Functional Form Robustness:

Advancements in Multiverse Analysis

Cristobal Young, Cornell University

Sheridan A. Stewart, Stanford University

Aug. 6, 2021. W = 9,096

Abstract

Social scientists face a dual problem of methodological abundance and model uncertainty. There are many ways to conduct an analysis, but the true model is unknown. Multiverse analysis embraces this challenge by recognizing 'many worlds' of modeling assumptions, using computational power to reveal a large set of plausible estimates. We advance from possible control variables to the functional form multiverse, considering OLS, logit, Poisson, inverse probability weighting, and various forms of matching. How much do results depend on the choice of functional form? Are some functional forms more stable across sets of controls than others? The multiverse estimator takes all plausible combinations of functional forms and control variables, showing what the data can support under alternative assumptions. Estimating thousands of specifications, we study three empirical cases: the effect of unemployment on wellbeing, the role of education in voting for Donald Trump, and discrimination by skin tone in professional soccer. Finally, multiverse analysis is compared with a "many analysts" crowdsourcing project. In our cases, we find that (1) most functional forms yield similar estimates on average, meaning that omitted variable bias is likely more worrisome than incorrect form; (2) functional forms differ by estimate stability / researcher degrees of freedom; (3) no functional form tested here is more stable than OLS; (4) matching estimators are notably unstable; and (5) multiverse analysis shows the same mean as a many-analysts crowdsourcing study, but still does not capture the full range of idiosyncratic results that human researchers can produce from the same data set.

Keywords: Methodology; inequality; unemployment; voting; discrimination

Corresponding author: Cristobal Young, 384 Uris Hall, Cornell University, Ithaca, New York, 14853. All code and data for this article will be publicly available upon publication. Contact the authors at cristobal.young@cornell.edu or sastew@stanford.edu.

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. Model specification involves two core elements: choosing which controls to include, and choosing a functional form that relates the left-hand and right-hand sides of the equation. Many different modeling combinations are plausible, creating a garden of forking path that offers multiplying chances to find statistically significant results (Gelman and Loken 2014; Borges 1962). 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). There is wide concern today that many statistically-significant results published in scientific literatures are not robust and fail to replicate (Camerer et al. 2018; Christensen, Freese, and Miguel 2019; Engzell, Mood, and Jonsson 2020). 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 (Young 2009). In an age of computational power, we need better ways of revealing the range of estimates the data can support.

This study is part of a program to build comprehensive multiverse analysis encompassing all the major analytical decisions that feed into an empirical estimate (Muñoz and Young 2018; Young and Holsteen 2017; Steegan et al. 2016; Brodeur et al. 2020; Leamer 1983). The central goal is to improve the transparency of research. 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. Multiverse analysis allows one to backtrack along the garden of forking paths, showing the sensitivity of each modeling assumption to reasonable alternative specifications. It uses computational power to estimate hundreds or thousands of theoretically-reasonable models, estimating all unique combinations of model ingredients. The result is a modeling distribution of estimates, similar to a bootstrap sampling distribution – a bootstrap of the model assumptions. This can show how each model ingredient affects the coefficient of interest.

We build on existing methods for the multiverse of control variables (Young and Holsteen 2017), expanding to include functional form decisions. Often, there is limited consensus on what functional form researchers should use: logit, probit, the linear model, or something else? In matching models, researchers face a choice among many approaches to matching, such as propensity-score and coarsened-exact matching (King and Nielson 2019; Rippollone et al 2020). All of these offer different ways of estimating an effect by conditioning on observables. How much do the results depend on choice among these possible estimators? Is the choice of functional forms similar to choosing among control variables? In applied data, is "incorrect functional form" bias similar to omitted variable bias?

We demonstrate the multiverse methodology in three sociological applications: the effect of job loss on subjective wellbeing, the role of education in voting for Donald Trump in 2016, and the effect of skin tone on receiving red cards in professional soccer. In these applications we find that some methodological concerns – such as linear regression v. logit – may be overwrought, while matching estimators in particular should be treated with caution. We also compare the computational multiverse to a pool of estimates generated by many analysts in a crowdsourcing study. Does the multiverse algorithm reproduce the work of many human analysts?

1. Methodological Abundance

The variety, novelty, and abundance of statistical techniques available in social science is remarkable. The *Handbook of Econometrics*, for example, runs across 77 chapters and more than 5,000 pages (Heckman and Leamer 2007). New methods continually aim to improve estimation; however, these methods come with unknown sources of error and bias; they expand the garden of forking paths and increase the freedom of researchers to discover false positive results (Glaeser 2008; Muñoz and Young 2018).

Different analysts studying a topic scarcely ever use the same model specifications. The problem is not simply that some scholars are doing bad analysis: divergent results are common even among equivalent levels of competence and methodological quality. The "many analysts, one data set" project, for example, recruited 29 teams of social scientists to probe discrimination by skin tone in professional soccer (Silberzahn et al. 2018; c.f. Magnus and Morgan 1999). The result was wide-ranging empirical strategies, diverse results, and a roughly 70-30 split on the

empirical conclusion. The range of findings was not readily explainable: neither the researchers' level of experience 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. Different researchers studying a data set do not consistently produce the same result.

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 (Gribbin 2010).

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.' 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.' Scholars suffer from the problem of motivated reasoning, highlighted in classic books like *The Art of Self-Persuasion* (Boudon 1994). Researchers can convince themselves that favorable model assumptions are clearly superior. Objectively incorrect beliefs can be 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 trials of new medical treatments, researchers are naturally hopeful that the treatment will work – which could bias the results in multiple ways. However, during the study, researchers do not know who is in the treatment or control group (Shultz and Grimes 2002). Blinding procedures allow scholars to have non-objective beliefs, while still producing objective evidence: the methods of science insulate the evidence from the personal views of the scientist. When blinding is not viable, however, 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

Generating a model specification involves two central elements: identifying controls to include, and a functional form that relates the left-hand and right-hand sides of the equation. We assume there is a true model that generated the data: a correct set of controls, and a correct functional form. Research is evaluated and contested based on how well applied work approximates these elements of the (unknown) true model. We begin with the multiverse of controls and then extend to functional form.

