, logit models give estimates as log-odds (alternatively, odds-ratios), while probit estimates are reported on the z-score scale. Because changing functional forms simultaneously changes the scale of the coefficients, it is not possible to directly compare logit or probit results to OLS estimates. We use the average marginal effects procedure to make results from different functional forms comparable (Williams 2011).

Table 2 illustrates the problem of scale across functional forms. Consider an analysis of the voting for Donald Trump in the 2016 presidential election as a function of race (white versus non-white). The logit coefficient for white is 3.3, meaning that whites have more than triple the odds of voting Trump compared to non-whites. The probit coefficient is 0.64, which is hard to interpret but looks very different from 3.3. The next row shows the OLS estimate, 0.10, meaning that whites are 10 percentage points more likely to vote Trump than nonwhites. Are all these estimates saying the same thing, or are they giving different results?

Table 2. Comparison of Coefficient Magnitudes

	White (vs. nonwhite)
Logit (odds-ratios)	3.30
Probit	0.64
OLS	0.10
Logit (AME)	0.11
Probit (AME)	0.11

Note: Models control for income, gender, age, age squared, and marital status. Data: ANES 2016 Time Series. N = 1,701.

In the final rows of Table 2, we convert the logit odds ratios and probit coefficients onto the probability scale as average marginal effects.[6] Once these estimates are on the same scale as OLS, we see that the logit and probit results are much the same as the OLS estimates. When coefficients are placed on the same scale, it may well be that different functional forms produce equivalent results – a fact that is not transparent before converting coefficients to marginal effects. In practice, how often do logit, probit, and OLS models give the same underlying results? If logit coefficients were routinely converted to marginal effects and compared to OLS, how often would the differences be large enough to matter?

2.2 Binary Treatments: Matching versus OLS

Matching is a common method for effects estimation when the treatment variable of interest is binary. Think of the variable x as indicating groups of treatment ($x = 1$) and control ($x = 0$) cases. If assignment to 0 or 1 is random, then the causal effect of the treatment is estimated simply by the mean difference between the groups on the outcome variable, i.e. E[Y | x = 1] – E[Y | x = 0]. When assignment is non-random, matching aims for balance in covariates between the two groups. The goal is to match each treatment case with at least one control case that has (very) similar observable characteristics. For example, if one group experiences job loss while another other does not, matching aims produce balanced ‘treatment’ and ‘control’ groups that have the same gender and age make up, education levels, occupational history, and the like.

[6] The average marginal effect is a post-estimation procedure that computes the expected difference in outcome probability due to a unit increase in the treatment variable. In this case, the AME is calculated from the logit / probit regression results by predicting the outcome probability for each observation ($\widehat{Vote\_Trump_i}$), treating all cases as if they were white respondents. Then predict the outcome probabilities as if each respondent were non-white. The difference between these two probabilities is the marginal effect *for each case*. Averaging the difference across all cases gives the average marginal effect, which can then be compared to the OLS coefficient (Williams 2011).

Early studies often viewed matching as a method that directly provides causal estimates (reviewed in Arceneaux et al. 2010). It is now recognized that matching relies on the same unconfoundedness assumption as OLS: matching offers no solution to problems of endogeneity or omitted variable bias (Morgan and Harding 2006; Imbens 2015; King and Nielson 2019). Unobserved or endogenous covariates bias matching estimators just as they do for OLS.

Still, matching is often invoked for other properties that may offer superior estimation over OLS. Matching is non-parametric, or perhaps quasi-non-parametric. Matching does not impose the strict linearity assumption of OLS, and offers ways to restrict the analysis to the basis of common support, and exclude or down-weight observations that are poorly matched – thus potentially improving covariate balance between comparison groups. These are welcome features. However, a downside of matching is the abundance of proposals of how to implement it in practice. The *idea* of matching has been much more popular than any single approach of how to do it. While the method of least squares is calculated today more or less exactly as it was laid out some 200 years ago, authors using matching have wide discretion in choosing a preferred version. Indeed, there remains “very little specific guidance in the literature on which of these matching algorithms works best” (Morgan and Harding 2006:34; see also Angrist and Pischke 2009:86).

In general, matching is a two-stage process. First, one estimates the probability that a unit is in the treatment group, based on the observed covariate z and selected elements of Q_n. In propensity score matching (PSM), the propensity score is

$$(9) \quad \pi_i = \Pr(X_i = 1 |\, z, Q_n)$$

π_i is often estimated using predictions from a logit model. In the second stage, the estimated propensity score $\widehat{\pi_i}$ is used to analyze the outcome equation, which could be simply the mean difference

$$(10) \quad \Pr(Y = 1 |\, x, z, Q_n) = E(Y|\widehat{\pi}_i = 1) - E(Y|\widehat{\pi}_i = 0)$$

In coarsened exact matching (CEM), rather than using the first-stage regression of equation (9), an algorithm temporarily coarsens the continuous z and $Q_n$variables into bins that allow treated cases to be matched to similar control cases. Cases with the same values for all coarsened variables are grouped together into a stratum (e.g., 'high,' 'medium,' and 'low' education). The goal is to coarsen z and Q_n just enough to match treatment ($x = 1$) and control ($x = 0$) observations into comparable strata. The CEM estimate comes from comparing treatment and control cases within each stratum. For further details of CEM see Iacus, King, and Porro (2012). As King and colleagues write, "PSM and CEM each represent the most common member of one of the two known classes of matching methods" (p 1). Moreover, the contrast between these two is particularly interesting given recent work asserting the superiority of CEM and that "propensity scores should not be used for matching" (King and Nielson 2019).

In some ways, matching methods are emblematic of the challenge of novelty in statistical methods. While new techniques can offer improved estimation or causal inference, they come with new uncertainties and poorly understood biases. New techniques often generate considerable enthusiasm, but should also come with large doses of skepticism because of the researcher degrees of freedom that novelty brings (Glaeser 2008). Some prominent scholars have argued that matching reduces model dependence in empirical results – in other words, matching offers more reliable effect estimates than regression (Ho, Imai, King, and Stuart 2007; King and Nielson 2019; Dehejia and Wahba 1999). Other applied researchers have sometimes noted – perhaps with disappointment – that "regression and matching yield rather similar patterns of results" (Brand and Halaby 2006:767). Other work, moreover, cautions that matching can produce biased or even nonsensical results in applied settings (Arceneaux et al. 2010). This well captures a wide range of views about matching: the method might improve the quality and consistency of estimation over OLS, might worsen it, or might simply give the same results as OLS. Our applications are not meant to definitively settle this matter, but rather provide a framework to acknowledge and report on it: a multiverse analysis that incorporates both matching and OLS.

3. Applications

We now apply the functional form multiverse analysis in a series of applications. These show (1) how comparisons that can be made across functional forms, (2) demonstrate the often-remarkable similarity of estimates from different functional forms when the coefficients are on the same scale, and (3) illustrate how to make use of those comparisons when interpreting results and presenting findings. We start by contrasting linear and non-linear models and then extend to matching estimators.

Application I: Discrimination in Mortgage Lending

Are commercial banks less likely to accept mortgage applications from African Americans? We examine data on discrimination in mortgage lending originally collected by the Boston Federal Reserve (Munnell et al., 1996). The outcome is a binary indicator of mortgage acceptance; the explanatory variable of interest is a binary indicator for race (black = 1, white = 0). We also consider eight plausibly important controls: gender, marital status, self-employment, and a set of variables especially important to mortgage lenders: the applicant's credit history, their housing expense ratio, payment-to-income ratio, loan-to-value ratio, and whether the applicant was denied private mortgage insurance. Table 3 shows the baseline results with all controls included using OLS, logit, and probit.

