Designing Surveys to Account for Endogenous Non-Response

Michael A. Bailey*
Georgetown University
Michael.Bailey@georgetown.edu

June 2018

Abstract

Non-response is a large and growing problem in survey research. Weighting can address non-response associated with observable variables, but cannot solve - and may exacerbate - non-response bias associated with unmeasured factors. Selection models can correct for non-response related to both measured and unmeasured factors, but prove either unwieldy or impossible for most conventional survey data. This paper argues that surveys should be designed to provide the information needed to make selection models function properly. In particular, this paper focuses on two tools that enable survey data to be used to assess selection on unmeasured factors. First, surveys can include questions that elicit willingness to respond independent of content of response. Second, by randomly treating some potential respondents with opt-in questions, we produce a variable that explains response, but does not affect outcome variables directly. Taken together, these tools allow us to easily assess weighting models' assumption that willingness to respond is unrelated to opinions. Two empirical applications demonstrate the potential for non-response bias to exaggerate polarization and turnout.

*Prepared for the International Total Survey Error Workshop, June 4 - 6 in Durham, North Carolina. I am grateful for helpful conversations with Mike Alvarez, Adam Berinsky, Mike Hanmer, Erin Hartman, Dan Hopkins, Jon Ladd, John Lapinski, Lilly Mason, Marc Meredith, Mike Miller, Hans Noel, Ellie Powell and talks at the 2017 Political Methodology Meetings at the University of Wisconsin Madison, Georgetown University, the University of Pennsylvania and the University of Maryland. All errors are mine.

Understanding how public opinion polls work (and fail) in the modern polling environment is a foundational issue for the study of public opinion. Clearly, there is much to learn. Few, if any, polls suggested that Donald Trump would carry as many Midwestern states as he did in 2016. Polls also did not foresee the Brexit victory in June 2016, Benjamin Netanyahu's victory in Israel in March 2015, David Cameron's victory in the U.K. in May 2015, Matt Bevin's victory in the Kentucky gubernatorial race in November 2015, among other misses.

Many suspect that non-response bias is an important factor in these polling mishaps. Over the last ten years response rates in the U.S. have plummeted and now are under ten percent for landlines and under eight percent for cell phones (Dutwin and Lavrakas 2016; Pew Research Center 2012). Academic surveys are not immune from declining response rates; some important academic polls have even abandoned random sampling, at least as conventionally understood. Potential biases that emerge in such contexts may be less public than for election polls, but are highly troubling nonetheless.

The conventional way to address non-response is via weighting, which produces an effective sample that reflects the target population with respect to selected measurable attributes. Weighting comes with a substantial drawback however: it fails to correct for non-response associated with unmeasured attributes. That is, weighting fails if the propensity to respond to a survey is endogenous (or, non-ignorable), meaning that non-response is related to the content of opinions after controlling for measured variables.

These limits of weighting are widely recognized (see, e.g., Peress 2010). They are even more widely ignored. Survey researchers using weights rarely diagnose whether the conditions necessary for weighting to be useful are satisfied (Franco, Malhotra, Simonovits and Zigerell 2015). One reason why pollsters seldom test for endogenous selection is that selection models are so demanding of data that they are often unusably low-powered and unreliable for survey data.

This paper presents a two-fold strategy for designing surveys so that they produce the kind of data needed to identify endogenous selection. First, surveys can include questions that elicit respondents' propensity to discuss politics independent of their opinions about politics. This information can be used

to directly test weighting models assumption that non-response is ignorable conditional on covariates. Second, pollsters can randomly assign respondents to conditions that affect the probability of response, but do not affect the content of opinions. This can be done in many ways, but it is particularly easy to implement with randomized treatments that *inhibit* response. The randomization produces a variable that predicts response while not directly related to the opinion being measured.

While we must be realistic about how much we learn about people who never respond to surveys, these tools give us a much stronger basis for assessing the strong assumptions underlying weighting and related approaches or for analyzing data with models that allow for endogenous selection.

This paper provides results from two surveys that use these tools. These examples illustrate that the selection-sensitive survey design approach is simple, allowing us to assess endogenous selection in a context where conventional surveys would fail. The results also indicate that endogenous selection was an issue in both surveys. The survey designed identified clear non-response bias for turnout intention, a result that is consistent with previous work. The survey design also identified signs of severe selection bias among partisan subsamples. For example, the gap between Democrats and Republicans on feeling thermometers toward President Obama was 20 points higher among respondents with a high propensity to respond.

This paper proceeds as follows. Part 1 discusses weighting and selection models as distinct approaches to dealing with non-response. Part 2 presents survey design tools that survey researchers can use to confront these challenges. Part 3 presents simulation results that demonstrate how these design tools operate in theory. Part 4 shows how these design tools operate in practice by discussing results from two surveys utilizing these methods.

1 Weighting and Selection Models

Non-response is a large and growing problem in contemporary survey research. Figure 1 shows the non-response for large polling firms since 1998 from Dutwin and Lavrakas (2016) and Pew Research Center (2012). Scholars in the late nineties were already very concerned about non-response when 64

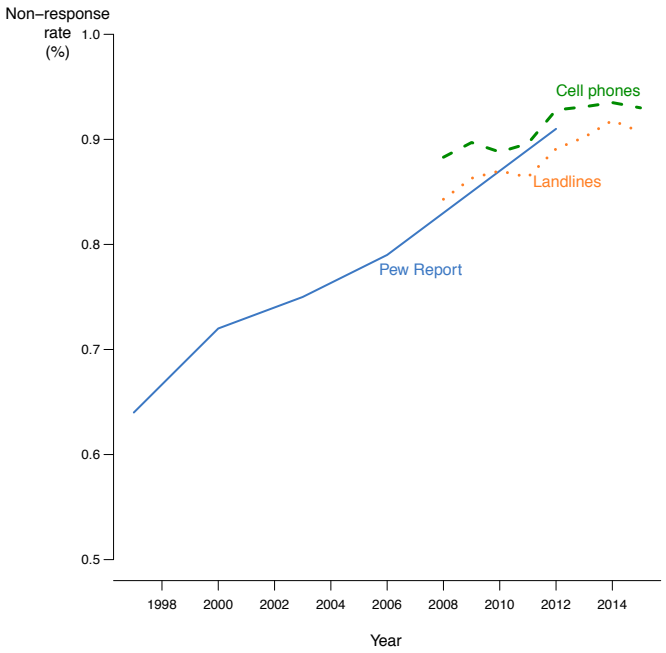

Figure 1: Non-response rates in political surveys

percent of those selected to be interviewed did not complete surveys. Things only got worse and now non-response rates above 90 percent are common. Simply by answering questions, survey respondents are indicating to us that they are willing to do something that the overwhelming majority of Americans are not willing to do: respond to pollsters.

The dangers of non-response are clear: the types of people who respond to surveys may systematically differ from others, yielding inaccurate descriptive and correlational statistics. This section provides a brief overview of the state of the non-response literature by contrasting weighting and selection models. Weighting models are generally feasible with data available to pollsters, but fail when non-response is affected by unmeasured factors that also affect opinion. Selection models can deal with non-response due to unmeasured, yet relevant, factors, but are often infeasible for conventional survey data.

We model survey response with the following two equations. The outcome of interest, Y, is the survey responses to a particular question. We model it as a function of covariates

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad (1)$$

where β_1 is a $1 \times p$ vector and X_i is a $p \times 1$ vector and ϵ is a mean-zero random variable uncorrelated with X. For simplicity, we do not at this point account for non-linearities and interactions.

We observe $Y_i|_{R_i=1}$, where R_i is an indicator variable equalling 1 for individuals who respond to the survey and 0 for those who do not. We model response in terms of R_i^*, which is the latent propensity to respond:

$$R_i^* = \gamma_0 + \gamma_1 Z_i + \tau_i \quad (2)$$

where γ_1 is a $1 \times k$ vector, Z_i is a $k \times 1$ vector and τ is a mean-zero random variable uncorrelated with Z.

Two mechanisms connect Equations 1 and 2. First, the covariates in Equation 1, the outcome equation, may contain (and possibly be the same as) the covariates in Equation 2, the selection equation. Second, the errors in the two equations may be correlated; we denote this correlation with ρ where $-1 < \rho < 1$.

Weighting is the most common approach to dealing with survey non-response. There are many ways to implement weighting models; here we focus on inverse propensity weighting models. In inverse propensity weighting models Equation 2 is used to generate a predicted probability of response, $\hat{p}_i$. Observations in the outcome equation are then weighted by $\frac{1}{\hat{p}_i}$ such that observations from people underrepresented in the survey sample (who have a low probability of response, $\hat{p}_i$), get high weights and observations from people overrepresented in the sample get lower weights.[1] If the assumptions underlying weighting models are correct, the weighted means in the sample for all independent variables will align with the underlying population means.

Weights are a staple in survey research. Virtually every commercial poll uses weights. Academic surveys such as the American National Election Study, the General Social Survey and the Congressional Cooperative Election Study provide weights and advise end-users to use them.

Scholars vary in their use of weights however. Franco, Malhotra, Simonovits and Zigerell (2015) assessed survey experiments in leading political science journals from 2000 to 2015 and found that 78 percent of the papers did not even report whether they used weights. Part of the variation is due to confusion about the purpose of weighting (Solon, Haider and Wooldridge 2013).[2]

[1] On the latest tools to create optimal weights, see Caughey and Hartman (2017).

[2] Note, for example, the tension between survey weighting and weighted least squares (WLS). The methods both involve weighting individual observations, yet are motivated quite differently. In WLS, the observations about which we

The key assumption underpinning weighting models is that the decision of individuals to respond to a poll is, conditional on covariates, unrelated to the content of the survey responses. Depending on the literature, this assumption is stated in different, yet essentially equivalent, terms. In the causal modeling literature, the conditional independence of propensity to respond and content of response occurs when non-response is *ignorable*; the condition is violated when non-response is *non-ignorable*. In the selection literature, this conditional independence is stated in terms of the correlation of the error terms in Equations 1 and 2. If τ, the error in the response equation, is uncorrelated with ϵ, the error in the outcome equation, then response is exogenous and weighting is appropriate. If these errors are correlated, response is endogenous and weighting (and OLS) will be subject to non-response bias.

Endogenous selection is a reasonable concern for many political questions. Consider a case in which white working-class men under the age of 30 are underrepresented in a survey sample, a common state of affairs. Suppose that based on their population, we would have expected to get ten such men in our sample, but only got five. A standard weighting scheme would double weight these five men such that they would effectively be ten men in the sample.

Weighting produces misleading results if our five respondents have systematically different opinions (the outcome of interest) than their demographic peers who did not respond. In the political realm, it is plausible that the five respondents were more politically engaged and that politically engaged young white working-class men (for example) have different political views than their less politically engaged peers. In such a situation, the error in the response equation will be correlated with the error in the outcome equation. The sample will suffer from non-response bias and the weighted sample may be *less* representative than an unweighted sample because it places more weight on the five unrepresentative young men.

One way to alleviate bias in weighting models is to have highly predictive covariates. Some panel-based surveys have historical data on respondents and non-respondents, enabling weighting based on a highly relevant covariate such as vote choice in a previous election (Lauderdale and Rivers 2016). Ho-

are most certain get the highest weight; in survey weighting, the observations from groups with the lowest response rates get the highest weight.