The set of possible control variables in an analysis can be represented as

$$(1) \quad y = f(x, Z, Q)$$

where 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 strict assumptions about controls to ensure the model space is credible. Q represents model uncertainty: prior theory does not inform which elements belong in the model, leaving different analysts to make their own debatable judgements (Young and Holsteen 2017; Leamer 1983).

The purpose of multiverse analysis is to understand how and how much 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 functional form assumed to be correct. 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. In this simple multiverse, a model could include q_1, or q_2, both, or neither. In this case, equation (1) expands to the following set of four possible models:

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

These equations represent different reasonable ways of specifying the model given the uncertainty, and offer four plausible estimates of β_1. As the number of Q model ingredients

increases, the model space grows 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 ten possible controls, there are $2^{10} = 1{,}024$ unique models, and with 20 possible controls there are over one million.

1.2 The Multiverse of Functional Forms

The well-known omitted variable bias formula gives guidance to what is expected when a control variable is omitted from a regression. However, there is no analogous formula for "wrong functional form bias," though scholars are often just as concerned about functional form as about choice of controls. What happens when using an incorrect functional form? Does that bias estimates in a similar way as a wrong (or omitted) control variable?

In the existing literature, selection among controls has been far more tractable to analyze than choice of functional form. We use brute force computation and, when needed, coefficient rescaling to show how much results depend on functional form assumptions.

To ground thinking about many possible functional forms, consider 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 that can link the left- and right-hand sides of an equation. The link could be the linear probability model (LPM), logit, probit, ordered logit or probit, multinomial logit, Poisson, negative binomial, and more. This is an analogy, as we extend this to consider algorithmic "functional forms" such as matching, as well as hybrid approaches like inverse probability weighting. 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,

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

where the function *f* is now a variable with subscript *m*, indicating a vector of models $[f_1, f_2, \ldots, f_M]$. 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 to allow a range of possibilities – each paired to different assumptions about which functional form would provide the least-biased estimate of β_1. For any link function that changes the scale of the β_1 coefficients, the resulting estimates are

converted into comparable units – either transforming odds ratios into average marginal effects, or marginal effects into implied odds ratios.

With *n* uncertain control variables and *m* plausible link functions and algorithms, there are $2^n \times m = J$ unique models. Retrieving the estimated effect of x from each of these models provides a modeling distribution that is 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.[1] 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 an analysis, the modeling distribution shows the range of estimates that would 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. Statisticians conventionally evaluate estimators by focusing on the mean and the variance of their *sampling* distributions – behavior in repeated sampling. We are 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, analogous to omitted variable bias.[2] 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 distributions. 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 (Coker, Rudin and King

[1] Recall that the term standard error simply means estimated standard deviation.

[2] For this analogy, imagine that an incorrect functional form has an omitted feature that biases the effect estimate, and correcting the functional form removes that bias.

2020). While sampling variance is conventionally reported in the confidence intervals or standard errors of a study, the modeling variance shows researcher degrees of freedom.

Below, we briefly discuss the functional forms empirically tested in this study.

2.1 Matching versus Regression

Matching is a common method for effect estimation when the treatment variable is binary. Matching aims for balance in covariates between 'treatment' and 'control' groups. The goal is to match each treatment case with at least one control case that has (very) similar observable characteristics. 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 Nielsen 2019). Still, matching is often invoked for properties that may offer superior estimation over OLS. Matching is non-parametric, or at least semi-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. A downside of matching, however, 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 match (Stuart 2010; Morgan and Harding 2006). This leaves considerable discretion in choosing a preferred method.

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

$$(7) \quad \pi_i = \Pr(X_i = 1 | \, z, Q_n)$$

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

$$(8) \quad E[Y|x, z, Q_n] = E(Y| \, x = 1, \widehat{\pi_i}) - E(Y|\mathrm{x} = 0, \widehat{\pi_i})$$

In coarsened exact matching (CEM), rather than using the first-stage regression of equation (7), an algorithm temporarily coarsens the continuous z and $Q_n$variables into bins that allow treated cases to be matched to similar control cases (Iacus, King, and Porro 2012). Cases with the same values for all coarsened variables are grouped together into a stratum (e.g., high, medium, and low). The goal is to coarsen z and Q_n just enough to match treated ($x = 1$) and control ($x = 0$) observations into comparable strata. The CEM estimate comes from aggregating all comparisons of treated and control cases within each stratum. There is an active debate over the properties of these two matching approaches. Some argue that "propensity scores should not be used for matching" (King and Nielsen 2019), while others report that CEM often leads to "high bias and low precision" relative to PSM (Ripollone et al 2019:613).

In some ways, matching methods are emblematic of the challenge of novelty in statistical methods. While new techniques may offer improved estimation or causal inference, they come with new uncertainties and poorly understood biases. New techniques often generate considerable enthusiasm, but merit large doses of skepticism because of the researcher degrees of freedom that novelty brings (Glaeser 2008). Some scholars argue 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 Nielsen 2019). Others caution 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, it might worsen them, or might not make much difference at all.

2.2 Inverse Probability Weighting

The propensity score π_i (eq. 7 above) is not only used for matching, but can also be used as a weight for a regression analysis when there is concern about missing data. When data are missing completely at random, listwise deletion is an unbiased solution. If one deletes all cases that have any missing information, the subsample will still be representative of the full sample. When data are missing due to unobserved systematic factors, there is no obvious solution for potential bias. However, there are intermediate cases of systematic missingness that can be corrected. Imagine that respondents are more likely to be missing if they are in the treated group

versus the control group (or when x is large rather than small). If that difference can be explained by observed variables (such as age, education, home ownership, etc), then inverse probability weighting can correct for selection bias in the treatment effect (Wooldridge 2007). In the case of binary treatments, for individuals in the treated group, the weight is $w_i = {}^{1}/_{\widehat{\pi_i}}$ and for those in the control group $w_i = {}^{1}/_{1-\widehat{\pi_i}}$. Re-specifying equation 2 with inverse probability weight gives

$$(9) \quad \frac{y_i}{w_i} = \frac{\beta_1 x_i}{w_i} + \frac{\beta_2 z_i}{w_i} + \frac{\varepsilon_i}{w_i}$$

Intuitively, IPW gives greater weight to control observations that are more similar in the probability of treatment as the treatment cases. This puts weight on ideal control cases: those most likely to receive the treatment, but whom were not treated.