The OLS estimate is -0.114, meaning that blacks are 11.4 percentage points less likely to have their mortgage application accepted (from an average acceptance of 88%). With a standard error of 0.018, the estimate is highly significant, with a t statistic greater than 6. We treat this as the 'preferred estimate' and as the reference point for robustness testing. The logit odds-ratio estimate is 0.373, but when converted to a marginal effect (model 3) the estimate is that blacks have a 9.7 percentage-point lower chance of having their mortgage application accepted. With probit, the marginal effect is a 10.0 percentage-point lower probability. While the logit and probit estimates are slightly smaller than the OLS estimate, all three show a clear pattern consistent with racial discrimination. Without converting the logit and probit coefficients to marginal effects, all we could say is that all the estimates are statistically significant – a weakly informative comparison.

Table 3. Effect of Race on Mortgage Acceptance

	Model 1: OLS	Model 2: Logit (odds-ratio)	Model 3: Logit (AME)	Model 4: Probit	Model 5: Probit (AME)
Black	-0.114***	0.373***	-0.097***	-0.534***	-0.100***
	(0.018)	(0.065)	(0.021)	(0.096)	(0.021)
Woman	0.037*	1.614*	0.035*	0.246*	0.035*
	(0.016)	(0.328)	(0.013)	(0.105)	(0.014)
Married	0.046***	1.719**	0.044**	0.287**	0.045**
	(0.013)	(0.274)	(0.013)	(0.083)	(0.013)
Self-employed	-0.056**	0.560**	-0.052*	-0.315**	-0.054*
	(0.018)	(0.116)	(0.021)	(0.110)	(0.021)
Bad credit history	-0.252***	0.187***	-0.208***	-0.951***	-0.220***
	(0.023)	(0.036)	(0.033)	(0.113)	(0.034)
Housing expense ratio	0.058	1.480	0.031	0.272	0.041
	(0.105)	(1.781)	(0.095)	(0.643)	(0.098)
Payment-income ratio	-0.500***	0.004***	-0.425***	-2.720***	-0.413***
	(0.093)	(0.005)	(0.083)	(0.553)	(0.084)
Loan-to-value ratio	-0.119**	0.125***	-0.164***	-0.966***	-0.147***
	(0.034)	(0.061)	(0.039)	(0.240)	(0.037)
Denied mortgage insurance	-0.712***	0.013***	-0.722***	-2.478***	-0.724***
	(0.042)	(0.007)	(0.071)	(0.282)	(0.066)
Constant	1.138***	280.188***	-	2.910***	-
	(0.034)	(143.800)	-	(0.249)	-
N	2,355	2,355	2,355	2,355	2,355
BIC	820.757	1,390.939	1,390.939	1,394.297	1,394.297

Note: Data: Federal Reserve Bank of Boston.

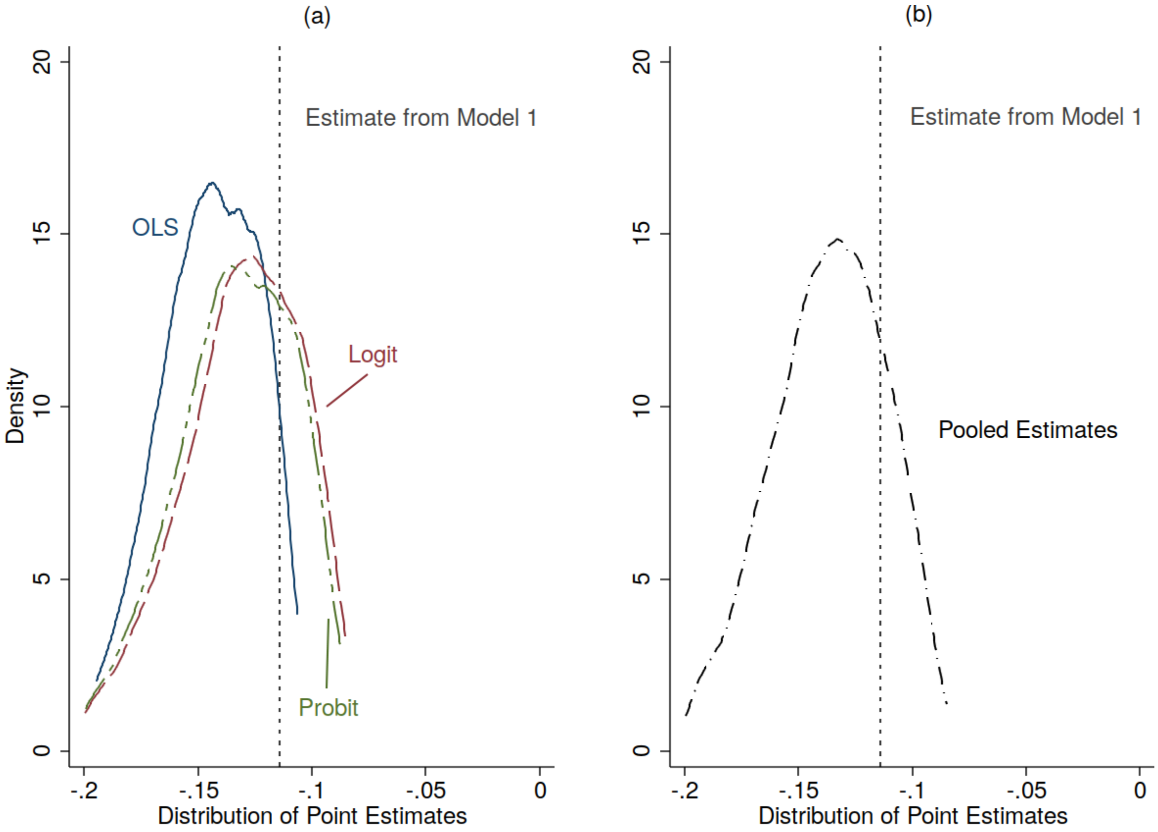
Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

To examine the robustness of the results to the choice of controls, we consider all combinations of the eight control variables, for a model space of $2^8 = 256$ possible specifications. To test functional form robustness, we run all 256 specifications separately using OLS, logit, and probit, converting the logit and probit estimates to marginal effects – for a combined model space of 768 unique specifications. In Figure 1, panel (a) depicts the distribution of estimates across the modeling space showing OLS, logit, and probit separately; panel (b) shows the pooled modeling distribution. Several conclusions emerge.

First, all three functional forms show a range of estimates depending on what controls are included in the model. Second, the estimates reported in Table 3 are among the most 'conservative' (i.e., closest to zero): excluding any control variable mostly makes the estimates larger in magnitude. The largest estimate of racial disparity the data can support in this model space is -20 percentage points, and the smallest is -9 percentage points. Third, the logit and probit estimates are virtually indistinguishable, even on simple a numeric basis. Fourth, the difference between the linear and non-linear models is systematic: the means of the non-linear distributions are closer to zero than for OLS, as is virtually every single point estimate in those distributions – though the difference is small. Fifth, the conclusion from the baseline regression result is highly robust: the large racial gap in mortgage acceptance is visible regardless of which control variables or functional form an analyst adopts. There is legitimate methodological debate to be had about the exact size of the racial disparity, but not about its existence.