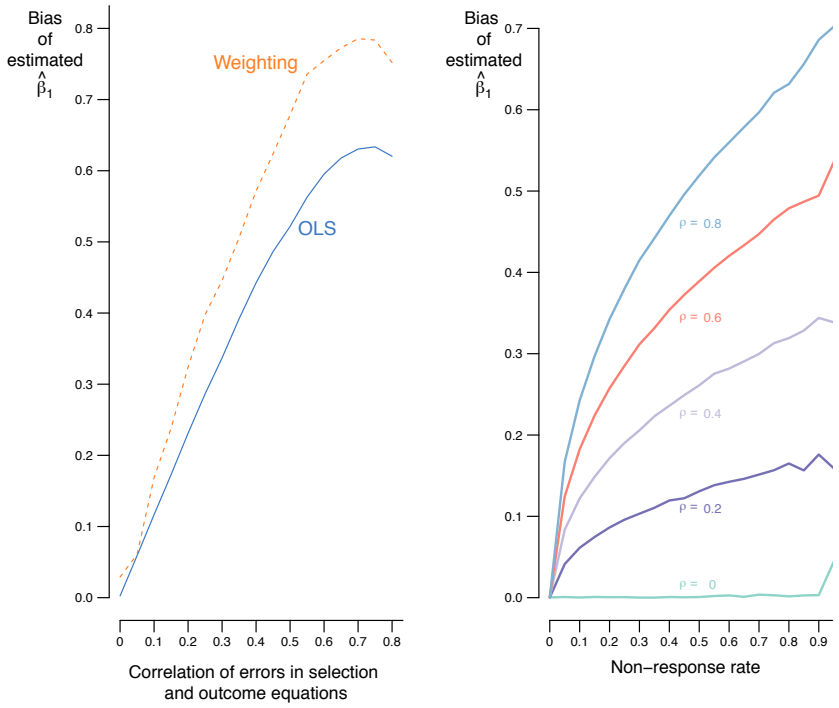

Figure 2: Bias in Weighting and OLS models

wever, even with such a variable, endogenous selection can render weighting problematic. For example, if we weight based on 2012 vote choice when analyzing 2016 polls, we can take into account some of the trends that may have led Obama 2012 voters to be more or less likely to respond to polls depending on what was happening in the campaign. However, such an approach assumes that the Obama 2012 voters who responded in 2016 were representative of all 2012 Obama voters. As with our demographic-based example, it is plausible that there was systematic difference in Obama 2012 voters interested in talking to pollsters compared to such voters who did not respond. For example, if white, working-class Obama 2012 supporters became disillusioned with politics and responded less, we would end up generalizing to the broader population of such voters based on a skewed subset of them.

Two factors produce rising bias in weighting approaches. The first is that the bias for weighting (and OLS) gets worse as the correlation between the errors of the response and outcome equations increases. The left panel of Figure 2 shows results for a simulation based on Equations 1 and 2. As the correlation of error increases, both OLS and weighted models produce more biased estimates of β_1, with the bias being worse in weighted models.[3]

[3] Errors distributed according to a joint normal distribution in the simulations. Non-response bias extends beyond this specific distributional assumption.

The second factor that affects bias is the degree of non-response. Obviously, non-response bias requires non-response. As the extent of non-response increases, non-response bias rises. The panel on the right in Figure 2 shows bias as a function of non-response rate. Each line corresponds to a specific value of the correlation of errors in the response and outcome equations. For the line at the bottom, there is no correlation and, as we have seen above, there is no bias. For non-zero correlations, however, the bias is increasing in non-response rates. Consider the line reflecting bias when ρ, the correlation of errors, is 0.4. At a non-response rate of 0.3, the bias is around 0.2; by the time the non-response rate reaches 0.9, the bias reaches 0.34. Such patterns are visible for all non-zero correlations of the errors. Given the rapid increase in non-response in Figure 1, Figure 2 suggests that we should be more vigilant than ever about non-response.[4]

Selection models offer an alternative way to deal with non-response. The expected value of for observations within our sample is

$$E[Y_i|_{Y_i \text{ observed}}] = \beta_0 + \beta_1 X_i + E[\epsilon_i|_{R_i=1}] \quad (3)$$

$$= \beta_0 + \beta_1 X_i + E[\epsilon_i|_{\tau_i > -\gamma_0 - \gamma_1 Z_i}] \quad (4)$$

If non-response is exogenous/ignorable, then $\rho = 0$ and the $E[\epsilon_i|_{\tau_i > -\gamma_0 - \gamma_1 Z_i}] = E[\epsilon_i] = 0$ and OLS and weighted models will be unbiased. If non-response is endogenous/non-ignorable, then the expected value of ϵ, the error term in the outcome equation, will be related to the value of τ, the error term in the response equation, making $E[\epsilon_i|_{\tau_i > -\gamma_0 - \gamma_1 Z_i}] \neq 0$ and potentially correlated with X.

Selection models provide a mechanism to incorporate our intuition about non-response into our statistical analysis. Suppose for simplicity that the error terms in each equation reflect only trust in the media, an unmeasured factor that increases both the propensity to respond and the value of Y. The expected value of the error term in the outcome equation will be greater than zero because respondents are more trusting of the media than non-respondents. In other words, non-response would lead us to

[4] Noting that non-response bias rises with non-response does not contradict evidence in Groves and Peytcheva (2008) that the degree of non-response bias is unrelated to the magnitude of non-response. They are referring to an analysis of a cross-section of surveys. For some surveys, there was bias with low non-response (corresponding perhaps to a high ρ and low non-response rate in Figure 2) and in other surveys there was no bias with a high non-response rate (corresponding, for example, to a low ρ and high non-response rate situation).

overestimate the value of Y in the population.[5]

Selection models offer a variety of approaches to dealing with $E[\epsilon_i|_{R_i=1}]$. Heckman (1979) presents a canonical approach in which he assumes that τ and ϵ are distributed bivariate normally with correlation ρ. In this case,

$$\begin{aligned} E[Y_i|_{Y_i \text{ observed}}] &= \beta_0 + \beta_1 X_i + E[\epsilon_i|_{R_i=1}] \\ &= \beta_0 + \beta_1 X_i + E[\epsilon_i|_{\tau_i > -\gamma_0 - \gamma_1 Z_i}] \\ &= \beta_0 + \beta_1 X_i + \rho\sigma_\epsilon \frac{\phi(-\gamma_0 - \gamma_1 Z_i)}{1 - \Phi(-\gamma_0 - \gamma_1 Z_i)} \\ &= \beta_0 + \beta_1 X_i + \rho\sigma_\epsilon \frac{\phi(\gamma_0 + \gamma_1 Z_i)}{\Phi(\gamma_0 + \gamma_1 Z_i)} \\ &= \beta_0 + \beta_1 X_i + \lambda M_i \end{aligned} \tag{5}$$

where $\lambda = \rho\sigma_\epsilon$, σ_ϵ is the variance of ϵ and M_i is the inverse Mill's ratio which is $\frac{\phi(\gamma_0+\gamma_1 Z_i)}{\Phi(\gamma_0+\gamma_1 Z_i)}$. The function in the numerator, $\phi()$, is the normal probability density function. The function in the numerator, $\Phi()$, is the normal cumulative density function, which is equivalent to the fitted probability of response from a first stage probit model of response.

Achen (1986) uses a linear probability model to estimate the response equation, an approach that requires stronger parametric assumptions than Heckman's model, but is easier to work with. Wooldridge (2002, 563) shows that the Heckman model works with weaker assumptions: τ is normally distributed and $E[\epsilon|\tau] = \delta\tau$. Das, Newey and Vella (2003) provide a more general formulation that does not require the bivariate normality assumption:

$$E[Y_i|_{Y_i \text{ observed}}] = \beta_0 + \beta_1 X_i + \gamma_1 p_i + \gamma_2 p_i^2 + ... + \gamma_k p_i^k \tag{6}$$

Selection models use very similar information as weighting models, just in different ways. Inverse-propensity weighting models divide all variables by p_i, the probability of response for an individual;

[5] Endogenous selection can also bias regression coefficients in a model that ignores selection. Coefficients are biased if the omitted variable, $E[\epsilon_i|_{R_i=1}]$, is correlated with X_i. Suppose the observed variable, X, is education and that more educated people are more likely to respond. This means that people with low levels of education have to have particularly high values of τ to respond. In other words, low education respondents in the sample will on average be more trusting of the media than highly educated respondents, inducing a negative correlation between education and the level of trust. Sartori (2003) formalizes a model in which the errors in the two equations are assumed to be identical.

selection models include a function of p_i as a covariate. This is directly clear in the Das, Newey and Vella model; for the Heckman model, note that the inverse Mill's ratio can be re-written as $\frac{\phi(\Phi^{-1}(p_i))}{p_i}$.

The different treatment of similar information is consequential. Consider a simple case in which we estimate a population mean with no control variables:

$$Y_i^* \quad = \quad \beta_0 + \epsilon_i \tag{7}$$

Table 1 displays the model, the estimate and the marginal effect of Y_i on the population mean for OLS (as a baseline), weighted regression and as Heckman models (as an example of selection models). (Appendix A shows that the same principles operate in models with control variables.)

OLS is straightforward. The right-hand column in Table 1 shows that a one unit increase in Y_i increases $\hat{\beta}_0$, the estimated mean, by $\frac{1}{N}$.

Table 1: Marginal effects of Y_i on $\hat{\beta}_0$ in different approaches

Approach	Model	Parameter	Estimate	$\frac{\partial Estimate}{\partial Y_i}$
OLS	$Y_i = \beta_0 + \epsilon_i$	β_0	$\hat{\beta}_0 = \frac{\sum Y_i}{N}$	$\frac{1}{N}$
Weighting	$\frac{Y_i}{p_i} = \beta_0 \frac{1}{p_i} + \frac{\epsilon_i}{p_i}$	β_0	$\hat{\beta}_0 = \frac{\sum \frac{Y_i}{p_i^2}}{\sum(\frac{1}{p_i^2})}$	$\frac{1}{p_i^2 \sum(\frac{1}{p_i^2})}$
Heckman	$Y_i = \beta_0 + \lambda M_i + \epsilon_i$	β_0	$\hat{\beta}_0 = \overline{Y} - \overline{M}\hat{\lambda}$	$\frac{1}{N} - \frac{\frac{N-1}{N}(M_i - \overline{M})}{\sum(M_i - \overline{M})^2}$
		λ	$\hat{\lambda}_0 = \frac{\sum(M_i - \overline{M})(Y_i - \overline{Y})}{\sum(M_i - \overline{M})^2}$	$\frac{\frac{N-1}{N}(M_i - \overline{M})}{\sum(M_i - \overline{M})^2}$

The effect of a single observation Y_i on the WLS estimate of the population mean is more involved, but intuitive. The square of the probability of observation, p_i, is in the denominator, implying that low probability observations have much more influence than high probability observations. The effect of Y_i on the WLS estimate of $\hat{\beta}$ is greater than the effect of Y_i on the OLS estimate of $\hat{\beta}_0$ as long as $\frac{1}{p_i^2}$ is greater than the average of all $\frac{1}{p_i^2}$, something that happens for small p_i values. This is intuitive as the point of weighting in this context is to give more weight to low probability observations.

