

# Measures of Selection Bias in Regression Coefficients Estimated from Non-Probability Samples

Brady T. West, Roderick J.A. Little, Rebecca R. Andridge, Phil Boonstra, Erin B. Ware,  
Anita Pandit, Fernanda Alvarado-Leiton

## ABSTRACT

Selection bias in survey estimates is a major concern, particularly for non-probability samples. Recent developments have provided survey researchers with model-based indices of the potential selection bias in estimates of means and proportions computed from non-probability samples that may be subject to non-ignorable selection mechanisms. To our knowledge, there are currently no systematic approaches for measuring selection bias for regression coefficients, a problem of great practical importance. Generalizing recent developments, we derive novel measures of selection bias for estimates of the coefficients in linear regression models. The measures arise from normal pattern-mixture models that allow analysts to examine the sensitivity of their inferences to assumptions about the extent of the non-ignorable selection, and they leverage auxiliary variables available for the population of interest that provide information about the variable being modeled when conditioning on the predictors of interest. A simulation study is described that assesses the effectiveness of the measures. We then apply the measures to data from two real studies: one that aims to predict health indicators with polygenic scores in a volunteer sample recruited via Facebook, and a second that aims to model months worked in the past year among smartphone users with lower education.

**Key Words:** Linear Regression, Genes for Good, Polygenic Scores, National Survey of Family Growth

## 1. Introduction

The random selection of elements from a finite population of interest into a probability sample, where all population elements have a known non-zero probability of selection, ensures that elements included in the sample, appropriately weighted, mirror the target population in expectation. That is, for all variables of interest, the mechanism of selection of a subset of elements into the sample is ignorable, following the theoretical framework for missing data mechanisms originally introduced by Rubin (1976). Unfortunately, the collection of survey data from probability samples, which has long been a staple of the social and health sciences, is becoming much more difficult. The modern survey research environment presents substantial challenges: sampled units are harder to contact, survey response rates continue to decline in all modes of administration (face-to-face, telephone, etc.; Brick and Williams, 2013; de Leeuw, Hox and Luiten., 2018; Williams and Brick, 2018), and the costs of collecting and maintaining scientific probability samples are steadily rising (Presser and McCulloch, 2011).

Because of these problems and challenges, researchers are turning to the collection and analysis of data from non-probability samples. These data may be scraped from social media platforms, collected from commercial databases, gathered from online searches, or recorded via online surveys of volunteers (Baker et al., 2013). In clinical trials, inferences about the effect of a treatment in a target population are often made based on volunteer samples. Whatever their source, these data are generally cheap to collect and enable researchers to compile large data sets quickly. However, the resulting samples are all ultimately non-probability samples, and design-based methods of making inference about finite target populations from probability samples are invalid when applied to data from non-probability samples (Elliott and Valliant, 2017). The protection of ignorable selection conveyed by

probability sampling no longer applies in these settings, and extremely low response rates in probability samples raise similar concerns. Estimates computed from non-probability samples may be biased, depending on the extent to which the underlying selection mechanism is non-ignorable, and model-based inferential methods for these samples are an active focus of current research (Elliott and Valliant, 2017; Valliant, 2019).

There is thus a critical need for diagnostic measures that researchers can use to both assess and correct for the bias in estimates that can arise from non-ignorable selection mechanisms in non-probability samples. Nishimura, Wagner and Elliott (2016) demonstrated that no existing measures could effectively detect the selection bias introduced by non-ignorable survey nonresponse. Little, West, Boonstra and Hu (2019) and Andridge, West, Little, Boonstra and Alvarado-Leiton (2019) addressed this absence of informative diagnostic measures for estimates of means and proportions, respectively. These authors leveraged estimators from Andridge and Little (2011) in the context of non-ignorable survey nonresponse to develop effective model-based indicators of the potential selection bias in descriptive estimates generated from non-probability samples. These measures have now been shown to outperform alternative diagnostic measures such as the R-indicator (Schouten, Cobben and Bethlehem, 2009) in simulations (Boonstra, Andridge, West, Little and Alvarado-Leiton, 2019). A limitation is that all the work in this area has focused on measuring the potential bias in means and proportions.

Selection bias can also affect estimates of the relationships between variables, which are often of interest. In particular, good measures are needed of the extent to which estimates of regression coefficients from a given non-probability sample are subject to bias due to non-ignorable selection. In this article, we extend recently-developed indices of non-ignorable

selection bias for estimates of means and proportions to regression coefficients. We evaluate the ability of these measures to detect selection bias in a simulation study and then apply them to estimated coefficients in two non-probability samples (from genetics and survey settings) where true values of the coefficients of interest are available for benchmarking.

## 2. Models and Methods

Assume that a non-probability sample has data  $D = \{Y_i, Z_i, A_i, i = 1, \dots, n\}$ , where  $i$  is the unit of analysis, the sample is of size  $n$ ,  $Y_i$  is an outcome variable of interest,  $Z_i$  is a  $p \times 1$  vector of predictor variables of interest, and  $A_i$  is a vector of auxiliary variables. The analysis of interest is a linear regression of  $Y$  (possibly with missing values) on  $Z$ . The auxiliary variables  $A$  are not included in the regression model of interest but are still predictive of  $Y$  after conditioning on  $Z$ . Deviations from ignorable selection are not detectable from the sample alone (Little et al., 2019), so measures of selection bias require information from an external data source. Here we assume that summary statistics (means, variances, and covariances) for  $Z$  and  $A$  are available for the entire population, either from auxiliary data or some other external source (e.g., Census data, administrative records, or a large probability sample like the American Community Survey).

Our proposed indices are based on a pattern-mixture model for  $Y$  and  $A$  given  $Z$  (Little, 1994).

Let  $S$  be a selection indicator, equal to 1 for population units in the non-probability sample

and 0 otherwise. We suppose that  $E(Y | Z, A, S = 1) = \beta_{y0:za}^{(1)} + \beta_{yz:za}^{(1)t} Z + \beta_{ya:za}^{(1)t} A$ , where

$X = (\beta_{ya:za}^{(1)t} A)$  is the best predictor of  $Y$  in the non-probability sample after conditioning on  $Z$ .