2.3 Linear vs. Non-linear: logit, probit, LPM

Modeling choices are often interconnected with the structure of the data. When an outcome variable is binary, sociologists commonly adopt a logit model. Economists traditionally favored probit in such cases (Angrist and Pischke 2009:102-7). Either way, research with binary outcomes often uses a nonlinear functional form for a variety of reasons (Long 1997). In contrast to the linear equation 2 above, logit and probit are written as

$$(10) \quad \text{Logit:} \quad \Pr(Y = 1|\, x, z) = (1 + e^{-(\beta_1 x_i + \beta_2 z_i)})^{-1}$$

$$(11) \quad \text{Probit:} \quad \Pr(Y = 1|\, x, z) = \Phi(\beta_1 x_i + \beta_2 z_i)$$

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 nonlinear effects of x on y. However, there are also serious shortcomings of these nonlinear models, including problems that can prevent convergence[3], and challenges in interpreting results across nested models[4] (King and

[3] Contributing factors in convergence failure 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).

[4] Briefly explain the issue of non-comparable coefficients – rescaling due to changes in the error term across specifications (Mood 2010).

Zeng 2001; Mood 2010). In light of various problems with nonlinear models for binary outcomes, Breen, Karlson, and Holm (2018) recommend the linear probability model as the default.

2.3.1 Scale of Coefficients

A challenge in functional form robustness is that different functional forms may report estimates on different scales (Williams 2011; Mize, Doan, and Long 2019). The linear model produces 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. However, in post-estimation, odds ratios can be converted into average marginal effects, and alternatively marginal effects can be converted into implied odds ratios (Greene 2012:689-90). The conversion is possible since the odds of an outcome (Y=1) is the probability the outcome occurs divided by the probability that it does not occur. If we think of the probability (P) of an outcome in treated (t) and control (c) groups, then the marginal effect is $ME = P_t - P_c$, while the odds ratio is $OR = \frac{P_t/1-P_t}{P_c/1-P_c}$. The necessary probabilities for either metric can always be calculated regardless of preferred functional form (see Appendix I for more detail).

In summary, there are many different functions or algorithms that relate the left- and right-hand sides of a model. This variety of functional forms 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 may provide 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.

3. Applications

We now apply the functional form multiverse analysis in a series of applied cases. First, we analyze data from the Panel Study of Income Dynamics on unemployment and wellbeing. This features a binary treatment (job loss) allowing comparison between matching and regression. Next, we examine American National Election Study data on the effect of a college

degree in voting for Donald Trump in 2016; as both the outcome and explanatory variables are binary, we can compare linear regression, logit and matching. Our third application considers the effect of skin tone on receiving red cards in professional soccer. These data are drawn from a crowdsourcing study in which many analysts considered logit, linear regression, and Poisson as possible functional forms (as well as choosing among a set of possible controls). This invites a correspondence test for the range of results a in computational multiverse analysis, compared to the estimates of many different scholars each contributing one preferred estimate.

Application I: Unemployment and Wellbeing

How does job loss affect an individual's subjective wellbeing? We draw on data from the Panel Study of Income Dynamics to address this question. With a binary treatment variable (job loss = 1, no job loss = 0), we compare OLS, inverse probability weighting, propensity score matching, and coarsened exact matching estimates of the effect of losing a job in a two-wave panel. We use fixed-effects analysis examining the transition from working to unemployed (Young 2012: 622, Table 3).[5] A set of 10 plausible time-varying controls includes income, part-time work, children in the home, marital status, homeowner status, 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 (Montgomery, Nyhan, and Torres 2019; Elwert and Winship 2019). It is 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 four different functional forms: OLS, IPW, PSM, and CEM. This broader multiverse yields a modeling distribution of 4,090 point estimates. In our baseline linear regression model, including all 10 controls, job loss leads to a 0.4 standard deviation drop in wellbeing (table not shown). As Young (2012) notes, job loss has a larger effect on well-being than any other variable.

[5] 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 1. Modeling Distributions: Effect of Job Loss on Subjective Wellbeing

Panel (a) of Figure 1 shows the modeling distributions from OLS, IPW, CEM, and PSM separately, illustrating how the multiverse of controls depends on functional form. The modeling distribution from OLS is highly concentrated – appearing largely as a spike in estimates near the baseline of -0.4: none of the controls meaningfully change the results. IPW, CEM, and PSM all have similar means, but show a much wider range of estimates. Table 2 provides more detail: the average estimate by functional form is -.44 (OLS), -.46 (IPW), -.040 (CEM), and -0.39 (PSM), revealing little systematic difference. However, the ranges are far larger using non-OLS functional forms. The modeling standard errors are many times greater using IPW (0.07), CEM (0.08) or PSM (0.09) than for OLS (0.01). This reflects idiosyncratic differences across functional forms. In particular, PSM permits point estimates that are more than twice as large in magnitude than any OLS estimate, as well as some small and non-significant estimates. These non-systematic differences mean that the results are more open to debate among scholars on relatively loose methodological grounds. It also means that researchers have more (undisclosed) flexibility to choose a preferred estimate.

Table 1. Functional Form Robustness of Effect of Job Loss on Wellbeing

Nevertheless, as Figure 1, panel (b) shows, the core empirical finding of psychological harm from unemployment is not in serious question; rival functional forms produce debate about the magnitude, but not the existence of painful psychological effects of job loss (Young 2012; Brand 2015). Pooling across all four functional forms, all of the estimates are negative, and 96 percent are statistically significant – strongly robust by the Raftery (1995) standard. In these 4,000+ models, fewer than one in 20 report a non-significant result. If an author preferred a null result, they could curate a table of models supporting a null conclusion. However, it is clear those null results are fragile specifications that would be very difficult to substantively justify. In contrast, 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 I show that even if the *average* estimate may be quite similar across functional forms – where these is little *systematic* difference – some functional forms may be much more sensitive to the choice of controls. Figure 2 offers one visualization of model influence: how the number of controls in a specification affects the coefficient of interest, by functional form type. OLS estimates are almost entirely insensitive to the number of controls included in a specification. For coarsened exact matching, the estimates trend slightly closer to zero as the number of controls rises, while in IPW they trend slightly *away* from zero. For propensity score matching, the estimated effect of job loss grows dramatically when there are more than 8 controls included. These three functional forms are adjusting for covariates in different ways, with some surprisingly different impacts. For example, in OLS, controlling for family income has no influence on the effect of job loss on wellbeing; in 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 II, Table A3). The intersection of functional form and control variables often leverages chance associations in the data in different ways.