The effects of a single observation Y_i on parameters in the Heckman model are counterintuitive. The effect of Y_i on $\hat{\beta}_0$ is the effect in OLS ($\frac{1}{N}$) minus the marginal effect of Y_i on $\hat{\lambda}$. The effect of Y_i

on $\hat{\lambda}$ depends on p_i. If p_i is low, then $M_i > \overline{M}$ (because p_i is in the denominator of M_i) and the effect of increasing Y_i on $\hat{\lambda}$ is positive, meaning the effect of Y_i on $\hat{\beta}_0$ is *less than* the effect in OLS. In other words, low probability observations have a smaller effect on the estimated mean than they do in OLS (and, therefore, a smaller effect than in WLS models).

A one-unit increase in Y_i actually can *lower* the Heckman estimate of the population mean value. Figure 3 shows a stylized example of a scatterplot of Y_i and M_i values for a Heckman model. The solid line indicates the fitted line from a Heckman model for the five observations. The intercept is 0.95; this is the estimated mean for the population, $\hat{\beta}_0^{(1)}$. The slope is the estimated normalized correlation between the errors in the selection and outcome equations for the five observations, $\hat{\lambda}^{(1)}$. If we increase the value of Y_i by one for the observation with the highest value of the inverse Mill's ratio (a low probability observation given the definition of the inverse Mill's ratio), the estimated line will become steeper. This means that the estimated correlation of errors ($\hat{\lambda}^{(2)}$) is higher and the estimated mean for the population ($\hat{\beta}_0^{(2)}$) is lower. In other words, a higher value of Y_i for a low probability observation leads the Heckman model to estimate that the overall mean value of Y in the population is *lower*.[6]

This can happen because a high value of Y_i for a low probability observation is evidence of a correlation error in the outcome equation and in the selection equation, potentially pushing the Heckman model to estimate a higher value of $\hat{\lambda}$ which, in turn, pushes down the estimate of $\hat{\beta}_0$.

The key distinction between the Heckman and weighting models is that data in the Heckman model is simultaneously informative about the correlation of errors and about the relationship between the independent and dependent variables. In some cases, a low probability observation will indicate that there is correlation in the error terms, rather than indicate the nature of the relationship between X and Y. In weighted models, in contrast, low probability observations are always taken to be highly and solely informative about the relationship between the independent and dependent variables.

Selection models therefore present a theoretically appealing, and distinct, alternative to weighting models. There is a catch however: these models do "not always give sensible answers and [are] now

[6] If we increase the value of Y_i for a high probability (and, therefore, low M_i) observation, the estimate of the population mean will rise and the estimate of the correlation of errors will fall. Appendix A shows the relative influence of observations in the OLS, WLS and Heckman models for a model with a covariate.

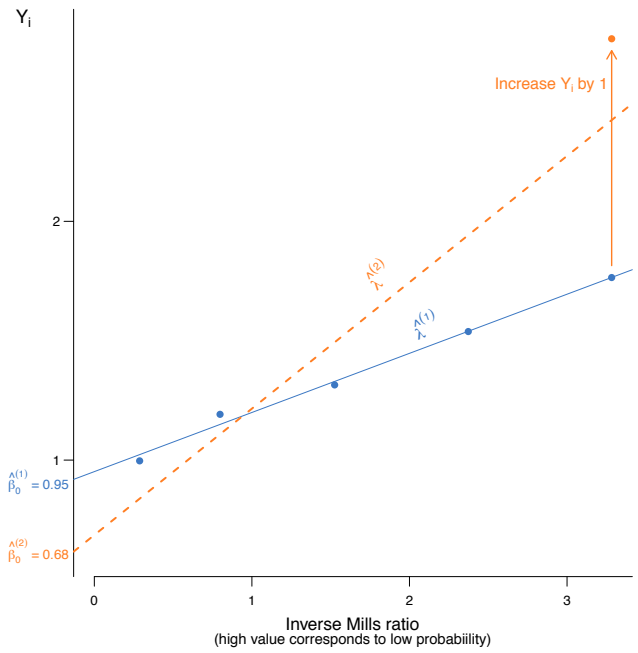

Figure 3: Effect of increasing Y_i for low probability observation on fitted line in Heckman model

no longer regarded as the panacea for all data selection problems" (Copas and Li 1997, 59). One problem is that we may lack data on the non-respondents. A more vexing problem is that the non-linearity of the inverse Mills ratio notwithstanding, it is very common to see extraordinary high levels of multicollinearity between the inverse Mills ratio and the other covariates. This is especially true when the same variables are used in the response and outcome equations, as is common in survey research (Bushway, Johnson, and Slocum 2007; Puhani 2000). Appendix B elaborates on the sources of this problem. Later, we will show examples in which Heckman models applied to conventional survey data fail. Section 3 shows this for simulations and Section 4 shows this for real data sets.

2 Selection Sensitive Survey Design

What, then, is a survey researcher supposed to do in the face of widespread non-response? There are, it seems, three unappealing alternatives: ignore the problem, assume ignorable non-response and weight data or implement selection models that perform poorly (if at all) on typical survey data.

This section presents an alternative approach, based on the idea that "design trumps analysis"

(Rubin 2008). The goal is to re-design surveys so that they produce the kind of data that will allow selection models to perform adequately. I focus on two strategies in particular. The first is to include questions that elicit information about the propensity to respond independent of the content of response. Doing so will allow us to directly assess whether the relationship between response propensity and opinion; given a parametric model such as Peress (2010) we can also estimate non-response bias. The second strategy is to incorporate into the survey a randomized treatment that affects response propensity, but not the content of opinions. Doing so will give us the statistical power necessary to estimate a selection model.

The specific questions used here are illustrative of the broader, crucial point: we need not – and indeed, should not – be passive in the face of potential endogenous/non-ignorable non-response when we design our surveys. Designing surveys that enable us to address non-ignorable non-response is simple and can produce very useful information about the nature of non-response. We may find that there is no evidence of endogenous selection; or we may (as we do below) find evidence that calls into question weighting as a solution to non-response.

Figure 4 illustrates how we can design our surveys to identify endogenous non-response. I include hypothetical response numbers to provide a sense of how the process works. A sample of 15,000 is drawn from some population and, consistent with recent response rates, 10% of these individuals complete the survey. The 1,500 respondents are randomly assigned to control and treatment groups. The 750 individuals in the control group are asked a set of political questions.[7] The 750 individuals in the treatment group are first asked to choose a category about which to answer questions. For example, they may be asked to choose to provide feedback on individuals associated with politics, sports or movies as shown in Figure 5. The 375 respondents who choose politics (50% of the treatment group in this example) are asked the same political questions as the control group. The 375 individuals who choose a non-political topic are given a series of questions on that topic and then they are asked the political questions.

[7] For now we ignore non-response in the control group.

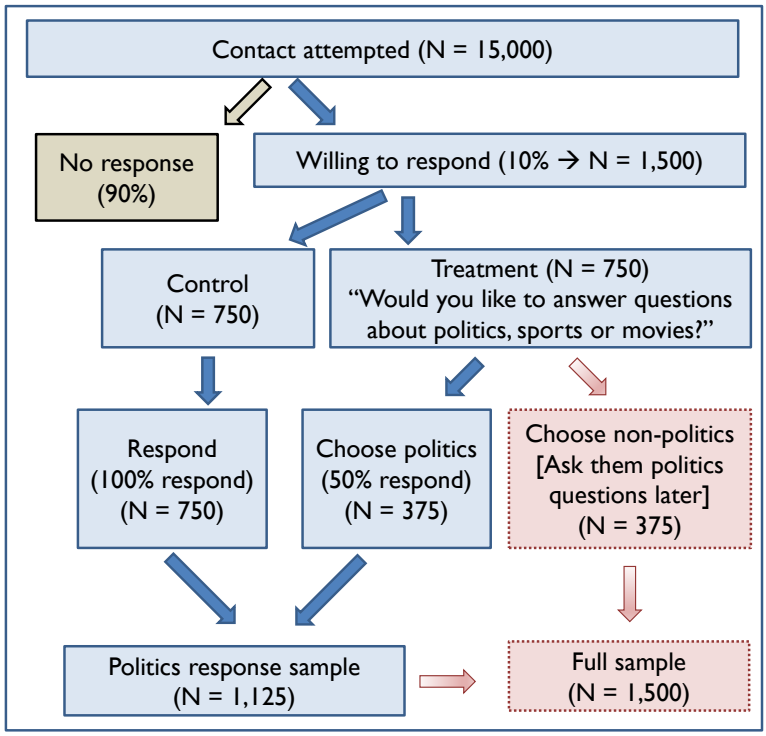

Figure 4: Selection sensitive survey design

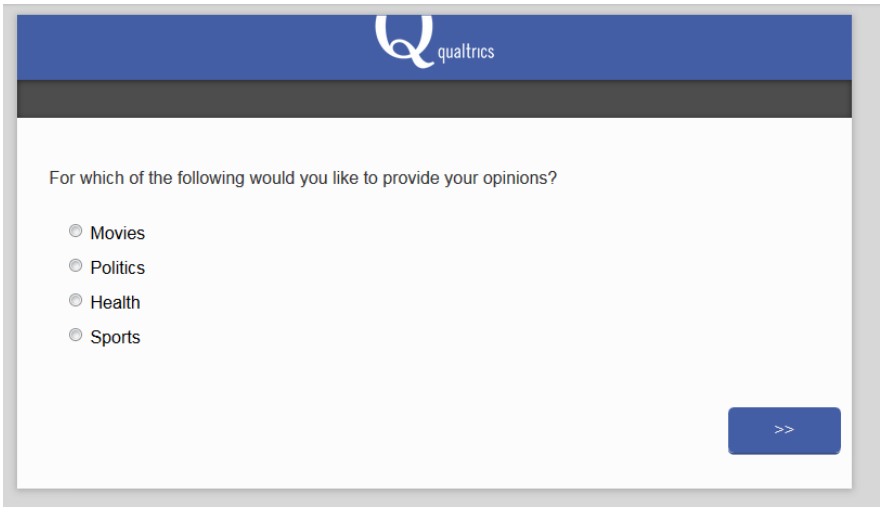

Figure 5: Opt-in question

This design provides several types of data that are very useful for assessing and, if necessary, correcting for endogenous selection. First, we can directly diagnose endogenous selection. If willingness to respond has no effect on Y, then the 375 treated respondents who chose politics should, conditional on covariates, be no different with regard to the outcome variable than the 375 treated respondents who initially chose to answer non-political questions. This can be tested with a simple OLS model.

$$Y_i^* = \beta_0 + \beta_1 \text{Choose politics}_i + \beta_X X_i + \epsilon_i \quad (8)$$

where *Choose politics* is a dummy variable indicating that the respondent chose to answer questions on politics and X is a vector of other covariates. If a willingness to discuss politics is associated with the content of political opinions, we have direct evidence against the ignorability assumption needed for weighting models.

Second, using the political opinions expressed by the 1,125 respondents in the politics response sample, we have data that includes a randomized first stage treatment variable that is uncorrelated with Y, but affects the probability of response. In this case, for example, those in the treatment group have a probability of response that is 0.5 lower than those in the control group. This enables us to estimate a first stage using the information necessary to identify endogenous selection. We simulate such a model in the next section.

3 Simulations

This section presents simulation results that illustrate how selection-sensitive survey design can enable selection models to function properly.

In the baseline models, the response rate is 10% and pollsters contact enough people to yield a sample of roughly 1,500 observations. A covariate X affects both response probability and the outcome.