Here and throughout the paper, we use the notation  $\beta_{yz:za}^{(1)t}$  to refer to the coefficients for  $Z$  in a

regression model for  $Y$  given  $Z$  and  $A$ , fitted to the selected sample ( $S = 1$ ). Similarly,  $\beta_{y0:za}^{(1)}$

refers to the intercept in a model for  $Y$  given  $Z$  and  $A$ . We further define  $X^* = X \sqrt{\sigma_{yy \cdot z}^{(1)} / \sigma_{xx \cdot z}^{(1)}}$  as the *auxiliary proxy* for  $Y$ , scaled to have the same residual variance as  $Y$  when conditioning on  $Z$ .  $V$  denotes the variables in  $A$  that are orthogonal to  $X$  given  $Z$  (i.e.,  $\beta_{xv \cdot zv}^{(1)t} = 0$ ). We assume a normal pattern-mixture model (Little, 1994) for  $Y$  and  $X$  given  $Z$ ,  $V$  and  $S$ :

$$\begin{pmatrix} X \\ Y \end{pmatrix} | Z, V, S \sim N \left( \begin{pmatrix} \beta_{x0 \cdot zv}^{(s)} + \beta_{xz \cdot zv}^{(s)t} Z + \beta_{xv \cdot zv}^{(s)t} V \\ \beta_{y0 \cdot zv}^{(s)} + \beta_{yz \cdot zv}^{(s)t} Z + \beta_{yv \cdot zv}^{(s)t} V \end{pmatrix}, \begin{pmatrix} \sigma_{xx \cdot zv}^{(s)} & \sigma_{xy \cdot zv}^{(s)} \\ \sigma_{xy \cdot zv}^{(s)} & \sigma_{yy \cdot zv}^{(s)} \end{pmatrix} \right), \quad (2)$$

where

$$\Pr(S=1 | X, Y, Z, V) = g(Y^*, Z, V), \quad Y^* = (1-\phi)X^* + \phi Y \quad (3)$$

and  $g$  is an unknown function. The parameter  $\phi$  is an unknown scalar, and because  $X^*$  is a proxy for  $Y$ , we assume  $\phi$  is positive, that is,  $0 \leq \phi \leq 1$ . The parameter  $\phi$  is a measure of the “degree of non-random selection,” after conditioning on  $X^*$ , and no information is available on  $\phi$  in the data.

We make the following assumptions about the model specified in (2) and (3):

- a)  $E(Y | Z, A, S=0) = \beta_{y0 \cdot za}^{(0)} + \beta_{yz \cdot za}^{(0)t} Z + \beta_{ya \cdot za}^{(0)t} A$ , where  $\beta_{ya \cdot za}^{(0)t} = \lambda \beta_{ya \cdot za}^{(1)t}$ , that is,

$X = (\beta_{ya \cdot za}^{(1)t} A)$  is the best predictor of  $Y$ , after conditioning on  $Z$ , for both selected and non-selected cases; and

- b)  $V$  is orthogonal to  $X$  given  $Z$  for non-selected cases, that is,  $\beta_{xv \cdot zv}^{(s)t} = 0$  for  $S=0,1$

(one can test this assumption given microdata on the non-selected cases).

Under a) and b),  $\beta_{yv \cdot zv}^{(s)} = \beta_{yv \cdot zv}^{(s)} - \frac{\sigma_{xy \cdot zv}^{(s)} \beta_{xv \cdot zv}^{(s)}}{\sigma_{xx \cdot zv}^{(s)}} = \beta_{yv \cdot zv}^{(s)} = 0$  for  $S=0,1$ , so (2) reduces to

$$\begin{pmatrix} X \\ Y \end{pmatrix} | Z, V, S \sim N \left( \begin{pmatrix} \beta_{x0 \cdot zv}^{(s)} + \beta_{xz \cdot zv}^{(s)t} Z \\ \beta_{y0 \cdot zv}^{(s)} + \beta_{yz \cdot zv}^{(s)t} Z \end{pmatrix}, \begin{pmatrix} \sigma_{xx \cdot zv}^{(s)} & \sigma_{xy \cdot zv}^{(s)} \\ \sigma_{xy \cdot zv}^{(s)} & \sigma_{yy \cdot zv}^{(s)} \end{pmatrix} \right). \quad (4)$$

Because  $X$  and  $Y$  are independent of  $V$  given  $Z$  and  $S$ , we can simplify the notation in (4) by dropping the subscript  $v$  in the parameters, resulting in

$$\left( \begin{pmatrix} X \\ Y \end{pmatrix} \middle| Z, V, S \right) \sim N \left( \begin{pmatrix} \beta_{x0 \cdot z}^{(s)} + \beta_{xz \cdot z}^{(s)t} Z \\ \beta_{y0 \cdot z}^{(s)} + \beta_{yz \cdot z}^{(s)t} Z \end{pmatrix}, \begin{pmatrix} \sigma_{xx \cdot z}^{(s)} & \sigma_{xy \cdot z}^{(s)} \\ \sigma_{xy \cdot z}^{(s)} & \sigma_{yy \cdot z}^{(s)} \end{pmatrix} \right). \quad (5)$$

Consider first the setting where  $\phi = 1$  in (3). This implies that selection depends only on  $Y$ ,  $Z$ , and  $V$ , and therefore the regression of  $X$  on  $Y$ ,  $Z$ , and  $V$  is the same for both patterns defined by  $S$ . Hence, we have

$$\beta_{x0 \cdot yz}^{(1)} = \beta_{x0 \cdot yz}^{(0)}, \beta_{xy \cdot yz}^{(1)} = \beta_{xy \cdot yz}^{(0)}, \text{ and } \beta_{xz \cdot yz}^{(1)} = \beta_{xz \cdot yz}^{(0)}. \quad (6)$$

Maximum likelihood (ML) estimates (or draws) of these parameters can therefore be obtained from the regression of  $X$  on  $Y$  and  $Z$  for the non-probability sample ( $S = 1$ ). ML estimates (or draws) of the parameters  $(\beta_{x0 \cdot z}^{(s)}, \beta_{xz \cdot z}^{(s)}, \sigma_{xx \cdot z}^{(s)})$  from the regression of  $X$  on  $Z$  are obtained from the regression models fitted to each respective pattern ( $S = 0, 1$ ), and can be computed given only means, variances, and covariances for the non-selected cases.