Figure 1. Modeling Distributions: Effect of Race on Mortgage Acceptance



Note: Estimates from 768 model specifications. Data: Federal Reserve Bank of Boston. N = 2,355.

Table 4 provides summary statistics of the modeling distribution and helps to quantify these conclusions. The average OLS estimate is -14.4pp, while the average logit estimate is -13.1pp. This is a modest but systematic difference of 1.3pp (or about 10 percent of the mean OLS estimate). Probit, with an average estimate of -13.4pp, differs from OLS by only 1pp and from logit by only 0.3pp. For each functional form, the modeling standard error is larger than the sampling standard error. This means there is greater variance in the results through repeated modeling than from repeated sampling. Re-iterating the robustness of the core conclusion, the estimate for race is statistically significant in 100 percent of model specifications across functional forms.

Table 4. Functional Form Robustness of: Effect of Race on Mortgage Acceptance

	Average estimate	*Range*	*Average sample SE*	*Modeling SE*	*Significance rate ($p < 0.05$)*
OLS	-0.144	[-0.195, -0.106]	0.018	0.021	100.0%
Logit (AME)	-0.131	[-0.200, -0.085]	0.023	0.026	100.0%
Probit (AME)	-0.134	[-0.200, -0.087]	0.023	0.025	100.0%
Pooled distribution	-0.136	[-0.200, -0.085]	0.021	0.025	100.0%

Note: Estimates from 768 model specifications. Data: Federal Reserve Bank of Boston. N = 2,355.

Does the choice of functional form matter? In this case, hardly at all. The choice between logit and probit appears entirely stylistic. Between a linear and a non-linear model, the difference is greater, but selecting between them could at best 'improve' or weaken a marginally significant estimate. This would only matter when the effect size in question is small.

Application II: Voting for Trump in the 2016 Election

Who voted for candidate Donald Trump in the 2016 presidential election? Historically, college graduates have tended to vote republican while working-class voters leaned democrat (Pew 2018). Intuition tells us that 2016 did not follow that pattern. We draw on the American National Election Study to analyze the effect of having a college degree on voting for Trump.

Both the outcome and the treatment variables of interest are binary. In this application, we compare OLS, logit, propensity-score matching (PSM), and coarsened exact matching (CEM). In the interests of space, we set aside probit and focus on logit to represent the nonlinear model. We consider seven plausible control variables: race, gender, age, square of age, marital status, party affiliation, and income. Taking all possible combinations of these controls gives a modeling space of $2^7 = 128$ unique specifications. This set is then estimated with four different functional forms: OLS, logit, and the matching estimators (PSM and CEM). This yields 510 unique estimates of the effect of a college degree on voting for Trump in 2016.[7]

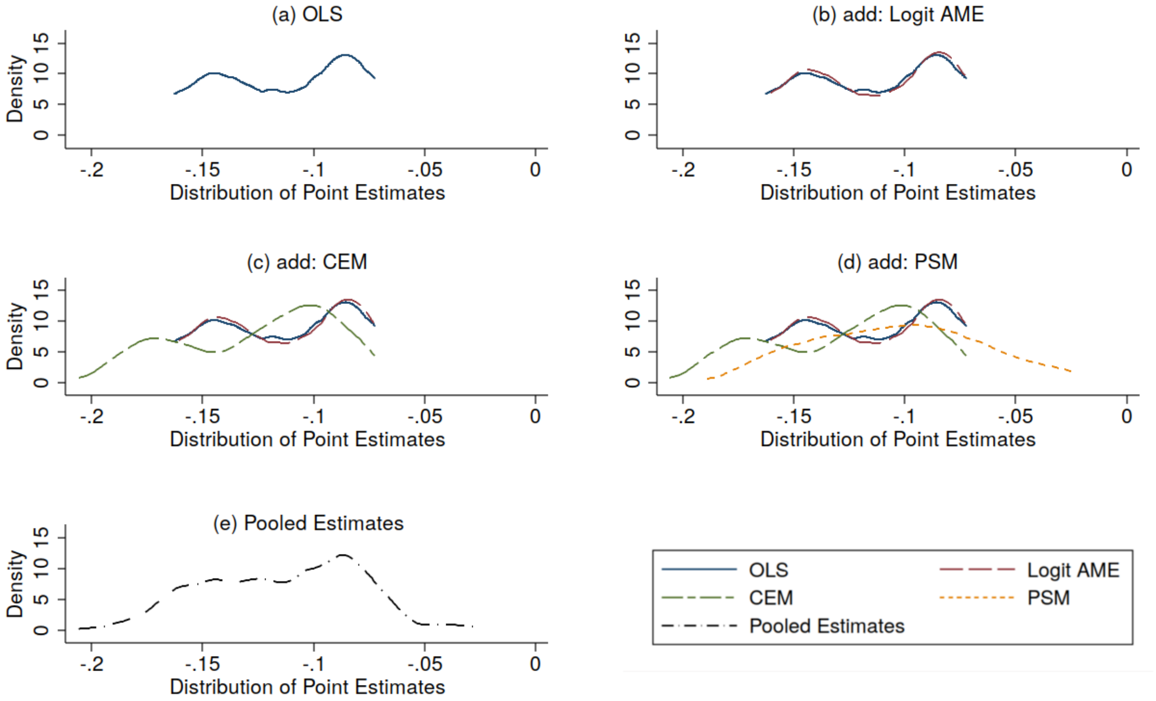
Figure 2 plots the results. The modeling distributions from all four functional forms clearly overlap. Looking at the plots of the modeling distributions, we can begin to draw inferences about the behavior of the models. In cell a, we report the modeling distribution from OLS. Depending on the set of controls used, one can find estimates ranging from 7 to 16

[7] Note that the matching estimators – CEM and PSM – do not estimate the bivariate model with no controls, as this provides nothing to match on. Thus, the matching models each produce 127, rather than 128, estimates.

percentage points lower probability of voting for Trump among college graduates. The distribution is bimodal, indicating that one of the control variables has strong influence on the results. In Appendix I, the influence analysis shows this variable is party affiliation: including this variable shrinks the estimate for education by about 50 percent. Party affiliation may well be an endogenous control/collider variable, something more like an intermediate outcome than an exogenous factor to be controlled. Our point is that informed consumers of this research should be aware that the decision to control for party affiliation is highly influential for the results. Race and income also have some modest influence, while gender, age, and marital status are irrelevant to the results. While gender and age are often important factors in support for Trump, they do not matter for the current analysis (the effect of education in voting for Trump).

In cell b, we add the modeling distribution from logit, reporting average marginal effects. These two functional forms offer identical modeling distributions: nothing is at stake in the analytic choice between OLS and logit in this case. In cell c, we add the distribution from coarsened exact matching (CEM). While this distribution largely overlaps with those of OLS and logit, it is wider, with some estimates that are larger in magnitude (longer left tail). In cell d, we add the modeling distribution from propensity score matching, which further expands the distribution with some estimates that are closer to zero (longer right tail). Finally, in cell e we show the pooled distribution of estimates: results from all possible combinations of seven plausible control variables and the four different functional forms. Matching produces both larger and smaller estimates, depending on what set of controls are included. In other words, the matching estimators have idiosyncratic influence on the modeling distribution.