The selection model for the opt-in design is

$$R_i^* = \gamma_0 + \gamma_1 \text{Opt-in treatment}_i + \gamma_2 X_i + \tau_i \quad (9)$$

where *Opt-in treatment*$_i = 1$ for those individuals randomly selected to be given a choice to politics

or some other topic.[8] We expect $\gamma_1 < 0$, meaning that we anticipate that respondents randomly presented with a choice of question topics will have a lower probability of responding to political items than respondents not given this choice. Because the opt-in treatment is randomly assigned it will produce variation in the probability of response that is unrelated to X and that will have no direct effect on Y.

The outcome equation is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{10}$$

where the correlation of τ and ϵ is defined as ρ and varies across simulations. We set $\beta_1 = 1$ in the simulations.

We analyze each simulated data set with OLS and weighting approaches, as described above. For the weighting model, we assume that we observe the covariate for all contacted individuals, whether they respond or not.

We also estimate two selection models. First, the "conventional Heckman" model includes X as the only variable in the first and second stage models; this corresponds to the common situation in which the variables that affect selection also affect outcomes. Second, in the selection-sensitive survey design, the roughly 1,500 respondents who are willing to respond are randomly divided into a control group and a treatment group. Everyone in the control group responds. Individuals in the treatment group are presented with an opt-in question that reduces their willingness to respond to political questions to 50% (conditional on the fact that they are willing to respond to the overall survey in the first place).[9]

We begin with a scenario in which the non-randomized covariate (X) has a relatively modest effect on selection. Specifically, we assume that $\gamma_2 = 0.3$ (and the variance of $\tau = 1$). As discussed in Appendix B, this induces a more-or-less linear relationship between the inverse Mill's ratio and X, which will undermine the ability of a conventional Heckman model to identify endogenous selection.

Figure 6 shows results for this case. The upper left panel shows the average value of $\hat{\beta}_1$ across 500

[8] Here the variable relates to the randomly assigned treatment. In Equation 8 the variable of interest refers to behavior of those exposed to the randomly assigned treatment.

[9] Allowing for non-response among the control group does not change results. The 50% response rate chosen for this simulation, could be higher or lower; future work could investigate what the optimal value of the drop-off is and how to generate questions that induce such a drop-off.

simulations for each value of ρ, the degree of correlation between the errors in the selection and outcome equations. The OLS and weighting models exhibit very similar patterns, with the bias increasing as the correlation of errors increases. The dashed line for the conventional Heckman model (which lacks a randomized treatment variable) is typically closer to the true value of one, but is highly variable. The solid line for the Heckman model with the randomized treatment variable is quite close to one, indicating no sign of bias.

The panel on the upper right shows the square root of the mean-squared error (RMSE) of the β_1 estimate for the various approaches. Weighting and OLS are similar, with OLS performing a bit better across the board. The RMSE for these two techniques increases as the correlation of errors increases.

The RMSE of the conventional Heckman model in Figure 6 is awful. Even though the upper left panel indicated that the conventional Heckman model is less biased than weighting and OLS, the conventional Heckman model is, in fact, essentially useless as the RMSE dwarfs the RMSE in the other models. This result will not surprise those with considerable experience with Heckman models as these models are prone to producing highly unstable and sometimes nonsensical results. The problem, of course, is that the inverse Mill's ratio from the first stage model is extremely highly correlated with the X variable in the outcome equation, making it very hard to identify both the effects of X and selection. This occurs even though the estimated effect of X in the first stage probit model is highly statistically significant, with the z-statistics averaging over 10.

The RMSE for the Heckman model with the randomized treatment instrument is excellent, coming in below the other approaches. This is the benefit of having a first stage variable that affects response, but is not correlated with the variable in the outcome equation. This strong performance occurs even though the selection-sensitive survey design model has about 25% *fewer* observations than the other models. The informational quality of the observations trumps the volume of observations in the other approaches. This state of affairs is analogous to a case in which a conventional survey based on random sampling proves more useful than a larger convenience sample as the randomness in the selection process offsets any advantages from having more non-randomized observations.

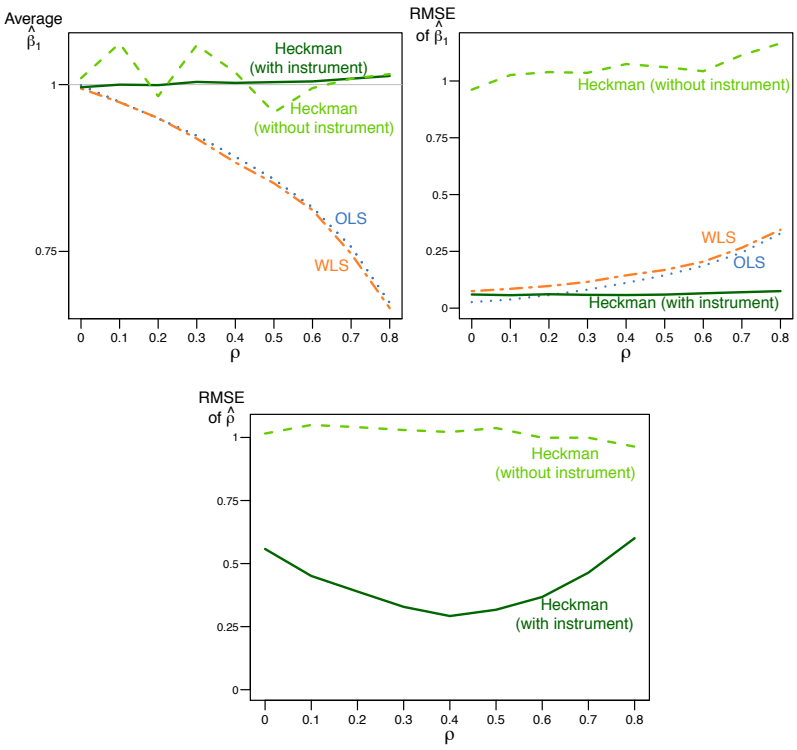

Figure 6: Simulation results for weak first stage covariate ($\gamma_2 = 0.3$)

The bottom panel of Figure 6 provides a clue toward what is happening by displaying the RMSE for the estimate of ρ for the two Heckman models. The selection-sensitive survey design Heckman model performs much better than the conventional Heckman model. In fact, the conventional Heckman model has a RMSE for ρ of around one, indicating that it is essentially useless in estimating the correlation of errors.

Our second set of simulations demonstrate what happens when the covariate has a stronger effect on selection by setting $\gamma_2 = 1.0$ (keeping variance of τ at 1). Figure 7 shows results for this scenario. The upper left panel plots the average value of $\hat{\beta}_1$ across 500 simulations for each value of ρ, the degree of correlation between the errors in the selection and outcome equations. OLS and weighting models become increasingly biased as the correlation of errors rises. The magnitude of the bias is larger than in Figure 6 because the larger γ_2 in the first stage induces a stronger relationship between X in outcome equation and the inverse Mill's ratio. There is less bias in the two Heckman models compared to OLS and weighting, with the conventional Heckman model showing the least bias.

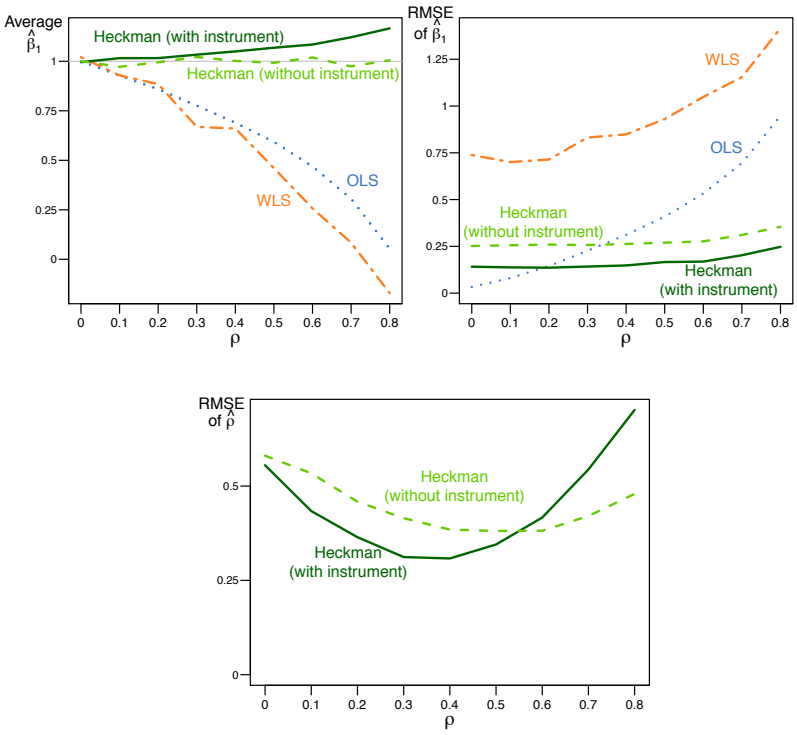

Figure 7: Simulation results for better first stage covariate ($\gamma_2 = 1.0$)

The panel in the upper right of Figure 6 shows the RMSE for the various approaches. The Heckman models are best, with the selection-sensitive survey design model performing the best (despite being based on less data), followed by the conventional Heckman model, OLS and weighting. Weighting performs the worst, indicating that when X has a large effect on selection, using a weight based on X can produce inaccurate results. Even though the Heckman model with the random first stage treatment has more bias than the conventional Heckman model, it has a lower RMSE. One reason for this is that the Heckman model with the randomized treatment variable generally estimates ρ more accurately, as evidenced by the bottom panel of Figure 7.

In summary, while these simulations confirm that selection models can indeed perform poorly, they also identify grounds for optimism. If we add an easy-to-implement randomized opt-in procedure to our survey design, selection models vastly outperform weighted models when there is non-trivial endogenous selection.

4 Survey examples

This section presents results from two surveys that used selection-sensitive survey design principles. The first survey was in the field March 9-15, 2016 (N = 1,075) and the second survey was in the field May 19-20, 2016 (N = 2,100). Both were fielded using Amazon's Mechanical Turk, an online service that pays people to answer surveys. This is a non-representative sample, but has been shown to provide reasonable characterizations of the U.S. population (Berinsky, Huber and Lenz 2010) and is especially useful to assess survey experiments where the researcher is more interested in characterizing treatment effects than in summarizing the U.S. population. The techniques described below need not be limited to Mechanical Turk, however.

In each survey, half of the respondents were randomly assigned to a control condition; these individuals were immediately asked to rate politicians on a feeling thermometer. The other half of respondents were assigned to a treatment condition; they were asked to pick a topic for questions from a list (see Figure 5). An individual who selected politics was given the same questions as the control group. A respondent who chose something else was first asked questions on the chosen topic and then asked the political questions.[10]

The political contexts of the two surveys were quite different. The first survey occurred in early March when the primaries were heating up and outcomes were uncertain. By late May, it was becoming clearer that Clinton would win the Democratic nomination and that Trump had a very realistic chance of winning the Republican nomination.[11]

[10] The May survey had two treatments: in one the alternatives to politics were sports and movies and in the other the alternatives to politics were sports, movies and health. For some questions, the effect of the treatments seems to differ, but at this point I have identified no systematic pattern and for simplicity model these two treatments as a single treatment. In the May survey, individuals who selected sports were asked to rate Bryce Harper, Serena Williams, Tom Brady and LeBron James. Individuals who selected movies were asked to rate Bradley Cooper, Will Smith, Jennifer Lawrence and Tina Fey. Individuals who selected health were asked a question about frequency of exercise and a question about how much nutrition affects their food choices. In the March survey, 538 respondents were given a choice of topics: 237 chose politics (44%), 213 chose movies (40%) and 88 chose sports (16%). In the May survey, 501 respondents were given three alternatives: 170 chose politics (34%), 240 chose movies (48%) and 91 chose sports (18%). Another 504 respondents were given four alternatives: 148 chose politics (29%), 160 chose movies (32%), 85 chose sports (17%) and 111 chose health (22%).