We can now express the unidentified parameters of the regression of  $Y$  on  $Z$  for  $S = 0$  (the non-selected cases) in terms of the identified parameters above. The intercept of the regression of  $Y$  on  $Z$  for  $S = 0$  can be written as

$$\beta_{y0 \cdot z}^{(0)} = \frac{\beta_{x0 \cdot z}^{(0)} - \beta_{x0 \cdot yz}^{(0)}}{\beta_{xy \cdot yz}^{(0)}} =_{(\text{by 6})} \frac{\beta_{x0 \cdot z}^{(0)} - \beta_{x0 \cdot yz}^{(1)}}{\beta_{xy \cdot yz}^{(1)}} = \frac{\beta_{x0 \cdot z}^{(0)} - (\beta_{x0 \cdot z}^{(1)} - \beta_{xy \cdot yz}^{(1)} \beta_{y0 \cdot z}^{(1)})}{\beta_{xy \cdot yz}^{(1)}} \quad (7)$$

and hence  $\beta_{y0 \cdot z}^{(0)} = \beta_{y0 \cdot z}^{(1)} + \frac{\beta_{x0 \cdot z}^{(0)} - \beta_{x0 \cdot z}^{(1)}}{\beta_{xy \cdot yz}^{(1)}}$ . Similarly, for the slope of  $Z$  and the residual variance

of the regression of  $Y$  on  $Z$ , we have:

$$\begin{aligned}\beta_{yz \cdot z}^{(0)} &= \beta_{yz \cdot z}^{(1)} + \frac{\beta_{xz \cdot z}^{(0)} - \beta_{xz \cdot z}^{(1)}}{\beta_{xy \cdot yz}^{(1)}} \text{ and} \\ \sigma_{yy \cdot z}^{(0)} &= \sigma_{yy \cdot z}^{(1)} + \frac{\sigma_{xx \cdot z}^{(0)} - \sigma_{xx \cdot z}^{(1)}}{(\beta_{xy \cdot yz}^{(1)})^2}.\end{aligned}\tag{8}$$

ML estimates (or draws) of these parameters can be obtained by substituting the ML estimates (or draws) of the identified parameters on the right-hand sides of these expressions.

For other values of  $\phi$ , the transformation  $Y_\phi = \phi Y + (1 - \phi) X^*$  yields

$$\beta_{y0 \cdot z}^{(0)} = \beta_{y0 \cdot z}^{(1)} + \left( \frac{\phi + (1 - \phi)\rho_{xy \cdot z}^{(1)}}{\phi\rho_{xy \cdot z}^{(1)} + (1 - \phi)} \right) \sqrt{\frac{\sigma_{yy \cdot z}^{(1)}}{\sigma_{xx \cdot z}^{(1)}}} (\beta_{x0 \cdot z}^{(0)} - \beta_{x0 \cdot z}^{(1)}),\tag{9}$$

where  $\rho_{xy \cdot z}^{(1)} = \sigma_{xy \cdot z}^{(1)} / \sqrt{\sigma_{xx \cdot z}^{(1)} \sigma_{yy \cdot z}^{(1)}}$ . We also have

$$\begin{aligned}\beta_{yz \cdot z}^{(0)} &= \beta_{yz \cdot z}^{(1)} + \left( \frac{\phi + (1 - \phi)\rho_{xy \cdot z}^{(1)}}{\phi\rho_{xy \cdot z}^{(1)} + (1 - \phi)} \right) \sqrt{\frac{\sigma_{yy \cdot z}^{(1)}}{\sigma_{xx \cdot z}^{(1)}}} (\beta_{xz \cdot z}^{(0)} - \beta_{xz \cdot z}^{(1)}) \text{ and} \\ \sigma_{yy \cdot z}^{(0)} &= \sigma_{yy \cdot z}^{(1)} + \left( \frac{\phi + (1 - \phi)\rho_{xy \cdot z}^{(1)}}{\phi\rho_{xy \cdot z}^{(1)} + (1 - \phi)} \right)^2 \left( \frac{\sigma_{yy \cdot z}^{(1)}}{\sigma_{xx \cdot z}^{(1)}} \right) (\sigma_{xx \cdot z}^{(0)} - \sigma_{xx \cdot z}^{(1)}).\end{aligned}\tag{10}$$

As before, ML estimates (or draws) of these parameters are obtained by substituting ML estimates (or draws) of the identified parameters above into these expressions.

We propose using the differences between the ML estimates of the regression parameters for the selected and non-selected cases (based on the pattern-mixture model) as a *Measure of Unadjusted Bias* for the regression coefficients as compared to the *Non-Selected* cases (MUBNS). Given the results above, our proposed MUBNS for the intercept can be written as

$$\text{MUBNS}_0(\phi) = \beta_{y0 \cdot z}^{(1)} - \beta_{y0 \cdot z}^{(0)} = \left( \frac{\phi + (1 - \phi)\hat{\rho}_{xy \cdot z}^{(1)}}{\phi\hat{\rho}_{xy \cdot z}^{(1)} + (1 - \phi)} \right) \sqrt{\frac{\hat{\sigma}_{yy \cdot z}^{(1)}}{\hat{\sigma}_{xx \cdot z}^{(1)}}} (\hat{\beta}_{x0 \cdot z}^{(1)} - \hat{\beta}_{x0 \cdot z}^{(0)})\tag{11}$$

and the MUBNS indices for the slopes can be written as

$$\text{MUBNS}_z(\phi) = \beta_{yz \cdot z}^{(1)} - \beta_{yz \cdot z}^{(0)} = \left( \frac{\phi + (1-\phi)\hat{\rho}_{xy \cdot z}^{(1)}}{\phi\hat{\rho}_{xy \cdot z}^{(1)} + (1-\phi)} \right) \sqrt{\frac{\hat{\sigma}_{yy \cdot z}^{(1)}}{\hat{\sigma}_{xx \cdot z}^{(1)}}} (\hat{\beta}_{xz \cdot z}^{(1)} - \hat{\beta}_{xz \cdot z}^{(0)}). \quad (12)$$

If the selection fraction is small (as is the case with many non-probability samples), the differences defining the MUBNS indices in (11) and (12) essentially capture the bias in the regression coefficients estimated from the selected cases relative to the regression coefficients based on the *entire population*. By the law of total probability, we know that

$$E(Y | Z, S = 1) - E(Y | Z) = [E(Y | Z, S = 1) - E(Y | Z, S = 0)] \times \Pr(S = 0 | Z). \quad (13)$$

The pattern mixture model specified in (2) – (5) provides a comparison of the regression coefficients for  $S = 1$  and  $S = 0$ , as in the first term on the right-hand side of (13). For a comparison with the regression coefficients for the whole population (the *entire* right-hand side of [13]), the impact of the difference in coefficients at a particular value of  $Z$  depends on the non-selection rate  $\Pr(S = 0 | Z)$  for that value of  $Z$ . We note that the relative impact of selection on coefficients for different  $Z$  variables does not depend on  $Z$ , that is,  $\Pr(S = 0 | Z)$  is a constant factor in this comparison. If the overall selection rate for the non-probability sample is non-negligible and is known or can be estimated, we propose a Measure of Unadjusted Bias (MUB) for the selected cases that compares the coefficients to those for the *entire population*, by multiplying the MUBNS indices by the overall non-selection rate:

$$\text{MUB}_0(\phi) = \text{MUBNS}_0(\phi) \times \Pr(S = 0) \text{ and } \text{MUB}_z(\phi) = \text{MUBNS}_z(\phi) \times \Pr(S = 0). \quad (14)$$

We make the following four remarks about the indices proposed in (11), (12), and (14):

1. In the case where the regression model of interest only includes an intercept (i.e.,  $Z$  does not exist), the MUB index defined in (14) equals the unstandardized MUB index presented in Little et al. (2019) for means of continuous variables.