Figure 2: Effect Size of Job Loss by Number of Control Variables

Application II: Voting for Trump in the 2016 Election

Who voted for 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, reflecting a realignment in party coalitions. We draw on the American National Election Study to analyze the effect of having a college degree on voting for Trump.

This application compares the linear model, logit, IPW, and the matching estimators. 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 estimated with four different functional forms: the linear model, logit, and propensity score (PSM) and coarsened exact (CEM) matching.

Figure 3 shows the results. The modeling distributions from all four functional forms clearly overlap, but with noteworthy differences. Panel (a) reports the distribution of results from

the linear model. Depending on the set of controls used, one can find estimates ranging from a 7 to 16 percentage point 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 II, influence analysis shows this variable is party affiliation (republican v democrat): 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 (Morgan 2018). Either way, informed consumers of this research should be aware that the decision to control for party affiliation is clearly influential for the results.[6]

In panel 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 LPM and logit in this case. In panel c, we add the distribution from coarsened exact matching (CEM). While this distribution largely overlaps with those of LPM and logit, it is wider, with some estimates that are larger in magnitude (longer left tail). In panel 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, panel (e) shows 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. This shows idiosyncratic influence of controls on the modeling distribution, regardless of functional form.

Table 2. Functional Form Robustness of Effect of College Degree on Voting for Donald Trump in 2016

Figure 3. Modeling Distributions: Effect of College Degree on Voting Donald Trump in 2016

[6] 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 as controls for the current analysis (the effect of education in voting for Trump).

Table 3 offers further detail. Logit and the linear model have nearly the same point estimates, sample standard errors, and modeling standard errors. Matching methods produce similar average estimates, but the differences are noticeable. PSM gives the smallest average estimate (-0.104), and for 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. CEM and especially PSM have larger ranges and modeling standard errors than the linear model or logit. These longer-tail distributions are worrisome because researchers have greater flexibility in their conclusion. Some PSM models allow a null conclusion, but 98 percent of this multiverse finds that a college degree reduces the chance of voting for Trump, by about 11 percentage points on average. The pooled distribution has twice the range of either the OLS or logit distributions.

Influence analysis continues to show that different functional forms control for covariates in somewhat different ways. For example, in both LPM and logit, controlling for income reduces the effect of education on voting for Trump 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 (Appendix II: Table A3). This illustrates how different functional forms can adjust for controls in different ways.

Figure 4 shows that the influence of adding additional controls depends on functional form. PSM is most sensitive to the number of controls (pushing the estimate towards zero), CEM is less sensitive (the number of controls is irrelevant), while the linear model and logit curves – which are virtually indistinguishable from one another – lie between these ranges.

Figure 4: Effect Size by Number of Control Variables. Voting Data

Application III: Crowdsource Comparison - Discrimination in Professional Soccer

Finally, we compare multiverse analysis to the results of a many-analyst crowdsourcing project (Silberzahn et al. 2018). The empirical case focuses on discrimination in professional soccer: whether players with darker skin tone receive more red card penalties from referees. Drawing on one data set, 29 teams of researchers conducted their own analysis and each submitted their preferred estimate. The key value of this crowdsourcing, the authors argue, is leveraging "skills, perspectives and approaches to data analysis that no single analyst or research

team can realistically muster alone" (ibid:354). However, the teams mostly chose between one of three functional forms and selected from the same set of 10 possible control variables – the exact territory of a functional form multiverse. Taking all combinations of the core modeling features, does the multiverse algorithm reproduce the distribution of results from many analysts?

Multiverse skeptics may be concerned that the computational approach generates many odd or implausible models that human researchers would never select. This implies that many analysts would provide a more credible and sensible distribution of results than a computational approach – a distribution more closely bound to the best model. On the other hand, an algorithm has no preference for the outcome, does not engage in motivated reasoning, conducts no further data processing, and will not do any 'extra work' after seeing the results. This offers an intriguing correspondence test between computational and human analytical approaches.

In professional soccer, red cards are penalties for player misconduct or violence on the field, and referees have wide discretion in assigning them. Are referees more likely to give red cards to dark-skinned players than to light-skinned players? To address this question, Silberzahn et al (2018) collected data on European leagues. Skin tone is measured on a five-point scale ranging from very light to very dark, coded from player photographs. The outcome variable is at the player-referee dyad level, recording the number of red cards each player received from each referee they played with in their career. There are 146,028 dyads of players and referees in the data. The number of red cards in each dyad ranges from 0 to 2, with a mean of 0.004 – red cards are very rare events at the dyad level. In other words, the outcome is nearly binary, but is technically a count variable.

The original research teams nearly all used one of three general functional forms: logistic (52 percent), Poisson (21 percent), and the linear model (21 percent). Following Silberzahn et al (2018), in all specifications the main effect of skin tone is calculated as an odds ratio – the odds of a very dark skin player receiving a red card over that of a very light skin player.[7] Original authors chose from a set of 10 possible right-hand side variables, with teams using between one and eight of those – reflecting a range of preferences for parsimonious versus richly-specified models. The controls include player height, weight, and age; the number of goals, ties, victories,

[7] Poisson models report the incident rate ratio (IRR). When the outcome is a rare event (eg, less than 10%) as in this case, the odds ratio and the IRR are on the same scale. For non-rare events, the OR is always larger than the IRR and direct comparisons require a conversion formula. For the linear model, we convert marginal effects to odds ratios using the margins command.

and defeats, the number of yellow cards they received from the referee, their position on the team, and their national league (France, Germany, England, or Spain). The number of games played with each referee is treated as a necessary control – serving as an exposure variable. With 10 plausible controls there are 1,024 possible specifications for each functional form.