[11] There are a number of other differences between the surveys that we do not analyze here. For example, the March survey asked the political questions of those who chose sports or movies at the end of the survey; it did this only for the feeling thermometer political questions. The May survey asked the political feeling thermometer questions immediately following the non-political feeling thermometer questions and followed up on all political questions for those who chose other topics later.

The selection-sensitive survey design approach allows a direct test of the assumption that willingness to respond is independent of the content of response. Because we later asked the political questions of even those who chose to answer questions about movies or sports, we can compare the responses of those who chose to respond to the political questions (whom we label as "respondents" for the purposes of this discussion) and those who chose other topics (whom we label as "non-respondents" for the purposes of this discussion). We include covariates in the models. This is a direct test of the weighting-model assumption that willingness to respond is conditionally independent of opinions.

Figure 8 shows results and 95% confidence intervals for various dependent variables for the March survey from models that control for age, gender, religiosity, education, race and Hispanic ethnicity. The results are typically more statistically significant in models without covariates. Republican respondents (Republicans who chose to answer questions about politics) were 10.4 points cooler toward Hillary Clinton than were Republican non-respondents (Republicans who chose to answer non-political questions), a highly statistically significant difference. Republican respondents were 14.6 points more negative toward President Obama and 8.2 points more negative toward Bernie Sanders. There were no significant differences among Republicans between respondents and non-respondents with regard to Republican candidates (of whom, only Trump, Cruz and Rubio are shown in the figure for simplicity). Rubio was the only Republican candidate for whom Republican respondents were less favorable, although the difference was not statistically significant.

The right half of Figure 8 shows that Democratic respondents were more favorable than non-respondents toward Obama and Sanders, but not toward Clinton. The lack of a clear difference for Clinton may reflect an ambivalence among politically active Democrats toward Hillary Clinton at that time. Democratic respondents were less favorable toward Trump, Cruz and Rubio by about five points (although the difference for Rubio was not statistically significant).

These results clearly contradict the weighting assumption that willingness to respond is unrelated to opinions, at least for partisan samples.

For the population as a whole, however, this partisan polarization produced no evidence of selection

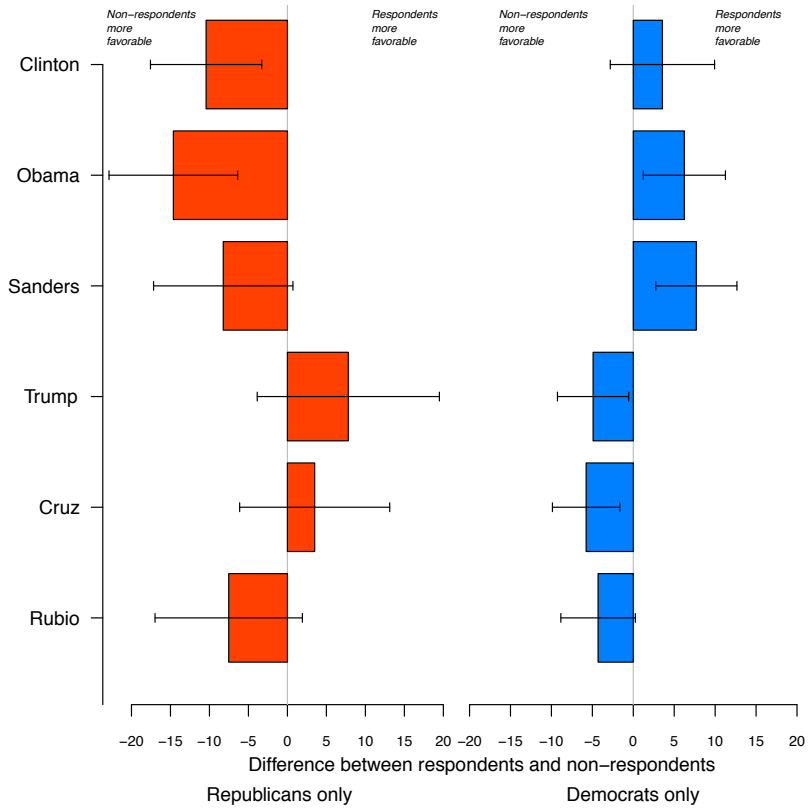

Figure 8: Difference in feeling thermometer ratings between those who chose political questions and those who do not in March 2016 survey, by party. Lines indicate 95% confidence intervals.

bias as the selection bias among Republicans countered the selection bias among Democrats. The one exception was Marco Rubio who was less popular among both Democratic and Republican respondents; for him (and only him) there was a statistically significant difference between respondents and non-respondents among the entire sample (Democrats, Republicans and independents).

Differences between respondents and non-respondents were different in the late May 2016 survey. When we included the same covariates as used in the models for the March survey, the differences between respondent and non-respondent Republicans were about half of what they were in the March survey (about minus 5 points for Clinton, minus 7 points for Obama and plus 5 points for Trump) and statistically significant for only Clinton and Obama. There were no differences among Democratic respondents and non-respondents.

The late May survey also included additional questions, allowing us to add more covariates, including a racial attitudes index (based on four questions about race), ideology (where high values indicate conservative ideology) and occupation (such as a dummy variable for self-identifying as working in a blue collar occupation).[12] When we added those covariates (including an interaction of age and ideology which was highly statistically significant across models), there were statistically significant differences between respondents and non-respondents when we looked at the entire sample. We report these results in Figure 9: respondents were less favorable toward Clinton and Obama and more favorable toward Trump.

The differences between the non-response patterns in the March and May surveys suggest that non-response patterns vary over time (see also Gelman, Goel, Rivers and Rothschild 2014). Others have uncovered similar patterns. Berinsky (2004) found non-response patterns that varied by question and over time while and Hopkins (2009) identified changes over time in the nature of non-response and other related biases.

Figure 10 reports the results for models in which answers to other questions on the May survey

[12]The racial attitudes battery asks respondents to respond on a strongly agree to strongly disagree scale to the following statements: "A history of slavery and discrimination makes it difficult for blacks to work their way out of the lower class." "If blacks would only try harder they could be as well off as whites." "Over the past few years, blacks have gotten less than they deserve." "Many minority groups in the U.S. have overcome prejudice. Blacks should do the same without any special favors."

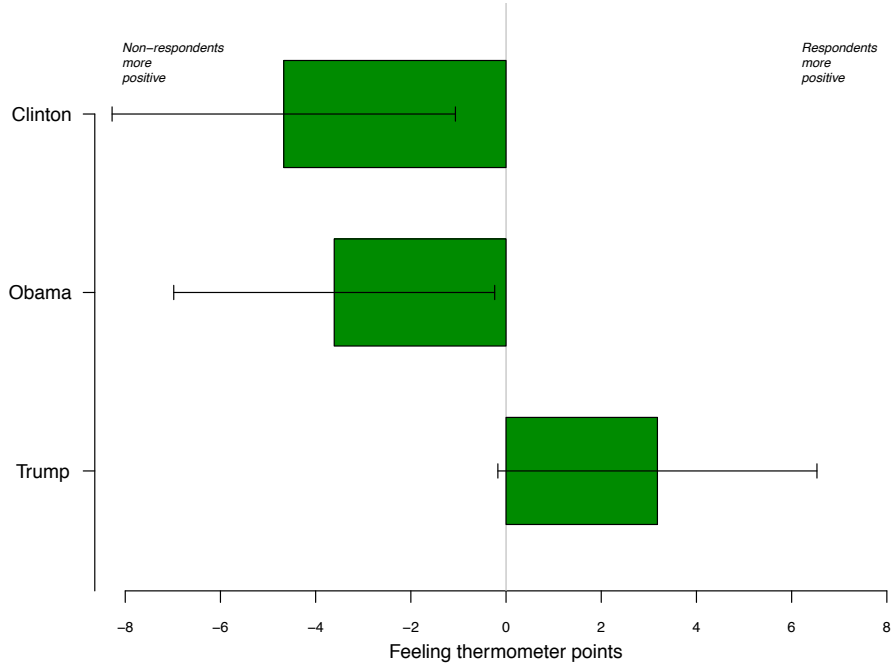

Figure 9: Difference in feeling thermometer ratings between those who chose non-politics questions and those who chose politics questions in May 2016 survey, all respondents. Lines indicate 95% confidence intervals.

are the dependent variables. All dependent variables are standardized so the effects are reasonably comparable. The survey asked respondents to indicate their likelihood of voting in the November 2016 election (coded from zero for not planning to vote to four for definitely planning to vote). Not surprisingly, respondents are more likely to indicate they will vote. This result serves as a validity check of the method as surveys regularly find that the proportion of survey respondents who say they will vote (or have voted) is much higher than the actual proportion of Americans who have actually voted (see, e.g., Brehm 1993).

Respondents were also less likely to say they would vote for Clinton over Trump and less likely to say that they expect Clinton would make a good president. These responses came at a time when Trump had considerable momentum in the primaries and before Democrats had unified around Clinton. On ideology, Democratic respondents were less conservative than Democratic non-respondents and Republican respondents were more conservative than Republican non-respondents.

Based on the whole population, respondents were less favorable toward the Black Lives Matter movement and were less likely to say that Congress should pass a law that addresses pay differences

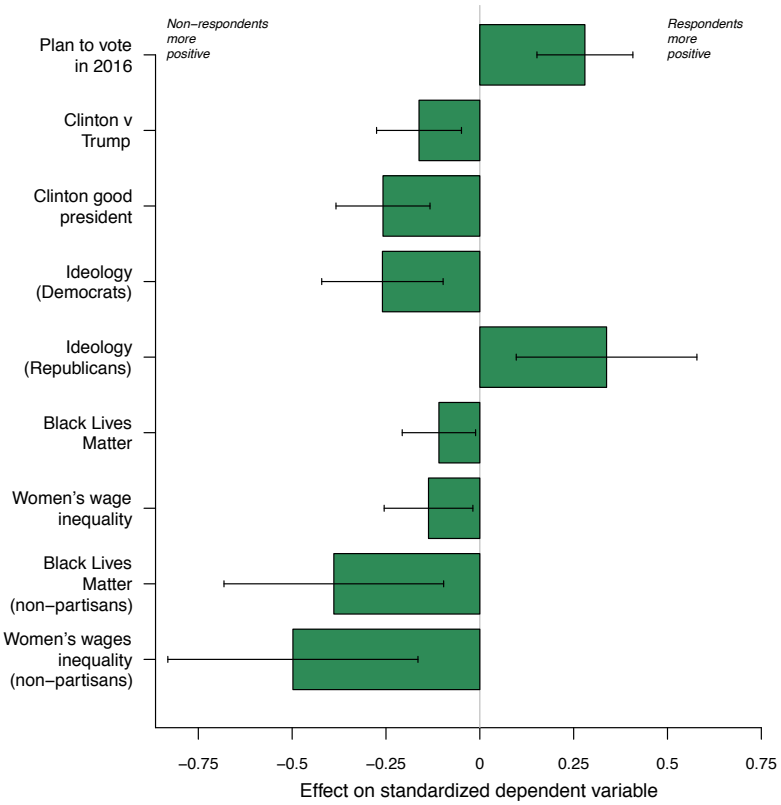

Figure 10: Difference in standardized responses between those who chose politics questions and those who did not in May 2016 survey. Lines indicate 95% confidence intervals.

between men and women. These effects appear concentrated among independents (N = 249 for the Black Lives Matter question and N = 265 for the wage inequality question).[13] There were no statistically significant differences between respondents and non-respondents on other questions on the May survey, including those relating to Muslim immigration, trade, bathroom accessibility for transgender people, the state of the U.S. economy and whether one would be upset if their child married a Democrat/Republican.