2. We recommend defining posterior distributions for these indices by performing a fully Bayesian analysis with a prior distribution on  $\phi$ , as described by Little et al. (2019) and Andridge et al. (2019). One can then use credible intervals for the MUBs defined in (11) and (12) to make inference about the selection bias. We consider this Bayesian approach, outlined in detail in the online supplementary materials, in our simulation study and our applications.
3. Little et al. (2019) and Andridge et al. (2019) also note the importance of having at least a moderate correlation between  $X$  and  $Y$ , which in our regression framework corresponds to having a moderate value of  $\rho_{xy \cdot z}^{(1)}$ , for these indices to be effective indicators of selection bias.
4. In the case where  $Y$  is a binary variable and the parameters of interest are the coefficients in a probit regression model of  $Y$  on  $Z$ , the pattern-mixture model above can be applied to an underlying latent standard normal variable  $U$  that gives rise to  $Y$  (where  $Y = 1$  if  $U > 0$ ).

### 3. Simulation Study

#### 3.1 Design of the Simulation Study

We assess the effectiveness of the proposed MUBNS indices via a simulation study; the study also serves as an assessment of the effectiveness of the MUB indices when the selection rate is known or can be estimated. Let  $Y$  be the outcome variable of interest, let  $Z_1$  and  $Z_2$  be the predictor variables of interest in the target regression model, and let  $A$  be an auxiliary variable that serves as the proxy variable. We repeatedly generate populations of size  $N = 10,000$  from the following superpopulation model:

$$\begin{pmatrix} Y \\ Z_1 \\ Z_2 \\ A \end{pmatrix} \sim N \left( \begin{pmatrix} 10 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 2\rho_{y1} & 2\rho_{y2} & \sigma_{ya} \\ 2\rho_{y1} & 1 & 0 & \rho_{1a} \\ 2\rho_{y2} & 0 & 1 & 0 \\ \sigma_{ya} & \rho_{1a} & 0 & 1 \end{pmatrix} \right) \quad (15)$$

Note that the predictor variables of interest are orthogonal to each other,  $Z_1$  is correlated with  $A$ , and  $Z_2$  is uncorrelated with  $A$ .

We set  $\rho_{y1}$  and  $\rho_{y2}$ , i.e., the correlations of  $Y$  with each of  $Z_1$  and  $Z_2$ , to be 0.2 (low), 0.4 (medium), and 0.6 (high). Specifically, these three possible correlations are crossed in our design. In addition, we choose three possible values for the correlation between  $Y$  and  $A$  given  $Z_1$  and  $Z_2$ : 0.2, 0.5, and 0.8. We then determine the values of  $\sigma_{ya}$  given these values. Finally, we choose three values for  $\rho_{1a}$ , the correlation between  $Z_1$  and  $A$ : 0.2, 0.4, and 0.6. This results in  $3 \times 3 \times 3 \times 3 = 81$  possible population distributions.

Next, for the sample selection mechanism applied to a given simulated population, we define the following selection model:

$$\text{logit}(P(S = 1|Y, Z_1, Z_2, A)) = \gamma_0 + \gamma_y Y + \gamma_{z1} Z_1 + \gamma_{z2} Z_2 + \gamma_a A \quad (16)$$

Here,  $S$  is the selection indicator (1 = selected, 0 = not selected), and conditional on the values of other parameters defined below, we select a value of  $\gamma_0$  that will result in a 5% selection fraction for a given simulated population. We then determine the selection mechanism by choosing a combination of the following parameters:

- $\gamma_y = \{0, \ln(1.1), \ln(2)\}$ ; here,  $\gamma_y = 0$  implies Selection At Random
- $\gamma_{z1} = \{\ln(1.1), \ln(2)\}$
- $\gamma_{z2} = \{\ln(1.1), \ln(2)\}$
- $\gamma_a = \{\ln(1.1), \ln(2)\}$

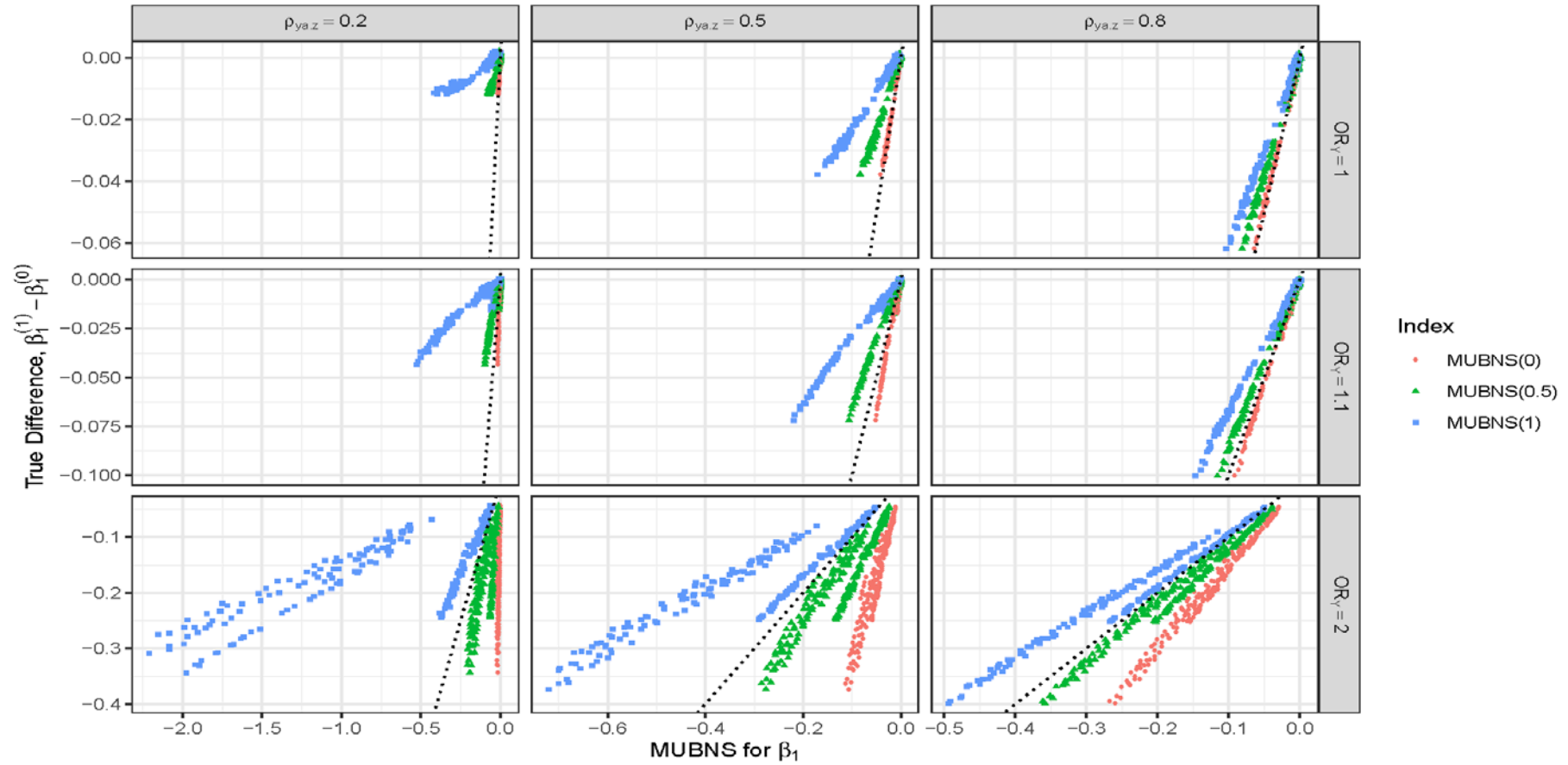
These values represent either no effects on selection ( $Y$  only;  $OR=1$ ), small effects on selection (all variables;  $OR=1.1$ ), and strong effects on selection (all variables;  $OR=2$ ). The various combinations of these parameters result in  $3 \times 2 \times 2 \times 2 = 24$  possible selection mechanisms. For a given selection mechanism, we refer a random  $UNIFORM(0,1)$  draw to the realized probability of selection for a given population unit based on (15), and set  $S = 1$  for that case if the draw is less than the realized probability for that unit. The complete simulation experiment therefore features  $81 \times 24 = 1,944$  combinations of data generation model and selection mechanism. For each of these combinations, we repeated the process of simulating a population of size  $N = 10,000$  and applying the specific selection mechanism 1,000 times. The simulations were programmed in R, and the simulation code is available at <https://github.com/bradytwest/IndicesOfNISB>.