Figure 2. Modeling Distributions: Effect of College Degree on Voting for Donald Trump in 2016



Note: Estimates from 510 model specifications. Data: ANES 2016 Time Series. N = 1,701.

The view provided in Figure 2 is reflected in the summary statistics of Table 5. Logit and OLS have nearly identical average point estimates, average sample standard errors, and modeling standard errors. Matching methods produce similar estimates as well, though the differences are noticeable. PSM gives the smallest average estimate (-0.104), and in nearly nine percent of models the estimate is not statistically significant. CEM gives the largest average estimate (-0.126), and all its estimates are significant. Both matching estimators have larger ranges and modeling standard errors than OLS or logit, especially PSM. These longer-tail modeling distributions are worrisome because it means researchers have greater flexibility in how large or small a point estimate they can choose to report, and ultimately what empirical conclusions they draw. Some PSM models allow a null conclusion, but nearly 98 percent of this multiverse finds that a college degree reduces the chance of voting for Trump, by about 11 percentage points on average.

Table 5. Functional Form Robustness of: Effect of College Degree on Voting for Donald Trump in 2016

	Average estimate	*Range*	*Average sample SE*	*Modeling SE*	*Significance rate (p < 0.05)*
OLS	-0.114	[-0.163, -0.072]	0.019	0.031	100.0%
Logit (AME)	-0.113	[-0.161, -0.073]	0.019	0.030	100.0%
PSM	-0.104	[-0.189, -0.025]	0.024	0.037	91.3%
CEM	-0.126	[-0.206, -0.072]	0.021	0.034	100.0%
Pooled distribution	-0.114	[-0.206, -0.025]	0.021	0.034	97.8%

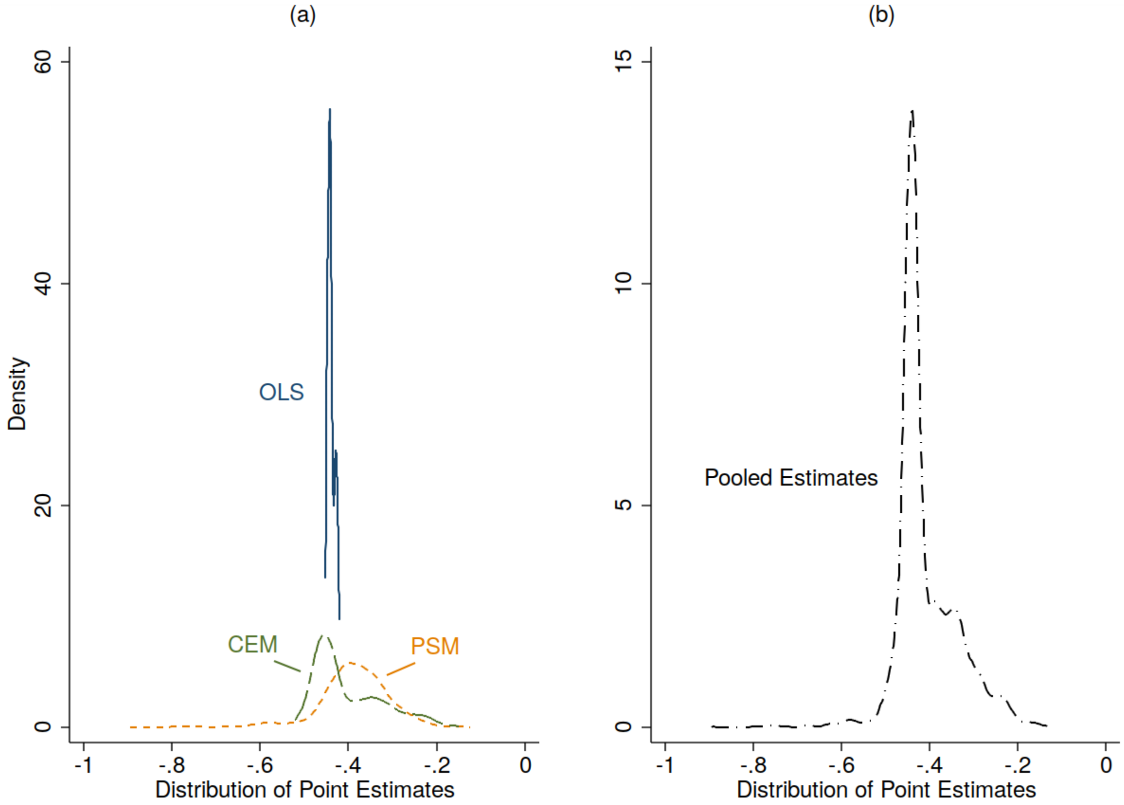
Note: Estimates from 510 model specifications. Data: ANES 2016 Time Series. N = 1,701.

Application III: Unemployment and Wellbeing

Finally, we draw on panel data examining the effect of job loss on subjective wellbeing. We compare OLS, propensity score matching, and coarsened exact matching estimates of the effect of losing a job in a two-wave panel. We do not include logit because the outcome (change in subjective well-being) is not binary, but we include matching estimators since the variable of interest (job loss) *is* binary. We use fixed-effects analysis examining the transition from working to unemployed (coded as one; zero for those not experiencing this transition) (Young 2012: 622, Table 3).[8] We focus on a set of 10 plausible time-varying controls, including working part-time, having children in the home, marital status, homeowner status, income, having zero/negative wealth, not having one's desired food, not having enough food, being eligible for unemployment insurance, and receiving food stamps. While each of these controls is plausibly related to changes in wellbeing (and employment status), some of them could be post-treatment or endogenous controls that could potentially bias the results. Consequently, it seems properly cautious to allow dropping any one of these controls, as might occur in a skeptical replication. All combinations of these 10 give a covariate multiverse of $2^{10} = 1{,}024$ unique specifications. Next, we extend the multiverse to incorporate three different functional forms: OLS, PSM, and CEM. This broader multiverse yields a modeling distribution of 3,070 point estimates. We standardize the outcome, changes in subjective well-being, for ease of interpretation. In our baseline OLS regression model, which includes all 10 controls, we find that job loss leads to a 0.4 standard deviation drop in wellbeing (table not shown).

[8] Specifically, the data are first-differenced so that all functional forms are analyzing the change in well-being between waves 2003-05 as a result of transitioning from employed in the 2003 wave to unemployed in the 2005 wave. See Young (2012) for a more complete discussion.

Figure 3. Modeling Distributions: Effect of Job Loss on Subjective Wellbeing



Note: Estimates from 3,070 model specifications. Data: PSID (2003-2005 waves). N = 6,192.

Panel (a) of Figure 3 shows the modeling distributions from OLS, PSM and CEM separately. The modeling distribution from OLS is highly concentrated – appearing largely as a spike in estimates near the baseline of -0.4. The two matching estimators, however, show a dramatically wider range of estimates, albeit with modes close to that of the OLS distribution. Table 6 provides more concrete detail: the average estimates across functional forms are all quite similar for OLS (-0.44), PSM (-0.39) and CEM (-.040), revealing little systematic difference. However, the ranges are far larger for matching than for OLS. The modeling standard errors are many times greater using PSM (0.09) and CEM (0.08) than for OLS (0.01). This reflects idiosyncratic differences across functional forms. These non-systematic differences mean that the results are more open to debate among scholars on any loose methodological grounds. It also means that, in the absence of a multiverse analysis, researchers will have more undisclosed leeway to choose a preferred estimate.