So far, we have presented evidence that suggests that willingness to respond is related to the content of political opinions. If so, weighting may be subject to considerable bias. The selection-sensitive survey design approach also produces data that will dramatically improve the performance of Heckman selection models, allowing us to create a parametric model that integrates the selection and outcome models. The core data set we use for this purpose is of individuals in the politics response set which includes everyone in the control group (except for the occasional individual in the control group who did not respond to a given question) and those in the treatment group who chose politics (which was about 40% of those given a choice of topics). The individuals who declined to answer the politics questions (those who occupy the dashed box in the lower right of Figure 4) are included in the first stage Heckman model, but not included in the second stage.

The results show that the randomized opt-in treatment question was necessary and useful. If the fit for the first stage selection model with only standard covariates is strong enough, the correlation between the inverse Mill's ratio and the other variables in the outcome equation may be manageably low. In this data, however, the Heckman models with only standard covariates are essentially useless. For feeling thermometers for each of the six politicians in Figure 8 I estimated a first stage probit using covariates for age, gender, education, race and Hispanic ethnicity and then estimated a model in which the inverse Mill's ratio from the first stage probit was regressed on those covariates. This corresponds to a common situation in which we believe the factors that affect opinion content may also

[13] We have not adjusted for multiple comparisons. On the one hand, doing so will widen the confidence intervals. On the other hand, if we are using these models to test the null hypothesis that willingness to respond is unrelated to political opinions, we may be more interested in avoiding Type II error that would occur if we say that there is no endogenous selection when there is endogenous selection. Multiple comparison adjustments focus on accurately assessing Type I error.

affect willingness to respond. The R^2_{IMR} in all cases is above 0.97 and for none of these dependent variables did a conventional Heckman model converge. In other words, a standard Heckman model is infeasible given the standard covariates. Given such a result, it is not surprising that analysts of conventional surveys do not use Heckman-type models to assess endogenous selection: such models fail due to lack of sufficiently informative data.

The selection-sensitive survey design approach provides the data needed to make a Heckman model work. It provides an additional covariate for the first stage, the randomly assigned treatment status. The R^2_{IMR} from models including the treatment variable are never higher than 0.21 for the six politicians we are investigating (recall that the lower this value, the better). In other words, the randomized opt-in treatment gives us enough separation between the inverse Mill's ratio and the other covariates to expect reliable estimation of Heckman models.

Figure 11 shows the estimates of ρ from Heckman models based on the politics sample which consists of everyone in the control group who responded and those in the treatment group who chose politics. Estimates based only on Republicans (N = 297) are on the left. The $\hat{\rho}$ estimates for Clinton and Obama are both around -0.8, indicating a strong tendency for Republicans who disliked these two to be more likely to answer political questions. The magnitudes for the other candidates are around 0.3 and statistically significant as well, also suggesting a clear relationship between willingness to respond and the content of opinions. Among Democrats (N = 606), the estimates of ρ are between 0.24 and 0.37 for Clinton, Obama and Sanders, all of which are statistically significant. The estimates for ρ for Trump, Cruz and Rubio based on the Democratic sample are statistically insignificant.

The estimates of ρ from the May survey are generally smaller. For Hillary Clinton feeling thermometers, $\hat{\rho}$ was -0.14 (p = 0.018), a pattern that was stronger among Republicans ($\hat{\rho} = -0.26$; p = 0.03) than Democrats ($\hat{\rho} = -0.13$; p = 0.06), but negative for partisans on both sides. Using the entire population, there was weak evidence of positive selection for Trump feeling thermometers ($\hat{\rho} = 0.11$; p = 0.09) and negative selection for Obama feeling thermometers ($\hat{\rho} = -0.10$; p = 0.11), both of which seemed to be concentrated among Democratic respondents. There was clear evidence of negative se-

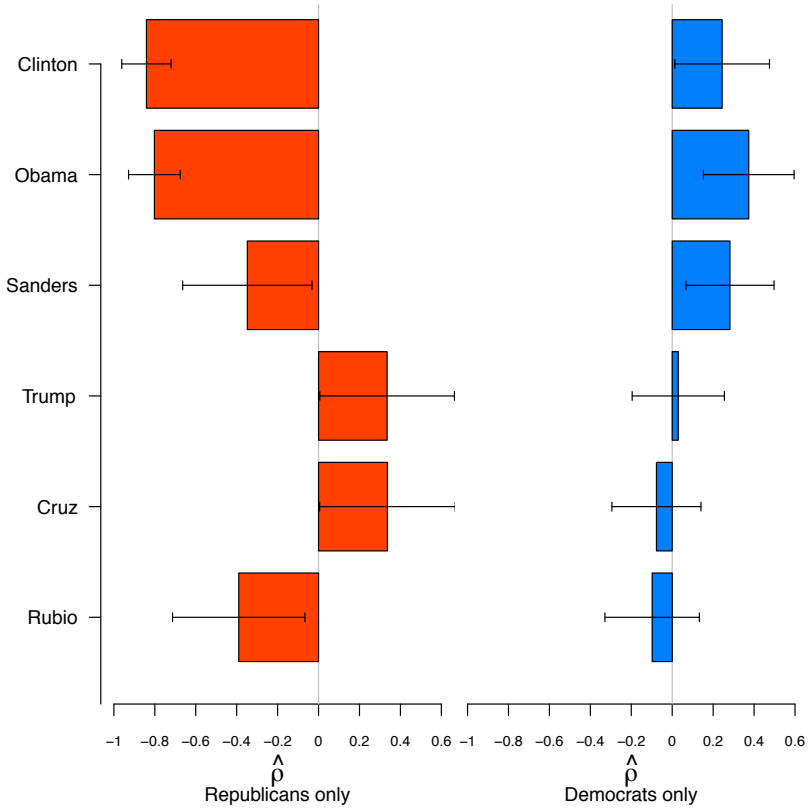

Figure 11: Estimate of ρ from Heckman model with randomized first stage treatment in March 2016 survey. Lines indicate 95% confidence intervals.

lection for answers about preferring Clinton over Trump in the general election ($\hat{\rho} = -0.22$; p = 0.008), with the selection parameter significant among Democratic respondents, but not Republican ones.

5 Conclusion

There are two ways to look at contemporary polling. One is to view the glass as half-full: despite low response rates and the strong possibility that people who respond to contemporary surveys are unrepresentative of the broader population, surveys have generally performed well, at least when properly weighted and when predicting national electoral outcomes. In fact, the biggest story of contemporary polling may be that surveys have survived the so-called death of random sampling. Based on this view, one could believe surveys are generally fine and to chalk up failures to the hard reality of life in an uncertain world.

Another view is more cautious: the problem – or possibility – of selection bias is relentless. Every survey, indeed every survey question, can suffer from non-response bias (Berinsky 2004; Groves et al 2009).

The stakes are high. Presidential election polls in 2016 systematically erred in ways that could have affected strategies and choices of campaigns and voters. While we do not definitively know whether non-response bias was the major contributor to polling errors, evidence suggests it mattered. Silver (2016) found that polling errors were systematically worse in states with higher percentages of white working class voters. If the white working class voters who responded to polls were representative of all such voters, then weighting would have taken care of any under (or over) responsiveness by white working class voters. However, if white working class Trump supporters were less likely to respond to surveys, weighting would do nothing but make us overconfident in incorrect results. Research by Enns, Schuldt, Lagodny and Rauter (2016) found that Trump voters were less willing to indicate their support for Trump on surveys.

The vast majority of pollsters deal with non-response by weighting their data. This will make polls more accurate if there is no endogenous selection. Otherwise, weighting can make poll results less

representative of the target population. This makes weighting about as attractive as driving without a seatbelt: it usually works out, but is a bad idea nonetheless.

This paper argues that we should proactively deal with the possibility of nonresponse bias. Every survey should incorporate explicit steps to diagnose and, if necessary, correct for endogenous selection. Two steps are easy to implement: we can inhibit response among a random subset of respondents and we can follow-up to get responses in the same survey for those who initially avoid answering political questions. Such data will produce data that makes it easy to test directly assumptions underlying weighting models and to estimate a selection model that allows non-response to be affected by unmeasured variables that also affect Y.

These steps do not cure non-response bias. No technique can definitively characterize the views of people who do not respond to surveys. It is possible, for example, that non-response patterns among people who are unreachable in the survey differ from the non-response patterns among those who are reachable by the survey. Selection sensitive survey design techniques can, however, test the hypothesis underlying weighting and related methods. Weighting models assume that there is no connection between response propensity and opinion. Selection-sensitive survey design techniques test that assumption for reachable people in the target population. If we reject the hypothesis of no response bias in this context, we have rejected the key assumption for weighting, at least for a set of reachable people in the target population. This allows us to move beyond the widespread, yet untenable, practice of simply assuming away endogenous non-response bias.

Future work can also explore combining the techniques presented here with methods that account for heterogeneous effects. For example, it is possible that the effect of the first-stage randomized treatment varies across subpopulations. Perhaps the instrument affects well-educated people less than others. Accounting for such heterogeneity could produce estimates that are more precise and potentially identify substantively interesting variation in behavior. The easiest way to investigate such possibilities is via interactions between the instrument and variables a researcher suspects may account for heterogeneous effects. A more expansive and, potentially, careful approach may involve LASSO or other techniques

for variable selection, especially when the number of possible covariates is high and one wants to avoid over-fitting the data.

Non-response is a challenge that cuts to the very heart of survey research. The current norm is to use weighting models that assume away endogenous non-response. A better approach is to design our surveys that allow us to diagnose and, when necessary, correct for bias that might arise due to endogenous non-response.