The intercept and slopes from the linear regression of  $Y$  on  $Z_1$  and  $Z_2$  were the parameters of interest, and thus for each simulated non-probability sample we computed the values of the proposed MUBNS indices at  $\phi = \{0, 0.5, 1\}$  for each of these parameters. We used the variable  $A$  to construct the proxy. We then compared the computed indices to the true estimated differences between the regression parameters for the selected and non-selected cases, which were available in each simulated dataset. For our first set of evaluations, we plotted the true values of the differences between coefficients against the computed values of the indices and compared the resulting relationships to a line representing a perfect 1:1 relationship. For our second set of evaluations, we examined side-by-side boxplots showing the distributions of Spearman correlations of the true differences between the coefficients for the selected cases and the population coefficients (i.e., the bias in the coefficients for the selected cases) and the MUBNS indices as a function of  $\phi$ .

Finally, following Little et al. (2019), we computed the percentage of simulated scenarios where intervals defined by  $[MUBNS(0), MUBNS(1)]$  (denoted by “MLE”) covered the true difference in the coefficients. We also evaluated the coverage properties and median widths of 95% credible intervals for MUBNS (based on the 2.5 and 97.5 percentiles of the distribution of posterior draws of MUBNS) following the fully Bayesian approach outlined in the online supplementary materials. We considered two potential approaches for drawing values of  $\phi$  when following the fully Bayesian approach: random draws from a UNIFORM(0,1) distribution (“Bayes-Uniform”) and random draws from a discrete distribution where values of 0, 0.5, and 1.0 have equal probability (“Bayes-Discrete”).

### ***3.2 Simulation Study Results***

Figure 1 presents results from all simulated scenarios and illustrates the associations between the median value of MUBNS across the 1,000 simulations and the true differences for the  $Z_1$  coefficient in the model of interest (very similar results were found for the other two coefficients). When  $Y$  is independent of the probability of selection (row 1 of Figure 1), MUBNS(0) correlates perfectly with the difference, as expected, and the MUBNS(0.5) and MUBNS(1) indices do not perform as well. Notably, the performance of MUBNS(0.5) and MUBNS(1) improves with stronger conditional correlations between  $A$  and  $Y$  in these and all other scenarios, i.e., these estimates are closer to the true difference. In the two non-ignorable scenarios (rows 2 and 3 of Figure 1), the performance of MUBNS(0) becomes weaker as the dependence of selection on  $Y$  becomes stronger (going down the rows of Figure 1), and we see that MUBNS(0.5) and MUBNS(1) tend to be closer to the actual differences. MUBNS(0.5) tends to work well in most scenarios, supporting the idea of computing this index as a starting point for assessing potential bias (consistent with the recommendations of Little et al., 2019). The poor performance of MUBNS(1) illustrated in the first two panels of