Table 6. Functional Form Robustness of: Effect of Job Loss on Wellbeing

	Average estimate	*Range*	*Average sample SE*	*Modeling SE*	*Significance rate (p < 0.05)*
OLS	-0.439	[-0.452, -0.421]	0.082	0.008	100.0%
PSM	-0.391	[-0.896, -0.124]	0.134	0.093	86.9%
CEM	-0.401	[-0.521, -0.149]	0.092	0.079	98.5%
Pooled distribution	-0.410	[-0.896, -0.124]	0.103	0.074	95.2%

Note: Estimates from 3,070 model specifications. Data: PSID (2003-2005 waves). N = 6,192.

Nevertheless, as panel (b) shows, the core empirical finding of psychological harm from unemployment is not in question; rival estimates produce debate about the magnitude, but not the existence of painful psychological effects of job loss (Young 2012; Brand 2015). Pooling across all three functional forms, all of the estimates are negative, and 95.2% are statistically significant – strongly robust by the Raftery standard. Only one in 20 models report a non-significant result. If an author preferred a null result, they could make a table with a handful of additional models supporting a null conclusion. However, once the multiverse analysis is revealed, it becomes clear that those null results are knife-edge specifications that would be very difficult to substantively justify. In the opposite case – if an author preferred one of the extremely large (long left tail) estimates reported in a few matching models – these would be equally difficult to justify once the multiverse of estimates is known.

More broadly, the findings of Application III show that even if the average estimate may be quite similar across functional forms – where these is no *systematic* difference – some functional forms may be much more sensitive than others to the choice of controls. For example, in OLS, controlling for family income has no influence on the effect of job loss on wellbeing. In contrast, in CEM, controlling for family income makes the effect of job loss smaller; in PSM, controlling for family income makes the effect much *larger* (see Appendix I for more detail on the influence analysis). These three functional forms are adjusting for covariates in different ways, and sometimes with surprisingly different impacts.

4. Discussion and Conclusion

Empirical research is often seen as ‘data analysis,’ implying that the data have priority in this process. In reality, authors’ choices about how to do the analysis can be just as important as the data. All empirical results are a joint product of both the data and the analytical assumptions.

In the ‘crisis of science’ today, many are recognizing that current research procedures fail to assure an impartial assessment of the evidence. Analysts are not blinded to how their methods support or disfavor a preferred conclusion, and the modeling process is opaque and difficult for readers to evaluate. Social science needs better practices that reduce the asymmetry of information between analysts and readers, and more transparently show what estimates are possible. Edward Leamer once called this a matter of “taking the con out of econometrics” (1983).

Multiverse analysis seeks to provide a comprehensive view into how empirical results depend on modeling assumptions and analytic choices. Decisions ranging from the choice of controls, the cleaning and processing of data, and the choice of functional form – among others – all contribute to a garden of forking paths. To see into the multiverse of possible results, one must incorporate as many elements of the modeling process as possible. This study is focused on incorporating functional form robustness into a broader multiverse analysis.

Functional form uncertainty is common when the data are dichotomous. Binary outcome variables suggest the use of logit or probit. Binary explanatory or treatment variables invite the use of various matching estimators. OLS remains plausible in all of these cases. How much does the choice of functional form affect empirical results? Do these decisions just reflect stylistic preferences for, say, non-linear models? Or do they change what conclusions can be drawn from the data?

We developed a multiverse estimator that will accept all plausible combinations of controls across many different functional forms. We then applied this method to a series of empirical cases, computing over 4,300 unique model specifications. Our interest is not in any single estimate, but rather in understanding how functional form uncertainty can change the apparent evidence and possible conclusions. We arrive at three key conclusions:

(1) There is no standard practice for analyzing binary data in sociology. This leaves considerable confusion about how to interpret results and about why different functional forms are being used. Reporting estimates as marginal effects – by putting effect sizes on the same probability scale – gives a standard metric and greatly clarifies what difference there is between

OLS, logit, and probit estimates. This common reporting metric is essential to understanding the empirical literature, and is a key foundation for functional form robustness.[9]

(2) Different functional forms, in our results, have similar central tendencies. Estimating across all plausible combinations of controls, the average (or modal) estimate from functional forms – including OLS, logit, probit, propensity score matching, and coarsened exact matching – are very similar. Sometimes a functional form can shift the entire modeling distribution in one direction or another, but we find systematic shifting to be modest. The largest differences we saw were only about 10 percent: if OLS on average gave estimates of 1, other functional forms might give mean estimates of 0.9 or 1.1 – differences that would matter only for a marginally-significant result. In particular, it may not matter much whether researchers use logit, probit, or OLS – as long as the marginal effects are reported.

(3) Different functional forms can also affect the *variance* of the modeling distribution, such as when an estimator is prone to extreme estimates. We find that matching models show more variance and much longer-tailed distributions of estimates. In one application, coarsened exact matching had longer *left* tail estimates than OLS or logit, while propensity score matching had longer *right* tail estimates than the others. In the panel data application, the long tails from matching models were even more dramatic. Propensity score matching, in particular, could yield estimates either twice as large, or half as small, as any possible OLS estimate. It is not that matching *per se* gives different answers than OLS, but matching can produce much more extreme estimates – both larger and smaller – when paired with specific combinations of controls.

In summary, we find the differences between functional forms are generally not systematic, but rather more idiosyncratic. These different estimators seem to generally produce the same result, albeit each with their own degree of flexibility. Different functional forms can sometimes adjust for controls in different ways, creating haphazard results for certain combinations of controls and functional forms. This is worrisome if researchers come to favor these extreme estimates and 'convince themselves' of their validity. In particular, we question whether matching offers benefits that are worth the downside of estimate instability and a

[9] Standardized code for conducting functional form robustness in Stata is included with this article's replication package.

tendency to offer unrepresentative, knife-edge results. At minimum, comprehensive robustness analysis seems especially important to understanding estimates from matching models.

In the past, realizing the entire multiverse – the whole garden of forking paths of an empirical analysis, where all the reasonable ‘paths not taken’ are brought to light – was computationally infeasible. However, this is becoming a reality as more elements of analysis – such as functional form – are brought into a multiverse framework. Future applications of the functional form multiverse could extend in natural ways. For outcomes measured on a Likert scale, how much do results depend on modeling those as continuous (OLS) or ordinal (ordered logit / probit)? For outcomes measured as percentages or fractions, how much does it influence results to use fractional regression rather than OLS (Wooldridge 2011)? With count data, one might compare OLS to negative binomial and Poisson. We can also generally think of models that are less sensitive to outliers than OLS, such as least absolute deviations or quantile regression (Andersen 2008).

Multiverse analysis provides a platform for transparency, but still requires researcher input to acknowledge alternative modeling assumptions. This means holding a view of constructive skepticism towards one’s preferred model specification, recognizing that in a community of scholars there will be differing reasonable views about what models deserve consideration – how to code variables, what controls belong, what is the best functional form. One constructive step may be to focus more on preferred *assumptions*, rather than preferred estimates, so as to emphasize what analytical issues call for greater attention and deliberation.

Social scientists face a dual problem of model uncertainty and methodological abundance. The challenge is not simply that there are many ways of doing a thoughtful analysis. The challenge is transparency: showing whether results are robust and compelling, or – as critics often suspect – are based on knife-edge model assumptions. Multiverse methods offer a way through the crisis of credibility in social science today: showing the other empirical worlds that different model assumptions allow.