Figure 5. Multiverse Results: Effect of Skin Tone on Red Cards

Figure 5 presents the functional form + controls multiverse, compared to the full set of results from the original study. Both distributions are strongly centered around an odds ratio of 1.3 – dark-skinned players have 30% higher odds of receiving a red card than light-skinned players. Functional form assumptions have little influence on the average estimate. However, results from many analysts show a considerably wider range and variance than the multiverse analysis. In particular, the many-analysts show a much longer left tail, including opposite-signed effects (odds ratios less than 1).[8] As Table 4 shows, the pooled multiverse effects range from 1.23 to 1.64 (i.e. from 23 to 64 percent higher odds), and all estimates are statistically significant. The many-analysts results range from 0.88 to 1.71, and only 69 percent are statistically significant. The modeling standard error in many analysts (0.170) is more than three times greater than in the computational multiverse analysis (0.048). In other words, human analysts produce a much wider range of results than does the multiverse algorithm.

Table 3. Functional Form Robustness of Effect of Skin Tone on Receiving Red Cards

In discussing their results, Silberzahn et al (2018) focused on choice of controls and functional form as representing the modeling differences across analysts and results. However, these core elements do not fully explain the diversity of their results. By comparison, the

[8] The original study included two outlier (far right tail) non-significant estimates which turned out have errors in scaling. These two estimates (2.93, 2.88) were more than twice the magnitude of all other estimates, despite being non- or barely-significant. On close inspection re-running the original authors' code, we identified and corrected the scaling errors. Team 21 used tobit, with the outcome as red cards / games played, setting tobit limits to censor the outcome at 0 and 1. This led to a large rescaling of the estimate and its standard error; we corrected this using fractional regression and replicated the same result using OLS (an odds ratio of 1.31). Team 27 interacted skin tone with all other controls, but still reported the coefficient on skin tone as a main effect. Since none of the interactions were significant, we simply re-ran their R code without the interaction terms – which rescaled both the coefficient and its standard error, without affecting the significance level (odds ratio of 1.29).

functional form multiverse is missing the left tail of the many analysts results – and ultimately appears as a conservative estimator of what results can be found in repeated modeling. This is reassuring for baseline validity of the computational multiverse: nothing entered this model space that a many-analysts group was not willing to consider. At the same time, this suggests that a truly rigorous multiverse needs to go further still, and incorporate more inputs into the modeling process – especially as data processing decisions like variable coding, treatment of outliers, and the handling of missing data (Steegen et al 2016).

4. Discussion

This study developed a multiverse estimator that models all plausible combinations of specified functional forms and control variables. We then apply this method to a series of empirical cases: the effect of unemployment on subjective wellbeing, and role of education in voting for Trump in 2016, and effect of skin tone on assigning red cards in professional soccer. Across these cases, we compute over 7,500 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. Some functional forms – such as logit and probit – report coefficients on non-comparable scales. Reporting estimates on a common scale (converting either to marginal effects or odds ratios) greatly clarifies what difference there is between LPM, logit, probit, Poisson, and matching estimates. Once the estimates from different models are on comparable scales, we can evaluate the modeling distribution analogously to the sampling distribution. We arrive at two (local) empirical conclusions, one for the mean and one for the variance of the modeling distributions we observe.

First, 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 linear regression, logit, probit, inverse probability weighting, Poisson, 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. On average, it may not matter much, for example, whether researchers use logit or the linear model as long as effect sizes are reported on a common scale (Breen et al. 2018).

Second, different functional forms can also affect the stability or *variance* of the modeling distribution, such as when an estimator is prone to more long-tailed estimates in repeated modeling. When functional forms have the same central tendency, scholars should favor those with the smallest modeling variance over those that give more dispersed and scattered estimates. This is analogous to the efficiency criteria in repeated sampling, and minimizes researcher degrees of freedom in repeated modeling. None of the functional forms considered here improve on linear regression: alternatives like logit, Poisson, inverse probability weighting, and matching could at best match the stability of OLS, not improve it. In Application II, coarsened exact matching had longer *left* tail estimates than LPM or logit, while propensity score matching had longer *right* tail estimates than the others. In application I using panel data, 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 linear regression estimate. Moreover, different functional forms can sometimes adjust for controls in different ways, creating haphazard results for certain combinations of controls and functional forms. This is troubling, given that matching in the past has been advocated as a way to reduce model dependence and improve estimate stability. In our view, these findings support the use of the linear model as a general workhorse, and as the default functional form from which alternatives should be justified in empirical practice.

Finally, we find that computational multiverse analysis does not over-represent modeling uncertainty in applied practice. Compared to a many-analyst crowdsourcing study, working with the same set of possible controls and functional forms, human analysts in fact produced a wider range of estimates. While the many-analysts and multiverse approaches converged on the same average estimate, many-analysts found many more left-tail and non-significant results than are available from all combinations of the basic modeling ingredients. If social scientists worry that the multiverse algorithm can produce idiosyncratic or hard-to-explain estimates, they should be even more worried about what their colleagues can do. This also calls for more, rather than less, multiverse expansion, especially into areas of data cleaning and processing.

5. 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. We routinely find that modeling variance is greater than the sampling variance – more uncertainty stems from the choice of model than from random sampling of the data. Conventional uncertainty about the sample data – captured by t-statistics, confidence intervals, or sampling standard errors, seems less important than uncertainty due to the multiverse of plausible models.

In the 'crisis of science' today, many recognize that current research practices fail to assure an impartial assessment of the evidence. Analysts are not blinded to how their methods support or disfavor a preferred conclusion, but for readers the modeling process is opaque and difficult 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" (1983:31).

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)? What about models that are less sensitive to outliers than OLS, such as least absolute deviations or quantile regression (Andersen 2008)? Finally, we advocate including data processing – alternative ways of cleaning, coding, and categorizing variables as a further dimension of the multiverse that may be more important than conventionally recognized (Leahey et al 2003; Steegan et al 2016; Young and Holsteen 2017).