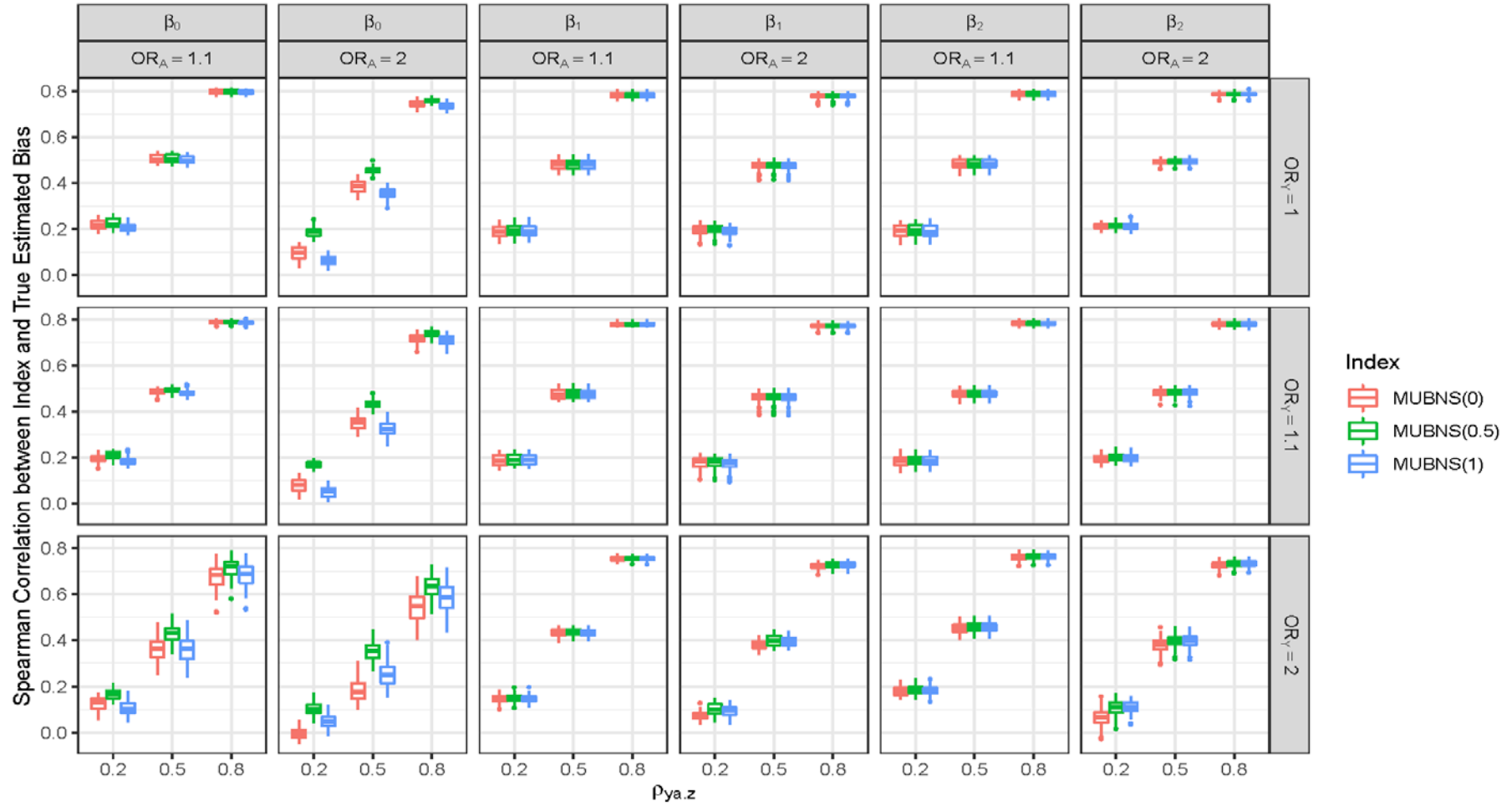


**Figure 1:** Scatter plots presenting associations between MUBNS and the true differences in coefficients between selected and non-selected units for the  $Z_1$  coefficient. Results are median MUBNS values across 1,000 simulated datasets for each of the 1,944 combinations of data generation model and selection mechanism; panels are separated by the level of dependence on  $Y$  in the selection model ( $OR_Y$ ; rows) and the correlation between  $Y$  and  $A$  given  $Z_1$  and  $Z_2$  (columns). The dotted black line represents the  $Y = X$  relationship.

the third row arises when  $A$  has a stronger association with selection and the conditional correlation between  $A$  and  $Y$  is weaker.

Figure 2 presents the distributions of the Spearman correlations between the MUBNS index values and the true biases under the different scenarios. The clear story that emerges from this set of results is the importance of the conditional correlation between  $A$  and  $Y$  for maximizing the Spearman correlation between the MUBNS index and the true difference in the coefficients for selected and non-selected cases. The MUBNS indices correlate reasonably well with the true difference (bias) when  $Cor(Y, A/Z_1, Z_2)$  is high but do worse as  $Cor(Y, A/Z_1, Z_2)$  decreases. The correlations between MUBNS and the true bias vary little across all possible scenarios considered in one of these 18 panels, with most of the uncertainty emerging for the intercept and when the conditional correlation is 0.2. Since each panel contains results that pool across values of  $\{\gamma_1, \gamma_2\}$ , which are the log-odds of selection for  $Z_1$  and  $Z_2$ , we can conclude that how strongly  $Z_1$  and  $Z_2$  are associated with selection does not have much impact on the performance of the MUBNS indices. Similarly, each of the 18 panels combines results across all values of  $\{\rho_{y1}, \rho_{y2}, \rho_{1a}\}$ , suggesting that the correlations of  $Y$  and  $A$  with  $Z_1$  and  $Z_2$  are not as influential as the conditional correlation between  $Y$  and  $A$  given  $Z_1$  and  $Z_2$ .

Figure 3 presents empirical distributions of the rates at which the proposed [MUBNS(0), MUBNS(1)] intervals (based on the MLEs) and the Bayesian intervals (“Bayes-Uniform” and “Bayes-Discrete”) cover the true differences in the coefficients across the different scenarios. As seen in Figure 3, the coverage of the proposed MLE-based interval improves when the dependence of selection on the dependent variable  $Y$  becomes stronger, and especially when selection depends more on the auxiliary proxy  $A$ . Notably, the coverage rates *decrease* when



**Figure 2:** Side-by-side box plots presenting distributions of the Spearman correlations between MUBNS and the true difference in the coefficients between selected and non-selected units. We estimate each correlation from 1,000 replicate populations for each combination of data generation model and selection model.  $OR_A$  = odds ratio for  $A$  in the selection model;  $OR_Y$  = odds ratio for  $Y$  in the selection model.