References

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

Brodeur, Abel, Nikolai Cook and Anthony Heyes. 2020. "A Proposed Specification Check for P-Hacking." *AEA Papers & Proceedings*. Vol. 110.

Arceneaux, Kevin, Alan Gerber, and Donald Green. 2010. "A Cautionary Note on the Use of Matching to Estimate Causal Effects: An Empirical Example Comparing Matching Estimates to an Experimental Benchmark." *Sociological Methods and Research*. Vol. 39(2): 256–282.

Bateman, Ian, Daniel Kahneman, Alistair Munro, Chris Starmer, and Robert Sugden. 2005. "Testing competing models of loss aversion: an adversarial collaboration." *Journal of Public Economics*. Volume 89(8):1561-80.

Battey, H., Cox, D.R., and Michelle Jackson. 2019. "On the Linear in Probability Model for Binary Data." *Royal Society Open Science*. Vol 6(5): 190067.

Blackwell, Matthew, Stefano Iacus, Gary King, and Giuseppe Porro. (2010). "Cem: Coarsened Exact Matching in Stata." Stata Users Group, BOS10 Stata Conference. 9.

Boudon, Raymond. 1994. *The Art of Self-Persuasion*. Cambridge, UK: Polity Press.

Brady, David, Ryan Finnigan, and Sabine Huebgen. 2017. "Rethinking the Risks of Poverty: A Framework for Analyzing Prevalences and Penalties." *American Journal of Sociology* 123(3):740-786

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*. Vol. 35(3):749-770.

Brand, Jennie. 2015. "The Far-Reaching Impact of Job Loss and Unemployment." *Annual Review of Sociology* 41:1.1-1.17.

Breen, R., Karlson, K. B., and Holm, A. 2018. "Interpreting and understanding logits, probits, and other nonlinear probability models." *Annual Review of Sociology*, 44, 39-54.

Camerer CF, et al. 2018. "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour*. Vol. 2: 637-644.

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

Gelman, Andrew, and Loken, E. 2014. “The Statistical Crisis in Science.” *American Scientist*, 102(6): 460-465.

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.

Gribbin, John. 2010. *In Search of the Multiverse: Parallel Worlds, Hidden Dimensions, and the Ultimate Quest for the Frontiers of Reality*.

James J. Heckman, Edward E. Leamer (eds.). 2007. *Handbook of Econometrics*. Volume 6, Part B. Elsevier.

Hellevik, O. 2009. “Linear versus logistic regression when the dependent variable is a dichotomy.” *Quality and Quantity*, *43,* 59-74.

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

Holcomb, W. L., Chaiworapongsa, T., Luke, D. A., and Burgdorf, K. D. 2001. An odd measure of risk: use and misuse of the odds ratio. *Obstetrics and Gynecology*, *94*(4), 685-688.

Iacus, Stefano, Gary King, and Giuseppe Porro. 2012. “Causal Inference Without Balance Checking: Coarsened Exact Matching.” *Political Analysis*. Volume 20(1):1-24.

Imbens, Guido. 2015. “Matching Methods in Practice: Three Examples.” *Journal of Human Resources*. 50:373-419.

Karlson, K. B., Holm, A., and Breen, R. 2012. “Comparing regression coefficients between same-sample nested models using logit and probit: A new method.” *Sociological Methodology*, *42:* 286-313.

King, Gary., and Zeng, L. 2001. Logistic regression and rare events data. *Political Analysis, 9*(2): 137-163.

Leahey, Erin, B Entwisle, P Einaudi. 2003. “Diversity in Everyday Research Practice: The case of data editing.” *Sociological Methods and Research*. Vol. 32(1): 64-89.

Long, J. Scott. 1997. *Regression models for categorical and limited dependent variables*. Thousand Oaks.

Mize, Trenton, Long Doan, and Scott Long. 2019. “A General Framework for Comparing Predictions and Marginal Effects across Models.” *Sociological Methodology*. Vol 49(1):152–189.

McCullagh, Peter, and John Nelder. 1989. *Generalized Linear Models (2nd ed.)*. Boca Raton, FL: Chapman and Hall/CRC.

Molina, Mario, and Filiz Garip. 2019. “Machine Learning for Sociology.” *Annual Review of Sociology*. Vol. 45(1):27-45.

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. “How Conditioning on Post-treatment Variables Can Ruin Your Experiment and What to Do about It.” *American Journal of Political Science*. 62: 760-775.

Mood, Carina. 2010. “Logistic regression: Why we cannot do what we think we can do, and what we can do about it.” *European Sociological Review*, *26*(1), 67-82.

Morgan, S., and David Harding. 2006. “Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice.” *Sociological Methods and Research*. Vol. 35(1): 3–60.

Mullainathan, Sendhil, and Jann Spiess. 2017. “Machine Learning: An Applied Econometric Approach.” *Journal of Economic Perspectives*. Vol. 31 (2): 87-106.

Muñoz, John and Cristobal Young. 2018a. “We ran 9 billion regressions: Eliminating false positives through computational model robustness.” *Sociological Methodology*, *48*(1), 1-33.

Muñoz, John and Cristobal Young. 2018b. “Rejoinder: Can We Weight by the Probability that the Model is True?” *Sociological Methodology*. Vol 48(1): 43–51.

Pew Research Center. 2018. “Wide Gender Gap, Growing Educational Divide in Voters’ Party Identification.” https://www.people-press.org/2018/03/20/1-trends-in-party-affiliation-among-demographic-groups/

Salganik, Matthew, et al. 2020. “Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration.” *PNAS*. Vol. 117(15):8398-403.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. 2011. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science*. Vol. 22(11): 1359–1366.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. “Increasing Transparency Through a Multiverse Analysis.” *Perspectives on Psychological Science*. Vol. 11(5): 702–712.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. 2018. “Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results.” *Advances in Methods and Practices in Psychological Science.* Vol. 1(3): 337–356.

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

Williams, Richard. 2011. “Using the margins command to estimate and interpret adjusted predictions and marginal effects.” *Stata Journal*. Volume 12(2): 308-331.

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

Wooldridge, Jeffrey. 2011. “Fractional response models with endogenous explanatory variables and heterogeneity.” http://www.stata.com/meeting/chicago11/materials/chi11_wooldridge.pdf.

Young, Cristobal. 2009. “Model uncertainty in sociological research: An application to religion and economic growth.” *American Sociological Review, 74,* 380-397.

Young, Cristobal. 2018. “Model Uncertainty and the Crisis in Science.” *Socius: Sociological Research for a Dynamic World*. Issue 4:1-7.

Young, Cristobal, and Kathleen Holsteen. 2017. “Model uncertainty and robustness: A computational framework for multimodel analysis.” *Sociological Methods and Research, 46*(1), 3-40.

Appendix I. Model Influence

Model influence analysis focuses on how the adoption of a control variable (or more broadly any model ingredient) changes the coefficient of interest. After calculating all estimates in the model space, influence analysis dissects the determinants of variation across models. Though not the main focus of this article, understanding model influence is a core part of multiverse analysis (Young and Holsteen 2017).

When deciding which control variables to include, the statistical significance of a control variable has little bearing on whether it influences the coefficient of interest. Sometimes highly-significant controls make no difference for the results, while other times including non-significant controls can change the results dramatically.