The challenge facing social science is not simply that there are many ways of doing a thoughtful analysis. The problem is transparency: showing whether results are robust and compelling, or – as critics sometimes suspect – are instead based on fragile model assumptions that are hard to justify or explain. Multiverse analysis offers a way through the crisis of credibility in social science today: showing the other empirical worlds that different analytical

assumptions can see. Incorporating functional form into a multiverse analysis is a key advancement, and we look forward to further progress in the computational analysis of modeling decisions.

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.

Bartus, Tamas. 2005. "Estimation of marginal effects using margeff." *The Stata Journal*. Vol. 5(3): 309–29.

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. 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.

Christensen, Garret, Jeremy Freese, and Edward Miguel. 2019. *Transparent and Reproducible Social Science: How to do Open Science*. Oakland, CA: UC Press.

Coker, Beau, Cynthia Rudin, and Gary King. 2020. "A Theory of Statistical Inference for Ensuring the Robustness of Scientific Results." Working Paper.

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.

Engzell, Per, Carina Mood, and Jan Jonsson. 2020. "It's All about the Parents: Inequality Transmission across Three Generations in Sweden." *Sociological Science*. Vol. 7: 242-267.

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.

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.

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 Nielsen, Richard. 2019. "Why propensity scores should not be used for matching." *Political Analysis*, *27*(4), 435-454.

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.

Leamer, Edward. 1983. “Let's Take the Con Out of Econometrics.” *American Economic Review*. Vol. 73(1):31-43.

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.

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, Stephen. 2018. "Status Threat, Material Interests, and the 2016 Presidential Vote." *Socius: Sociological Research for a Dynamic World* 4:1-17.

Morgan, Stephen, 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.

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

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/

Raftery, Adrian. 1995. “Bayesian Model Selection in Social Research.” Sociological Methodology 25:111-65.

Ripollone, John, Krista Huybrechts, Kenneth Rothman, Ryan Ferguson, Jessica Franklin. 2020. “Evaluating the Utility of Coarsened Exact Matching for Pharmacoepidemiology Using

Real and Simulated Claims Data." American Journal of Epidemiology. Vol. 189(6): 613–622.

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.

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. 2007. "Inverse Probability Weighted Estimation for General Missing Data Problems." *Journal of Econometrics*. Vol. 141(2): 1281–1301.

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.

Table 1. Functional Form Robustness of Effect of Job Loss on Wellbeing

	Average estimate	*Range*	*Average sample SE*	*Modeling SE*	*Significance rate ($p < 0.05$)*
OLS	-0.44	[-0.45, -0.42]	0.08	0.01	100.0%
IPW	-0.46	[-0.64, -0.32]	0.12	0.07	100.0%
CEM	-0.40	[-0.52, -0.15]	0.09	0.08	98.5%
PSM	-0.39	[-0.90, -0.12]	0.13	0.09	86.9%
Pooled distribution	-0.42	[-0.90, -0.12]	0.11	0.07	96.4%

Note: Estimates from 4,090 model specifications. Data: PSID (2003-2005 waves). N = 6,192.

Table 2. 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%

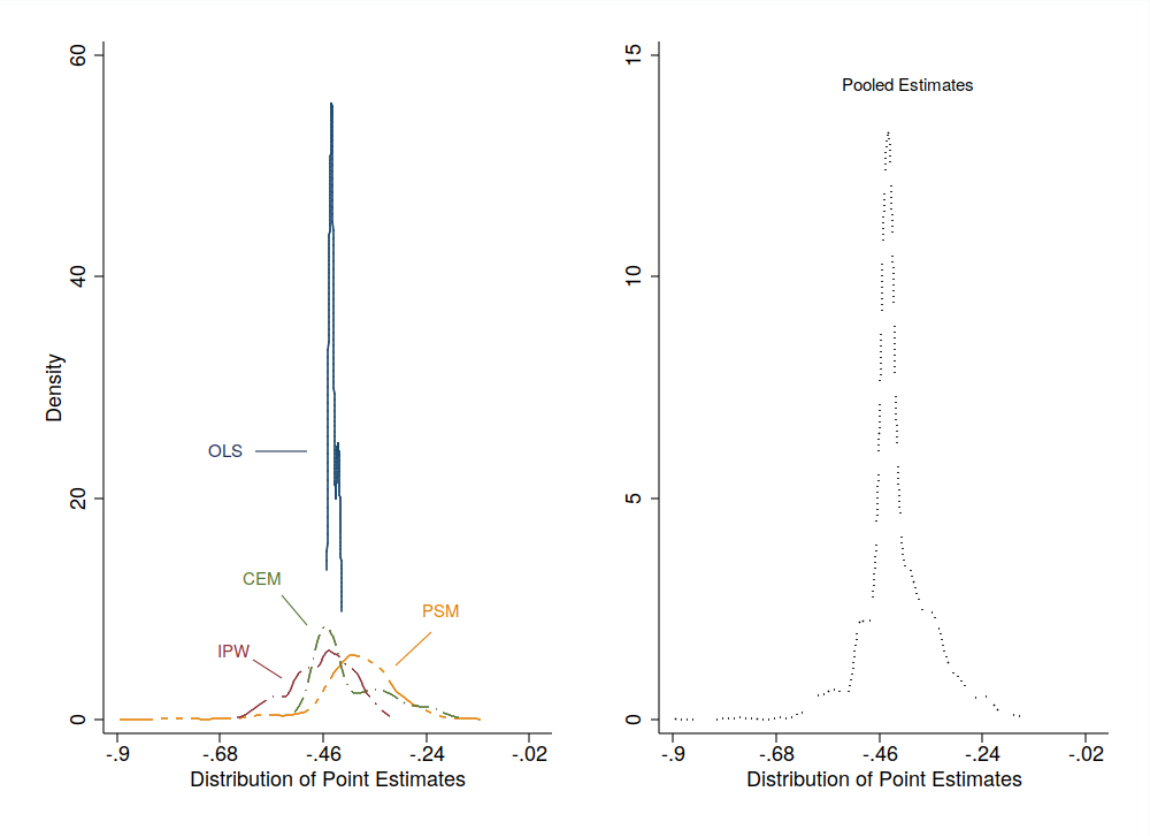
Note: Estimates from 510 model specifications. Data: ANES 2016 Time Series. N = 1,701.