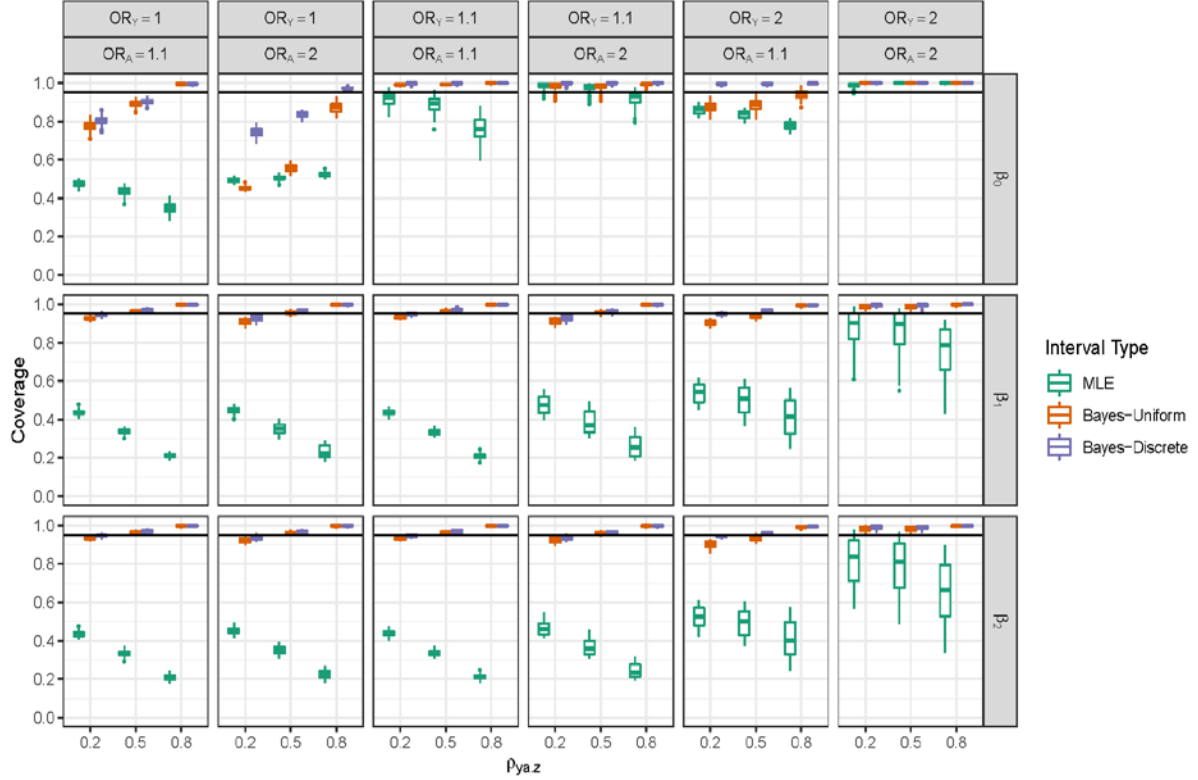
the auxiliary proxy  $A$  has a stronger conditional association with  $Y$ . This is because the MLE-based intervals become narrower in the presence of more informative auxiliary data, and for the selection mechanisms that are close to ignorable, the true MUBNS is close to the interval lower bound, i.e. close to  $\text{MUBNS}(0)$ .

Figure 3 also shows that the Bayesian intervals have improved coverage of the actual differences in the coefficients relative to the MLE-based intervals in nearly all scenarios, with coverage improving given stronger auxiliary proxies and declining only for the intercept when selection depends on  $A$  and not  $Y$  (the first two columns). We do note that in the specific scenario where  $\rho_{y1}$  is 0.2 ( $Z_1$  is weakly associated with  $Y$ ),  $\rho_{y2}$  is 0.6 ( $Z_2$  is strongly associated with  $Y$ ), the correlation between  $Y$  and  $A$  given  $Z_1$  and  $Z_2$  is 0.8 (we have access to a strong proxy / auxiliary information), and  $Z_1$  has a strong correlation with  $A$  (0.6), the Bayesian intervals tend to over-cover the difference in the  $Z_1$  coefficients (at least 0.99 coverage) across all missingness mechanisms. This high coverage needs to be weighed against the width of the resulting credible intervals, which we consider next.

To examine whether the good coverage of the Bayesian intervals in Figure 3 is simply arising from wide intervals, Figure 4 presents empirical distributions of the median widths of the intervals under the different scenarios. For context, the empirical ranges of median MUBNS values for the three coefficients across the different simulation scenarios were (0.02, 19.52), (-2.21, 0.01), and (-1.87, 0.02), respectively. If one were to consider “typical” MUBNS values of 3, -1, and -1 for each coefficient, an excessively wide 95% credible interval would have a width at least 33% larger than the estimate itself, meaning that widths of 4, 1.33, and 1.33 would be considered “excessive” for these typical MUBNS values.



Figure 4 shows that the median widths of the MLE-based and Bayesian credible intervals are generally quite reasonable across most of the scenarios (including the case of low conditional correlations of  $A$  with  $Y$ ). The credible intervals based on the discrete prior for  $\phi$  tend to become slightly wider in the presence of less informative auxiliary information. When the



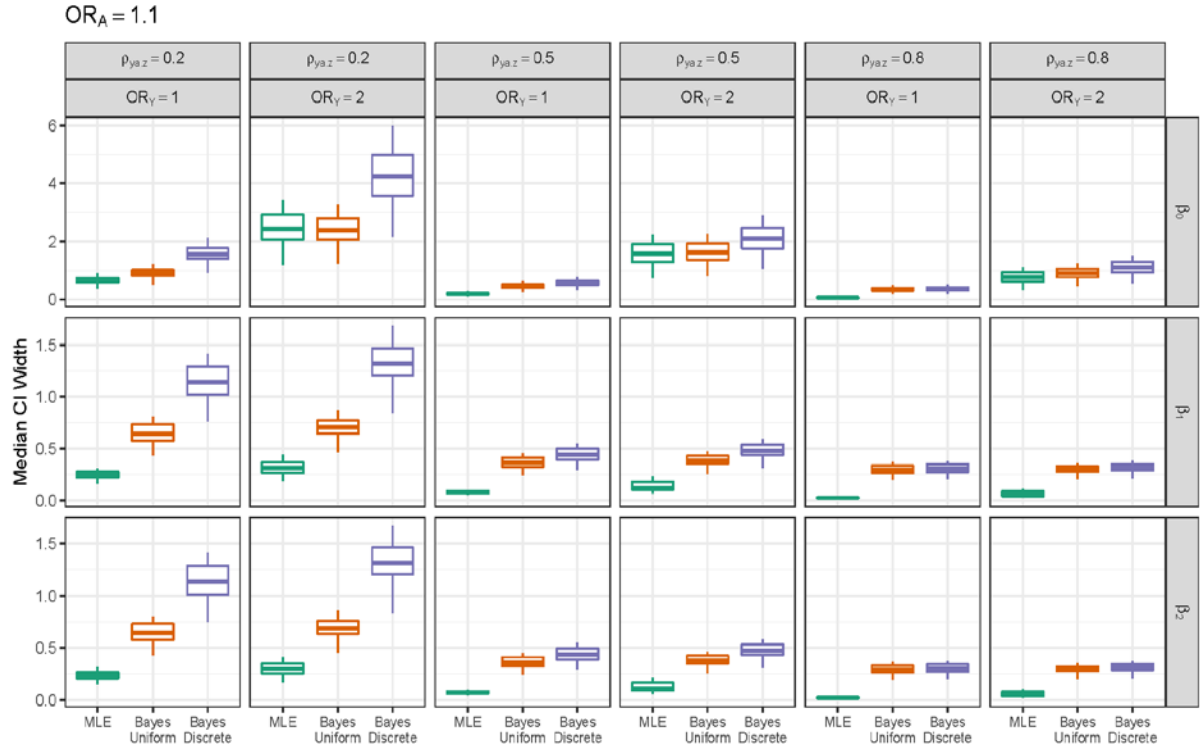
**Figure 3:** Side-by-side box plots presenting distributions of the empirical coverage rates for the alternative intervals. We estimate each coverage rate by computing the interval for each coefficient from 1,000 replicate populations for each combination of data generation model and selection model.  $OR_A$  = odds ratio for  $A$  in the selection model;  $OR_Y$  = odds ratio for  $Y$  in the selection model. The horizontal black line represents 0.95 coverage, for reference.

conditional correlation of the auxiliary proxy  $A$  with  $Y$  becomes 0.5 or higher, good coverage of the actual differences in coefficients is achieved with acceptably narrow intervals for most scenarios. Results in Figure 4 are for an intermediate association of  $A$  with selection; similar patterns were found for other scenarios. Collectively, the results of our simulation study provide support for the fully Bayesian approach with a  $UNIFORM(0,1)$  prior for  $\phi$ .