Consider two simple nested models:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad (A.1)$$

$$Y_i = \alpha + \beta^* X_i + \delta Z_i + \varepsilon_i^* \qquad (A.2)$$

We are interested in how changes in X_i affect the outcome, so β is the coefficient of interest. In equation $A.2$, Z_i is a control variable, and its relationship to the outcome, Y_i, is given by δ. When considering control variables, it is conventional to report the δ estimate. But we are more interested in 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. We calculate an influence score for each control variable (and ultimately, other aspects of model specification). This can be thought of as a meta-analysis of the model space (Stanley and Doucouliagos 2012). Using results from the full 2^n estimated models, what elements of the model specification are most influential for the results? We formulate an *influence regression* by using the coefficients of interest from all models as the outcome to be explained. The explanatory variables in the influence regression are dummies for the control variables included in each of the models. For *n* possible control variables, we create a set of dummy variables $\{D_1 \dots D_n\}$ to indicate when each control variable is in the model that generated the estimate. This meta influence regression has *J* = 2^n observations (i.e., coefficient estimates):

$$\widehat{\beta_J} = \alpha + \theta_1 D_{1j} + \theta_2 D_{2j} + \cdots + \theta_P D_{Pj} + \varepsilon_j \qquad (A.3)$$

In $A.3$, $\widehat{\beta_J}$ is the regression estimate from the *j*-th model. The influence coefficient θ_1 shows the expected change in the coefficient of interest ($\widehat{\beta_J}$) if the control variable corresponding to D_1 is included in the *j*-th model. Each coefficient estimates the conditional mean $\Delta\beta$ effect for each control variable. This is the statistic analysts and readers typically want to know about the impact of a control variable: how does it, on average, affect the coefficient of interest?

For functional form influence, our main goal is to see whether different functional forms adjust for controls in similar ways. In other words, does the influence of a control variable depend on which functional form is adopted?

Application I: Discrimination in Mortgage Lending

In application I, the $\Delta\beta$ estimates are all quite similar across functional forms – hence the high degree of functional form robustness. In all three functional forms, controlling for bad credit history has the largest effect on the estimates. In OLS, including this control raises the racial disparity estimate by 0.032. Since the baseline estimate is negative (i.e., lower mortgage acceptance for blacks) this positive influence effect means that after controlling for bad credit the disparity is closer to zero, falling by 22.1 percent. This percent change (column 2) is simply the influence effect divided by the average estimate from OLS (-0.144). Likewise, in all three functional forms, the most 'negative' influence comes from controlling for gender, which has an influence effect of -0.005. This means that when the dummy variable 'female' is in the model, the racial disparity estimate is half a percentage point larger in absolute magnitude.

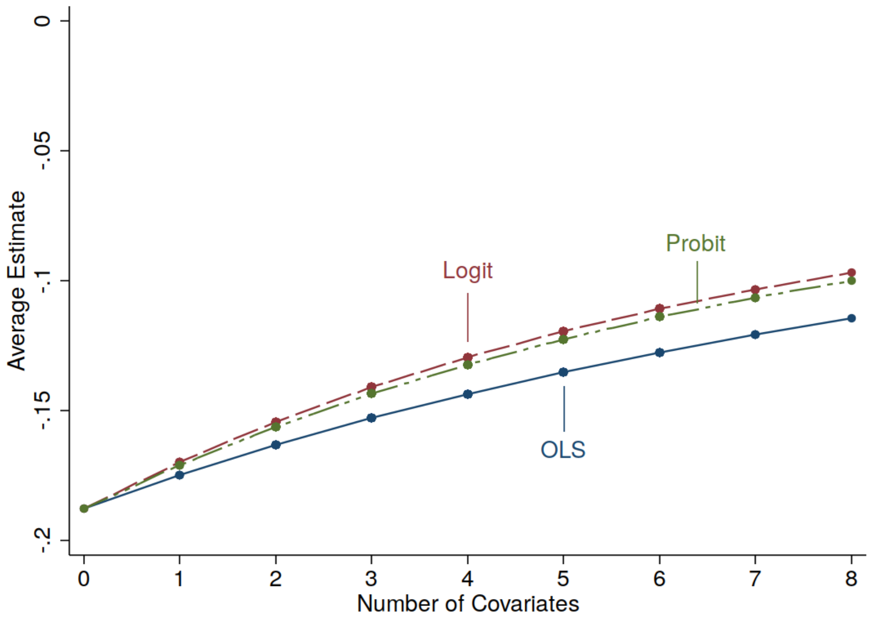
Table A1: Influence Effects for Effect of Race on Mortgage Lending

	OLS		Logit		Probit	
	effect	% change	effect	% change	effect	% change
Bad credit history	0.032	-22.1	0.036	-27.6	0.037	-27.5
Denied PMI	0.019	-13.3	0.014	-11.0	0.015	-11.3
Loan-to-value ratio	0.013	-9.2	0.026	-19.8	0.024	-17.9
PI ratio	0.010	-6.7	0.015	-11.1	0.014	-10.1
Married	0.003	-2.4	0.005	-4.0	0.005	-3.6
Housing exp. ratio	0.002	-1.3	0.003	-2.5	0.003	-2.3
Self employed	-0.002	1.5	-0.004	2.7	-0.004	3.1
Female	-0.005	3.6	-0.009	6.5	-0.008	6.1
Avg estimate	-0.144		-0.131		-0.134	
N	256		256		256	
R2	0.986		0.972		0.976	

Overall, the three functional forms give comparable influence scores for each variable – meaning that they control for the variables in very similar ways. The main difference across functional forms is seen in the loan-to-value ratio. In OLS, this variable has an influence of 1.3 percentage points; in Logit, the influence is 2.6 percentage points; in Probit, the influence is 2.4 percentage points. This is the main reason why the non-linear functional forms tend to give estimates closer to zero than OLS: the loan-to-value ratio is more influential (shrinking the estimate) in both the non-linear models. In other words, Logit is different from OLS because it differently controls for the loan-to-value ratio.

To give a more general sense of how control variables influence the findings, figure A1 shows the coefficient of interest depending on how many controls are included in the model. Adding control variables pushes the (negative) parameter estimate closer to zero, regardless of functional form.

Figure A1: Effect Size by Number of Control Variables



Application II: Voting for Trump in the 2016 Election

In application II, many of the $\Delta\beta$ estimates are similar across functional form. For example, conditioning on democratic party affiliation reduces the effect of college on voting for Trump by 49 percent in OLS, 50 percent in logit, 57 percent in PSM, and 43 percent in CEM.

One could debate whether party affiliation is a post-treatment, endogenous control in this case. In any event, this variable clearly has shared explanatory power with college degree.