Table 3. Functional Form Robustness of Effect of Skin Tone on Receiving Red Cards

	Average estimate	*Range*	*Average sample SE*	*Modeling SE*	*Significance rate ($p < 0.05$)*
Logit (OR)	1.30	[1.23,1.38]	0.110	0.035	100.0%
OLS (OR)	1.30	[1.23,1.36]	na	0.034	100.0%
Poisson (IRR)	1.34	[1.23, 1.64]	0.119	0.059	100.0%
Pooled distribution	1.31	[1.23, 1.64]	na	0.048	100.0%
"Many Analysts"	1.27	[0.88, 1.71]	na	0.170	69.0%

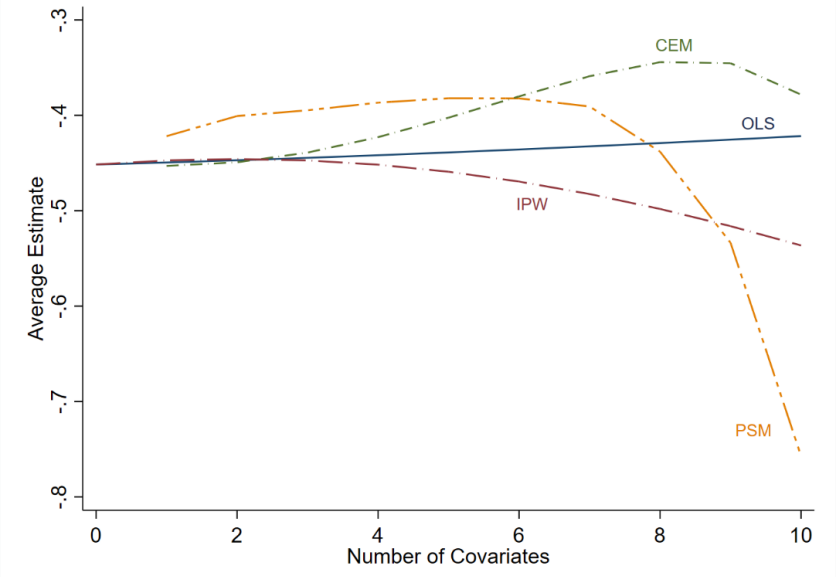
Note: Estimates from 3,072 model specifications: 3 functional forms each with 10 possible controls. Data: Silberzahn et al 2018. N ranges from 350,448 to 373,067, depending on selection of controls. Sampling standard errors were not computed for OLS odd ratio results, and were not directly reported in "many analysts" study.

Figure 1. Modeling Distributions: Effect of Job Loss on Subjective Wellbeing



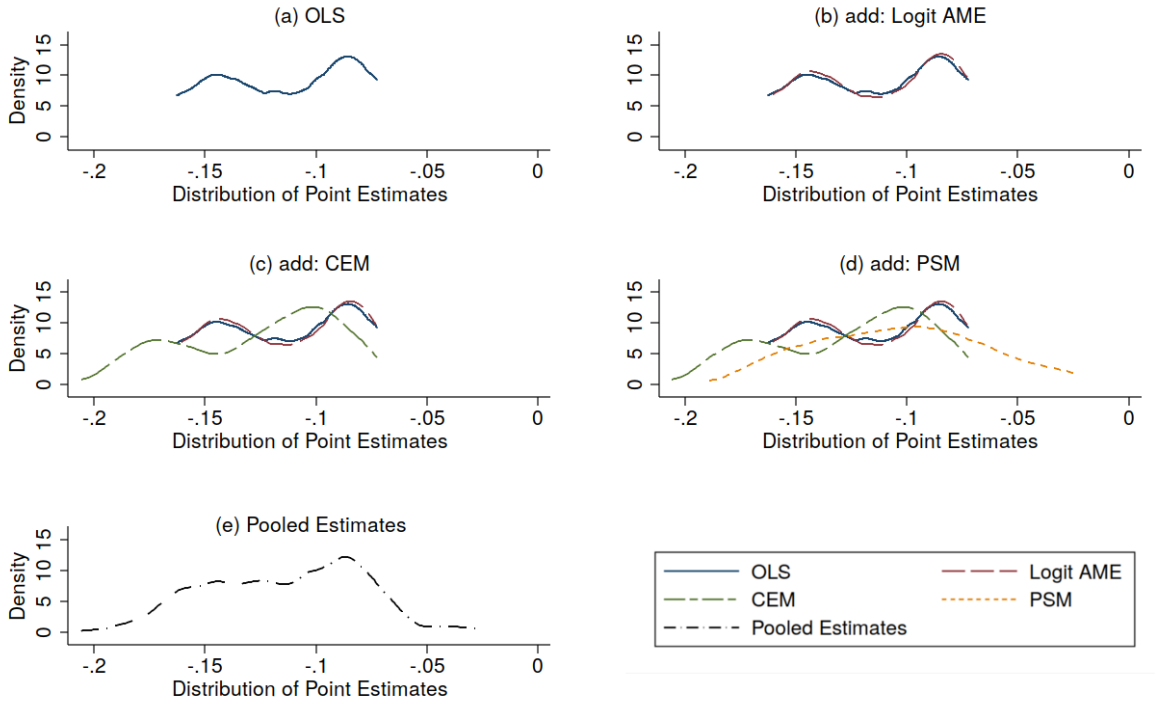
Note: Estimates from 4,090 model specifications. Data: PSID (2003-2005 waves). N = 6,192.