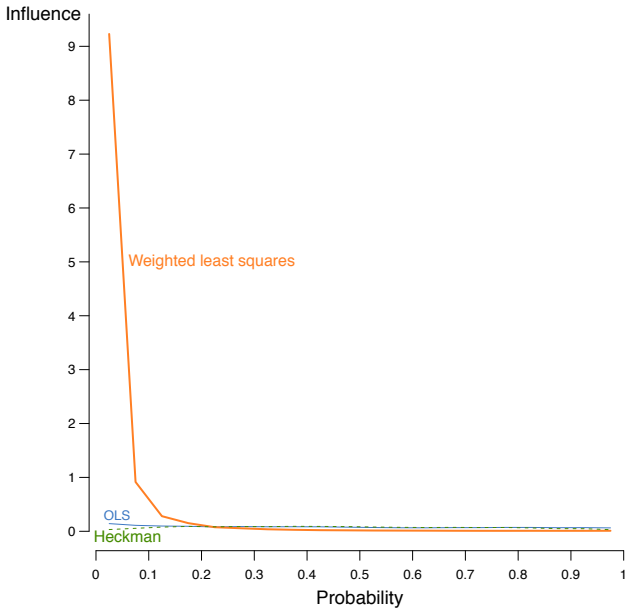

Figure A.1: Influence in OLS, weighting and Heckman models

Appendix

A Marginal effects in model with independent variable

The differential treatment of observations across models also occurs when we include control variables. Figure A.1 depicts simulation results for the influence of observations for OLS, weighting and Heckman models with one control variable. In each simulation, we simulate whether individuals respond or not based on Equation 2 and then estimate coefficients using OLS, weighting (Equation **??**) and selection models (Equation 5). The x-axis labels indicate the fitted probability of selection and the y-axis indicates the average influence on $\hat{\beta}_1$ for observations with in a bin near the indicated probability of selection. Each bin on the x-axis is 0.05 wide, meaning the plots reflect the average changes in the coefficient estimate for observations with probabilities in the specified ranges. Influence is measured with the absolute value of a dfbeta statistic for each observation for $\hat{\beta}_1$; this statistic captures the change in $\hat{\beta}_1$ that would occur if we were to exclude a given observation from the analysis. The results are based on 100 simulations of data sets with 300 individuals each.[14]

Low probability observations exert a huge effect in weighted models, with observations in the lowest probability bin (from 0.0 to 0.05) producing dfbetas that average above nine. The variation in dfbetas for the Heckman model pales in comparison. The figure will not surprise those who have dealt with instability in weighted models associated with extreme weights on low probability observations (see, e.g., Samii 2011, 19) but highlights the potentially dramatic effect of low probability observations in these models. Common sense suggests that we should be sure these models are appropriate before implementing them.

B Weaknesses in Heckman-type models

This section highlights the sources of the problems with the Heckman selection model. I focus on problems arising from poor model fit in the first stage model that causes severe multicollinearity between the inverse Mill's ratio and the other independent variables in the outcome equation. In

[14] The pattern in the figure is not sensitive to sample size or the magnitude of the correlation of errors.

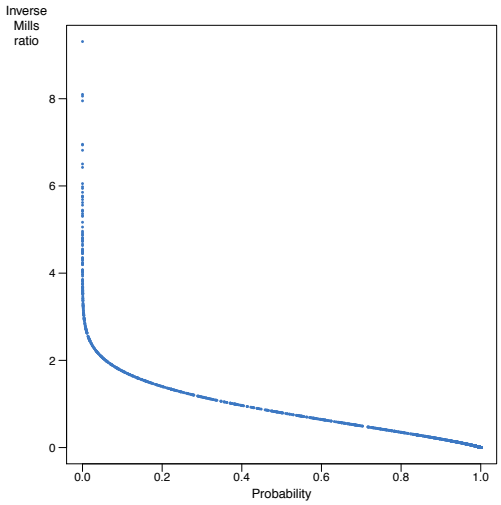

Figure B.1: Relationship of inverse Mill's ratio and fitted probability

perfectly ordinary circumstances, this multicollinearity ravages the statistical power for estimates of ρ and explodes the variance for the estimates of coefficients on independent variables.[15]

At first glance, the inverse Mill's ratio seems to be a highly non-linear function. However, linearity lurks, creating sometimes-catastrophic econometric problems. Figure B.1 shows the inverse Mill's ratio as a function of the probability of selection. It is essentially a straight line with a kink at lower probabilities. If our fitted values are largely confined to the higher probabilities, then our inverse Mill's ratio will be essentially a linear function of the fitted probabilities (Vella 1998).

Of course, the fitted probabilities are based on a probit model and are themselves non-linear. However, it is very common for fitted probit probabilities to be close to linear as well; it is, for example, unsurprising for a linear probability model (LPM) to produce fitted values that correlate very highly with probit fitted values.

Hence, if we have a situation in which the fitted probabilities are generally above 0.1 and in which a LPM produces fitted probabilities that approximate probit fitted values, the inverse Mill's ratio will be very close to being a linear function of the independent variables.

Figure B.2 illustrates the connection between first stage model results and the multicollinearity in the second stage by displaying the relationships between the probability of being selected, the inverse Mill's ratio and X for three cases. In each, the latent propensity of observing an observation is

$$R_i^* \quad = \quad \gamma_0 + \gamma_1 X_{1i} + \tau_i$$

Here we consider cases in which the covariate is the same in both the selection and outcome equations. This is quite common as factors that affect response propensity may also affect the outcome.

[15]There are other problems with the Heckman model. The Heckman model assumes errors in the selection and outcome equations are distributed bivariate normally. Das, Newey and Vella (2003) show that replacing the inverse Mill's ratio with a higher order function of the probability of selection will cover a broad range of possible joint distributions for the errors. Another issue is that model misspecification in the first stage biases Heckman coefficients toward OLS estimates. This means that the inverse Mill's ratio we estimate will have error and will suffer from same ills as any variable measured with error. In particular, the estimated $\hat{\lambda}$ coefficient will be attenuated relative to the true value, making it less likely that we will reject the null hypothesis that $\lambda = 0$. The inverse Mills ratio measurement errors will typically be correlated with the value of the true value of the inverse Mills ratio, making the consequences of measurement error more complicated than the typical example in which measurement error is assumed to be independent of the true value. Nonetheless, the attenuation bias associated with the independent error example carries over here under general conditions. Designing surveys with randomized treatments that affect the propensity to respond mitigates this problem by improving first stage model fit.

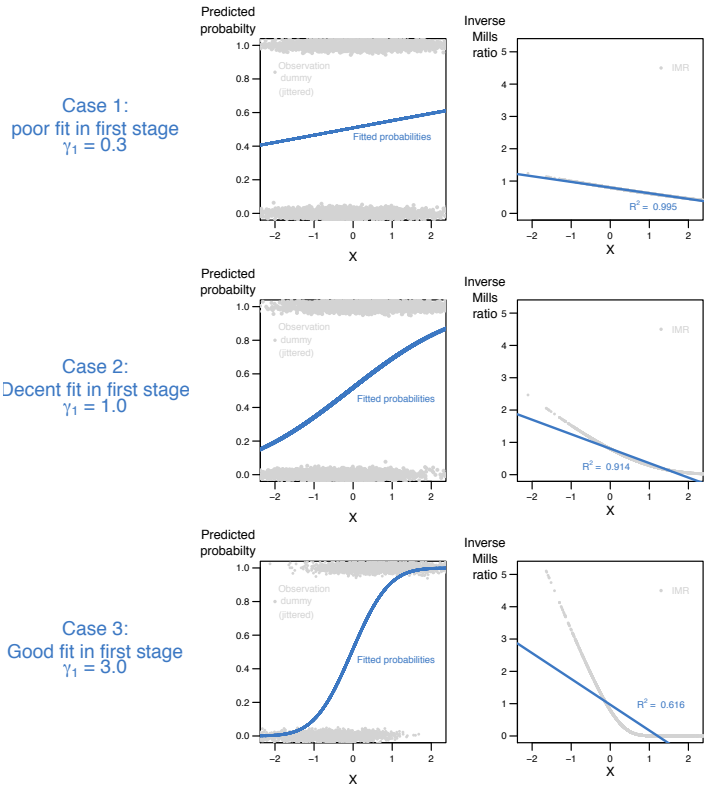

Figure B.2: Relationship of inverse Mill's ratio and X for three cases

In public opinion models, for example, factors such as income, education, age, and party identification plausibly affect both whether people respond to a pollster and the content of their opinions. (Later we explore models in which at least one first stage variable is not included in the second stage.)

In Case 1, $\gamma_1 = 0.3$, indicating that X has relatively limited influence in the first stage model predicting selection. The panel on the top left of Figure B.2 shows a scatterplot of a dummy variable indicating whether or not Y was observed as a function of X (the grey dots jittered at the top and bottom of the plot) and a fitted line from the first stage probit model predicting observation. There is only a weak relationship and the fitted values vary little. (Note, however, that the poor fit does not necessarily correspond to statistical significance; in the simulations below the effect of X in the selection equation is highly statistically significant even in the "poor-fit" simulations corresponding to the $\gamma_1 = 0.3$ case.)

The panel on the top right of Figure B.2 shows the relationship between X and the inverse Mill's ratio that will also be entered in the outcome equation for the Heckman model. Since X has limited explanatory power in the first stage, there is very little variation in predicted probabilities of observation, the key input into the inverse Mill's ratio. In this case, the inverse Mill's ratio variable is essentially a linear function of X, as indicated by the R^2 from an auxiliary regression of the inverse Mill's ratio on X of 0.995.

In Case 2, $\gamma_1 = 1$, suggesting a better fit in the first stage probit model. The fit from the first stage shows more variation and this in turn leads to an inverse Mill's ratio in the middle right panel of Figure B.2 that is not simply a linear function of X. There is still a strong linear relationship, but at least we can see some distinction between X and the inverse Mill's ratio.

In Case 3, $\gamma_1 = 3$, suggesting a strong relationship between X and selection. The fit from the first stage shown in the bottom right of Figure B.2 shows more variation and exhibits the classic "s shape" of probit models. This means that the fitted values that enter into the inverse Mill's ratio vary more substantially and therefore induce more variation in the inverse Mill's ratio, allowing it to be noticeably distinct from X.

From standard multicollinearity results, we know that the Heckman model will be a disaster in

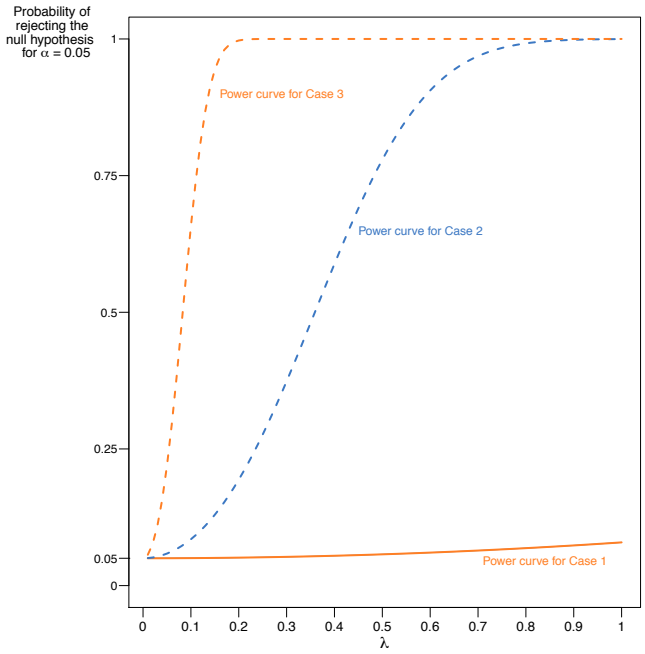

Figure B.3: Power curve for H_0: $\lambda = 0$ for three values of γ and N = 2,500

Case 1, as including X and the inverse Mill ratio in the same outcome equation will induce massive multicollinearity, inflating the standard errors and causing unreliable estimates. The problem persists in Case 2. For Case 3, the multicollinearity is in the range that we often see in observational studies.

Figure B.3 shows the power curves for testing the null hypothesis that $\lambda = 0$. This is a test of whether there is endogenous selection and the power varies considerably for the three cases. In Case 1, with the terrible first stage fit, we have virtually no statistical power to identify endogenous selection. In Case 2, we have decent power when λ is around 0.4 and above. We have good statistical power when we have good fit in the first stage model as we did in Case 3.