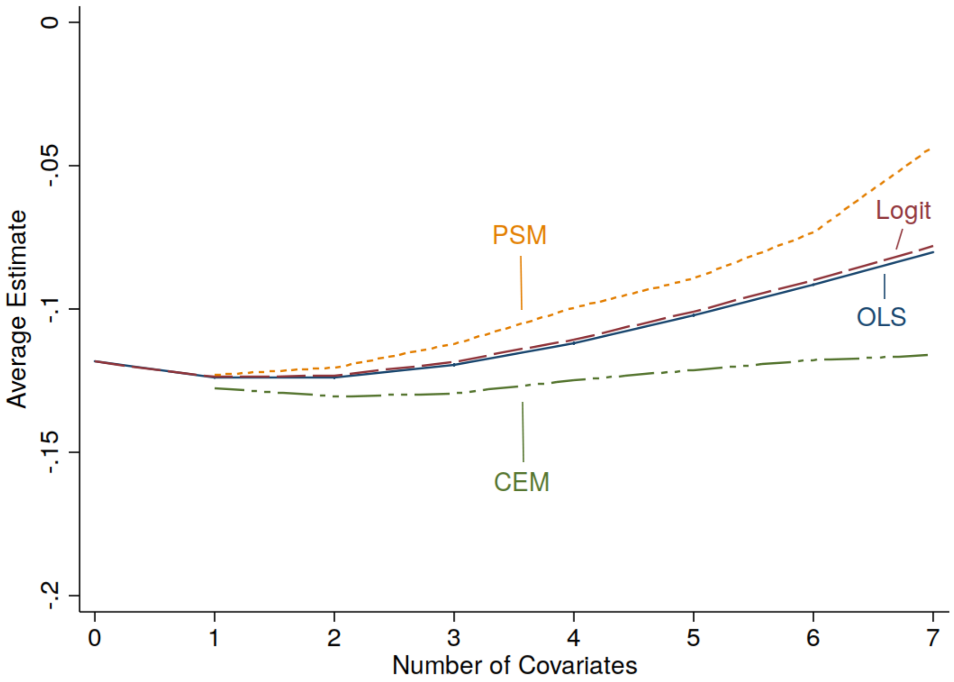
Table A2: Influence Effects for Effect of College Degree on Voting for Trump

	OLS		Logit		PSM		CEM	
	effect	% change	effect	% change	effect	% change	effect	% change
Democrat	0.056	-49%	0.060	-50%	0.060	-57%	0.055	-43%
Income	0.010	-8%	0.010	-8%	0.026	-25%	0.000	0%
Male	0.001	-1%	0.000	-1%	0.004	-4%	0.000	0%
Age	0.000	0%	0.000	0%	0.006	-6%	-0.003	2%
Age squared	0.000	0%	0.000	0%	0.005	-5%	0.000	0%
Marital status	-0.001	1%	0.000	2%	-0.002	2%	-0.003	3%
Non-white	-0.018	16%	-0.020	14%	-0.023	22%	-0.030	24%
Avg. estimate	-0.114		-0.113		-0.104		-0.126	
N	128		128		127		127	
R2	0.940		0.938		0.874		0.847	

The largest difference across functional form is seen when conditioning on income. In both OLS and logit, controlling for income reduces the coefficient of interest by 8 percent. However, in PSM, controlling for income reduces the effect size by 25 percent, and in CEM controlling for income has no effect on the estimate of interest. This illustrates how different functional forms can adjust for controls in different ways.

Figure A2 shows that the number of control variables, per se, has modest effect on the coefficient of interest. PSM is most sensitive to the number of controls (pushing the estimate towards zero), CEM is the least sensitives (the number of controls is irrelevant), while the OLS and logit curves lie between these ranges.

Figure A2: Effect Size by Number of Control Variables



Application III: Unemployment and Wellbeing

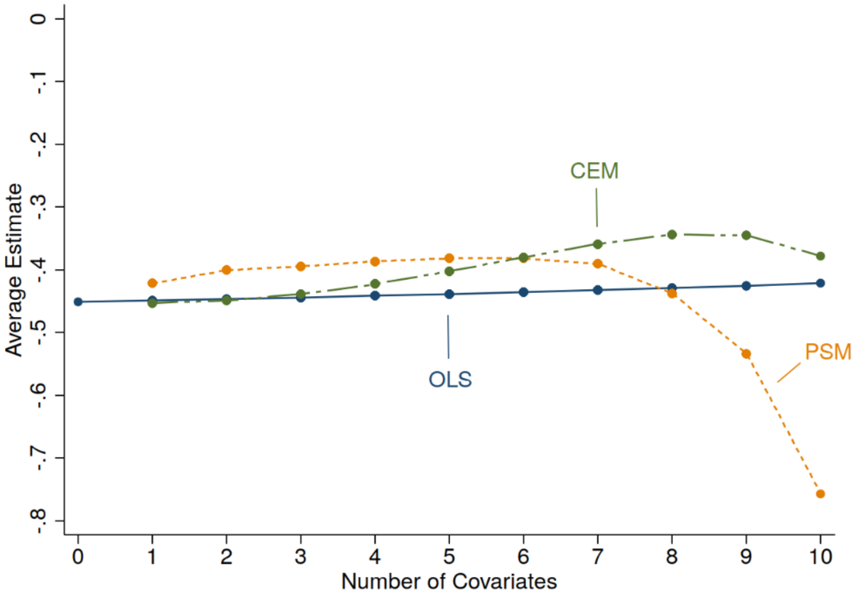
In application III, the influence of many control variables is different across functional forms, specifically OLS and matching. Even the signs of the influence scores frequently switch depending on the functional form. For example, how does controlling for self-reported food problems ('not having enough food') change the estimated effect of job loss? Propensity score matching shows a notable *negative* influence of this control (-0.021), while coarsened exact matching (0.030) and OLS (0.011) both find a positive influence of the control. A very similar pattern is seen for the variable 'children at home.' For homeowner status, CEM and PSM agree on the sign of its influence (negative), but in CEM the influence is more than ten times larger than in PSM (-.044 and -.003, respectively). Another striking difference across functional form is for income: it has no influence in OLS (-.001), positive influence in CEM (.017), and a large negative influence in PSM (-.046). These differences in the influence scores – in how a control variable affects the coefficient of interest – seem very idiosyncratic.

Figure A3 gives a broader view of these idiosyncratic influence effects. In OLS, the number of controls has no effect on the job loss coefficient. In CEM, additional controls tend to modestly shrink the estimate towards zero. In PSM, controls generally have no effect, until there are many of them, at which point adding controls greatly increases the estimate (grows it away from zero). In this troubling case, an author's (extreme) result could depend not on *which* controls are in the model, but simply on the discontinuous effect of including *many* controls. The PSM results, in particular, seem haphazard.

Table A3: Influence Effects for Effect of Job Loss on Subjective Wellbeing

	OLS		PSM		CEM	
	effect	% change	effect	% change	effect	% change
Not enough food	0.011	-2.5%	-0.021	5.4%	0.030	-7.5%
Not desired food	0.010	-2.3%	0.008	-2.0%	-0.005	1.4%
Eligible for unemployment	0.004	-0.8%	0.022	-5.7%	0.118	-29.4%
Children at home	0.002	-0.5%	-0.026	6.7%	0.011	-2.7%
Homeowner	0.001	-0.2%	-0.003	0.8%	-0.044	11.0%
Part-time work	0.001	-0.2%	0.036	-9.3%	0.030	-7.4%
Married	0.001	-0.1%	-0.003	0.8%	0.005	-1.3%
Zero/negative wealth	0.001	-0.1%	-0.002	0.4%	0.016	-4.0%
Food stamps	0.000	0%	-0.002	0.6%	0.006	-1.6%
Log income	0.000	0%	-0.046	11.7%	0.017	-4.2%
Avg. estimate	-0.439		-0.391		-0.401	
N	1,024		1,023		1,023	
R2	0.916		0.147		0.736	

Figure A3: Effect Size by Number of Control Variables


Average Estimate
CEM
OLS
PSM
Number of Covariates