Figure 2: Effect Size of Job Loss by Number of Control Variables



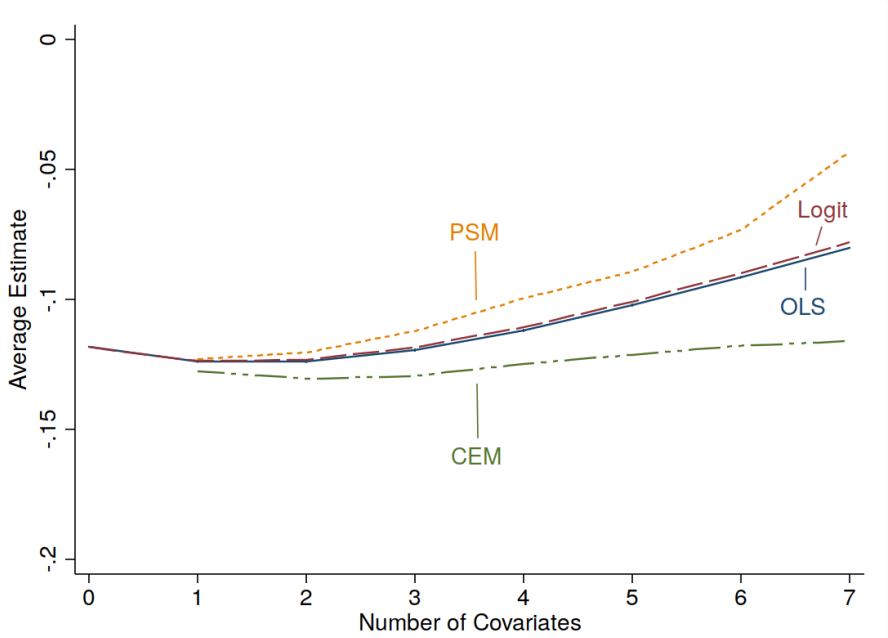
Note: based on results from 4,090 model specifications. Data: PSID 2003-05, N =6,192.

Figure 3. Modeling Distributions: Effect of College Degree on Voting Donald Trump in 2016



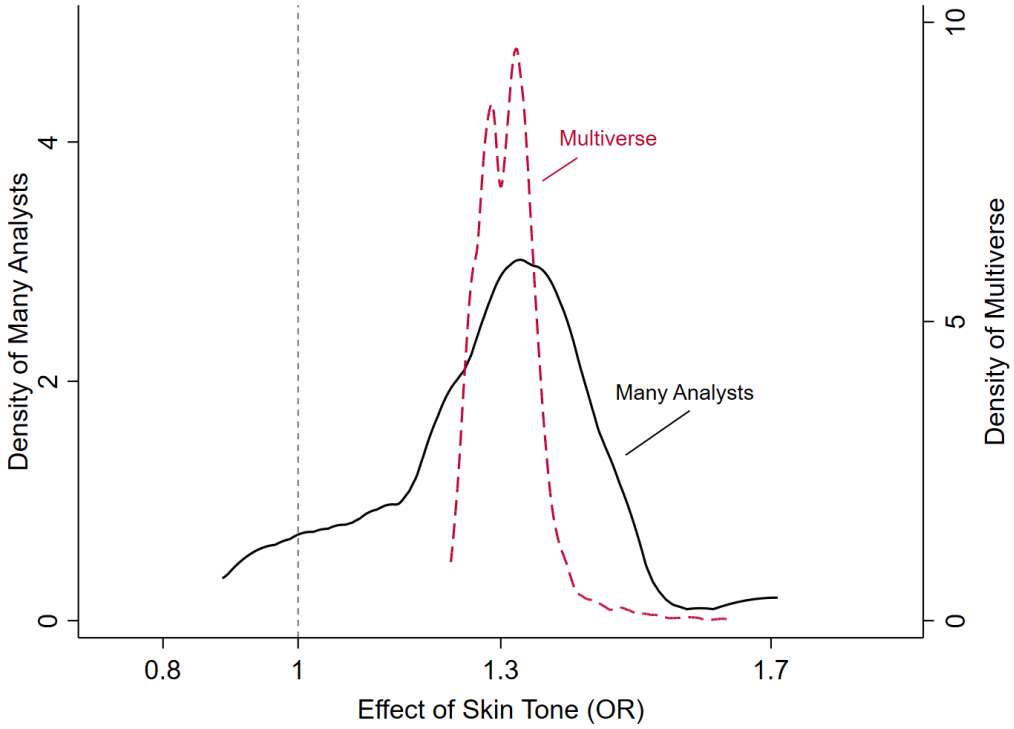
Note: Estimates from 510 model specifications. Data: ANES 2016 Time Series. N = 1,701.

Figure 4: Effect Size by Number of Control Variables. Voting Data



Note: based on results from 510 model specifications. Data: ANES 2016 (N = 1,701).

Figure 5. Multiverse Results: Effect of Skin Tone on Red Cards



Note. Multiverse: Odds ratio estimates from 3,072 model specifications. Many Analysts: 29 estimates reported in Silberzahn et al (2018). Two corrections noted in footnote 8. Data: Silberzahn et al 2018. N ranges from 350,448 to 373,067, depending on selection of controls.

Appendix I

Table 1 illustrates the problem of scale across functional forms. Consider an analysis of voting for Donald Trump in the 2016 presidential election as a function of voters' race (coded as white or 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 estimate from the linear model, 0.10, meaning that whites are 10 percentage points more likely to vote Trump than nonwhites (from an overall sample mean of 46 percent). Are all these estimates saying the same thing, or are they giving different results?

Table A1. Comparison of Coefficient Magnitudes

	White (vs. nonwhite)
Logit (odds ratio)	3.30
Probit	0.64
LPM	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 A1, we convert the logit odds-ratios and probit coefficients onto the probability scale as average marginal effects.[9] Once these estimates are on the same scale as LPM, we see that the logit and probit results are much the same as the estimates from the linear model. 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. If logit coefficients were routinely converted to marginal effects and compared to LPM (or vice versa), how often would the differences be large enough to matter?

[9] The average marginal effect is a post-estimation procedure that computes the expected difference in outcome probability due to a unit change 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. Next, the outcome probabilities are predicted 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 directly compared to the OLS coefficient (Williams 2011, see also Bartus 2005). Note that this procedure linearizes the effect of x on y (Mood 2010).

Appendix II. 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 (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 each of the control variables included in each of the models. For *n* possible control variables, we create a set of dummy variables $\{D_1 \ldots D_n\}$ 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 that 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: Unemployment and Wellbeing

In application I, the influence of many control variables is different across functional forms. 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 nearly fifteen 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), modest 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 A1 gives a broader view of these idiosyncratic influence effects. In OLS, the number of controls has no noticeable 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 unless there are many of them, in which case adding controls greatly increases the estimate (i.e., 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 A2: 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	

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 LPM, 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 A3: Influence Effects for Effect of College Degree on Voting for Trump

	LPM		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 LPM 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.