What is the best way to assess whether we have sufficient data for a Heckman model? Before presenting an answer, we can first reject the rule of thumb that it is necessary and possibly sufficient to have a variable that is statistically significant in the selection equation and excluded from the outcome equation. This logic is not unreasonable as having such a variable helps break the connection between the inverse Mill's ratio and the variables included in the outcome equation. This rule of thumb is also appealing because it corresponds to the intuition in two-stage least squares models, where an exclusion condition is necessary.

However, simply excluding a variable in the selection model from the outcome model is neither sufficient nor necessary to produce a Heckman model that will produce accurate estimates. In selection models, the (potential) non-linearity of the inverse Mill's ratio means that it can be perfectly valid to have the same variables in both equations. And, even if a variable excluded from the outcome equation is statistically significant in the selection equation, the inverse Mill's ratio can still suffer from model-killing multicollinearity.[16]

A better way to assess the sufficiency of data for a Heckman-type model is to assess directly the degree of multicollinearity in the outcome equation between the inverse Mill's ratio and the independent variables in the outcome equation. Multicollinearity can be measured via R^2_{IMR}, the R^2 produced in a regression of the inverse Mill's ratio on all the other variables in the outcome equation.[17] If the R^2_{IMR} is

[16] A common rule of thumb for two stage least squares models is that the F-statistic for a test of the null that the instruments exert no effect in the first stage needs to be greater than 10 (corresponding to a t-statistic over 3.2 for a single instrument). In Heckman models, it is possible to have a high z-statistic on a first-stage only variable and still have poor model properties. For example, in all the γ scenarios discussed above, we can almost always decisively reject the null hypothesis in the first stage regression that $\gamma_1 = 0$.

[17] The condition number of the matrix of independent variables in the outcome equation provides a very similar,

low enough that multicollinearity does not cause unworkably large standard errors on coefficients, then we can simply use the results from the Heckman model. If R^2_{IMR} is large, then we need to think about the statistical power of our tests. Analysts should report this (or a similar diagnostic) as a matter of course so that readers will know if it is even possible to know if there is endogenous selection.

We can do more than diagnose our ability to identify endogenous selection. We can take steps to improve our ability to identify and correct for endogenous selection. First, we can reduce the multicollinearity by including one or more variables in the selection equation that are excluded from the outcome equation. While this is difficult in observational studies, survey researchers design their surveys to allow for variation in response probabilities that depends on factors unrelated to factors that affect the outcome. Specifically, we can create a randomized treatment that affects propensity to respond, but does not affect the outcome of interest.[18]

Traditionally, pollsters have focused on trying to increase the probability of response, through incentives to respond or increased contact efforts (Singer and Ye 2013). Such efforts can produce useful information, but require more resources and can, if they involve persistent calls, degrade the quality of responses (Fricker and Tourangeau 2010). An attractive alternative is to reduce the multicollinearity in Heckman-type models by lowering the probability of response for a randomly selected subset of potential respondents. Figure B.1 demonstrated that the relationship between the inverse Mill's ratio and the probability of selection is quite linear above probabilities of 0.10 or so (Vella 1998, 135). The more observations we have in the low probability range, the more non-linearity we will introduce into the estimation process, thereby increasing power and reducing the chances of having a Heckman model that goes off the rails. Intuitively, while a high $\hat{p}_i$ person could have either a high or a low τ_i and still be observed, a low $\hat{p}_i$ person must have a high τ_i in order to be observed.[19]

although perhaps less intuitive, diagnostic measure (see, e.g., Bushway, Johnson, and Slocum 2007; Puhani 2000).

[18] Of course, every randomly sampled survey already creates a randomized variable that affects response, but observations with zero probability of response will have undefined inverse Mills ratio.

[19] Two additional approaches to dealing with high multicollinearity may be of limited use in practice. First, we could offset the multicollinearity with more data. However, if we simply add more of the same data, we will need a lot more data, something that is expensive and, often infeasible for data sets with fixed sizes. Second, we could improve the fit in the first stage, which would produce an inverse Mill's ratio that exhibits a weaker linear relationship with X. Usually, however, scholars have already done what they can to produce good model fit in the first stage with the variables at hand.

References

Achen, Chris. 1986. *The Statistical Analysis of Quasi-Experiments.* Berkeley: University of California Press.

Aron-Dine, Aviva, Liran Einav, and Amy Finkelstein. 2013. The RAND Health Insurance Experiment, Three Decades Later. *Journal of Economic Perspectives* 27(1): 197-222.

Aronow, Peter M., Alan S. Gerber, Donald P. Green, Holger Kern, and Michael J. LaCour. 2015. Double Sampling for Nonignorable Missing Outcome Data in Randomized Experiments. Manuscript.

Bailey, Michael A., Daniel J. Hopkins, and Todd Rogers. 2016. Unresponsive and Unpersuaded: The Unintended Consequences of Voter Persuasion Efforts. *Political Behavior* 38 (3): 713-746.

Berinsky, Adam J., Gregory A. Huber, Gabriel S. Lenz. 2010. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis* 20: 351368.

Berinsky, Adam J. 2004. *Silent Voices: Public Opinion and Political Participation in America.* Princeton University Press.

Bethlehem, Jelke G. 2002. Weighting Nonresponse Adjustments Based on Auxiliary Information. in Groves, Robert M., Don A. Dilman, John L. Eltinge, Roderick J.A. Little, ed. *Survey Nonresponse.* New York: Wiley Series in Probability and Statistics.

Bradburn, Norman M. 1992. Presidential Address: A Response to the Nonresponse Problem. *Public Opinion Quarterly* 56 (3): 391-397.

Brehm, John. 1993. *The Phantom Respondents: Opinion Surveys and Political Representation.* University of Michigan Press.

Bushway, Shawn, Brian D. Johnson, Lee Ann Slocum. 2007. Is the Magic Still There? The Use of the Heckman Two-Step Correction for Selection Bias in Criminology. *Journal of Quantitative Criminology* 23:151178.

Caughey, Devin and Erin Hartman. 2017. Target Selection as Variable Selection: Using the Lasso to Select Auxiliary Vectors for the Construction of Survey Weights. Paper presented at the 2017 Conference of the Society of Political Methodology Society.

Das, Mitali, Whitney K. Newey, and Francis Vella. 2003. Nonparametric Estimation of Sample Selection Models. *The Review of Economic Studies* 70(1): 33-58.

Dutwin, David and Paul J. Lavrakas. 2016. Trends in Telephone Outcomes, 2008 - 2015. *Survey Practice* 9(2).

Enns, Peter K., Jonathon P. Schuldt, Julius Lagodny and Alexander Rauter. 2016. Why the polls missed in 2016 Was it shy Trump supporters after all? MonkeyCage.com. December 13 at `https://www.washingtonpost.com/news/monkey-cage/wp/2016/12/13/why-the-polls-missed-in-2016-was-it-shy-trump-supporters-after-all`.

Franco, Annie, Neil Malhotra, Gabor Simonovits, L.J. Zigerell. 2015. Developing Standards for Post-Stratification Weighting in Population-Based Survey Experiments. Manuscript.

Fricker, S. and R. Tourangeau. 2010. Examining the Relationship between Nonresponse Propensity and Data Quality in Two National Household Surveys. *Public Opinion Quarterly* 74(5): 935-955.

Gelman, Andrew, Sharad Goel, Douglas Rivers, and David Rothschild. 2014. The Mythical Swing Voter. Manuscript, Columbia University, Stanford University and Microsoft.

Groves, Robert M., Don A. Dilman, John L. Eltinge, Roderick J.A. Little, ed. 2002. *Survey Nonresponse.* New York: Wiley.

Groves, Robert M. and Emilia Peytcheva. 2008. The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly* 72(2): 167 189.

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*, 2nd edition. New York: Wiley.

Gutsche, Tania L., Arie Kapteyn, Erik Meijer, and Bas Weerman. 2014. The RAND Continuous 2012 Presidential Election Poll. *Public Opinion Quarterly* 78, Special Issue: 233 - 254.

Heckman, James J. 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47(1): 153161.

Hopkins, Daniel J. 2009. No Wilder Effect, Never a Whitman Effect: When and Why Polls Mislead About Black and Female Candidates. *Journal of Politics* 71(3)769-781.

Lauderdale, Benjamin and Douglas Rivers. 2016. Beware the Phantom Swings: Why Dramatic Bounces in the Polls Aren't Always What They Seem. YouGov.com (November 1). today.yougov.com/news/2016/11/01/beware-phantom-swings-why-dramatic-swings-in-the-p/

Little, Roderick J. and Sonya Vartivarian. 2005. Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology* 31(2): 161-68.

Mason, Robert, Virginia Lesser and Michael W. Traugott. 2002. Effect of Item Nonresponse on Nonresponse Error and Inference. In Groves et al, ed. *Survey Nonresponse*. New York: Wiley.

Peress, Michael. 2010. Correcting for Survey Nonresponse Using Variable Response Propensity. *Journal of the American Statistical Association* 105(492): 1418-1430.

Pew Research Center. 2012. Assessing the Representativeness of Public Opinion Surveys. Available at http://www.people-press.org/2012/05/15/.

Peytchev, Andy. 2013. Consequences of Survey Nonresponse. *Annals of the American Academy of Political and Social Science*, 645 (1):88-111.

Puhani, Patrick. 2000. The Heckman Correction for Sample Selection and Its Critique. *Journal of Economic Surveys* 14, 1: 53-68.

Rubin, Donald B. 2008. For Objective Causal Inference, Design Trumps Analysis. *Annals of Applied Statistics* 2 (3): 808-840.

Samii, Cyrus. 2011. Weighting and Augmented Weighting for Causal Inference with Missing Data: New Directions. Working Paper, New York University.

Sartori, Anne. 2003. An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions. *Political Analysis* 11: 111 - 138.

Siddique, Juned and Thomas R. Belin. 2008. Using an Approximate Bayesian Bootstrap to Multiply Impute Nonignorable Missing Data. *Computational Statistics & Data Analysis* 53(2):405-415.

Silver, Nate. 2016. Pollsters Probably Didnt Talk To Enough White Voters Without College Degrees. FiveThirtyEight.com. December 1 at https://fivethirtyeight.com/features/pollsters-probably-didnt-talk-to-enough-white-voters-without-college-degrees/.

Singer, E. and C. Ye. 2013. The Use and Effects of Incentives in Surveys. *The Annals of the American Academy of Political and Social Science* 645(1): 112- 141.

Solon, Gary, Steven J. Haider, Jeffrey Wooldridge. 2013. What are We Weighting For? *Journal of Human Resources* 50, 2: 301-316.

Stolzenberg, Ross M. and Daniel A. Relles. 1997. Tools for Intuition about Sample Selection Bias and Its Correction. *American Sociological Review* 62: 494-507.

Thompson, B., M. Kyrillidou and C. Cook. 2009. Item Sampling in Service Quality Assessment Surveys to Improve Response Rates and Reduce Respondent Burden. *Performance Measurement and Metrics* 10(1): 6 - 16.

Vella, Francis. 1998. Estimating Models with Sample Selection Bias: A Survey. *Journal of Human Resources* 23(1): 127- 169.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Massachusetts: The MIT Press.

Wooldridge, Jeffrey M. 2007. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* 141:1281-1301.