## 4. Real Data Examples

### 4.1 Polygenic Score-Phenotype Relationships in the Genes for Good Study

Genes for Good (GfG) is a research study based at the University of Michigan that seeks to



**Figure 4:** Side-by-side box plots presenting distributions of the empirical median widths for the alternative intervals across the different scenarios. We obtain the median width by computing the interval for each coefficient from 1,000 replicate populations for each combination of data generation model and selection model.  $OR_A$  = odds ratio for  $A$  in the selection model;  $OR_Y$  = odds ratio for  $Y$  in the selection model.

engage the public in genetic research. Volunteers 18 years of age and above currently living in the United States enroll in the study via a Facebook app, which serves as a tool for them to engage in all aspects of the study (including the answering of health-related survey questions). Volunteers consent to be genotyped and provide saliva samples via mail after answering a minimum number of surveys. Researchers use the resulting genetic profiles to investigate the effects of certain genetic variants on health measures that volunteers self-report via the app. The study is based entirely on volunteers, of which there have been more than 77,000 to date (20,100 of which had been genotyped at the time of this analysis), and

therefore does not have an underlying probability sampling mechanism. One can find additional details on the GfG study at <https://genesforgood.org>.

Polygenic Scores (PGSs), or genetic risk scores (Belsky & Israel, 2014; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), are a quantitative tool for aggregating a large amount of otherwise unwieldy genetic information from genome-wide association studies (GWASs), which are usually based on meta-analyses of non-probability samples of volunteers (Han et al., 2009; Houlston et al., 2010; Lindgren et al., 2009; Nalls et al., 2014; Neale et al., 2010; Sklar et al., 2011). Some researchers have recently expressed the concern that GWASs are vulnerable to selection bias for their target populations (Martin et al., 2019), motivating our current application. For a given phenotype  $p$ , PGSs are generally computed as follows. First, drawing on specific GWASs focusing on that phenotype (e.g., Okbay et al., 2016), “weights” are computed for hundreds of thousands of single nucleotide polymorphisms (SNPs) from individual linear regression models. Each contributing study regresses the value for the phenotype of interest on the coded value for an individual SNP (0, 1, or 2), typically adjusting for cohort-specific covariates. These cohort-specific estimates are then meta-analyzed across all studies. Second, the resulting coefficient from the GWAS meta-analysis, denoted by  $\alpha_{i(p)}$  for SNP  $i$  and phenotype  $p$ , is treated as a weight in computing the PGS. The PGS for phenotype  $p$  for a given individual is then the linear combination of the products of the coded SNP values (denoted  $g_{i(p)}$ ) and the GWAS meta-analysis weights across all SNPs:

$$PGS_p = \sum_i \hat{\alpha}_{i(p)} g_{i(p)} \quad (17)$$

While one generally computes a PGS using (17), a variety of modifications may be used in practice, and this is an active area of methodological research. For example, researchers can

decide whether to use the correlation structure of the human genome to minimize the number of correlated variants in a score, and some researchers may employ  $p$ -value thresholds for identifying which weights are “important” for the computation; see Ware et al. (2017) for an in-depth discussion of these issues.

Underlying this increasingly popular research practice is the assumption that the PGS is in fact a strong correlate of the measures for the phenotype of interest, and this assumption is usually checked with simple regression models for the phenotypes that include the PGS as a predictor. PGSs have been found to be useful correlates of age at onset of alcohol dependence (Kapoor et al., 2016), selected psychiatric traits (Stein et al., 2017; Wray et al., 2014), schizophrenia and bipolar disorder (International Schizophrenia Consortium, 2009; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), and BMI (Locke et al., 2015), among other traits. However, the genetic data used to compute PGSs are generally collected from non-probability samples (usually composed of volunteers), and this raises important questions about whether the estimates of PGS-phenotype relationships are biased for the target population of interest. Recent work has suggested that the predictive ability of PGSs may be limited due to this selection bias (Martin et al., 2019).

We compute PGSs for various phenotypes (e.g., BMI, height, lifetime smoking, college education, etc.) for the 1,829 genotyped GfG participants who were age 50 and above and did not self-identify as Hispanic. Our primary interest is to estimate the relationships of the PGSs (our  $Z$  variables of interest) with their corresponding measured phenotypes (our  $Y$  variables of interest) and to quantify potential selection bias in these estimates. We use the Health and Retirement Study (HRS) as a source of auxiliary information for this target population (see the online supplementary materials for more details regarding the HRS). Applying the

proposed indices of selection bias requires means, variances, and covariances for the covariates of interest  $Z$  and the auxiliary variables  $A$  for the non-selected cases. Because non-probability samples generally have very small selection fractions when considering large target populations, estimates of these required quantities for non-selected cases can be computed from large probability samples that collected the same measures of  $Z$  and  $A$ . We therefore compute these PGSs using *identical SNPs* for both GfG, our non-probability sample, and a benchmark probability sample (HRS) that collected the exact same genetic information in addition to other auxiliary variables  $A$  (e.g., socio-demographics) that are also measured in GfG. See the online supplementary materials for details regarding the common variables available in both the HRS and the GfG study, including the size of the HRS sample and the process used to determine SNPs that were measured in both studies.

Table 1 presents calculations of the MUBNS indices (given that the sampling fraction for GfG is unknown and likely quite small) for the coefficients of five linear regression models fit to the GfG data. For each model, the dependent variable  $Y$  was a given continuous measure (height, BMI) or binary indicator (ever smoked more than 100 cigarettes, college degree [or greater], diabetes). We did not analyze the two binary indicators with the lowest prevalence (coronary artery disease, ever had a heart attack) since the appropriateness of a linear regression model for these indicators was questionable. The corresponding mean-centered PGS was the  $Z$  variable of primary interest in each model, and demographic measures (gender, education, birth cohort, age in years, race, and nativity) along with BMI and height (for all models aside from the BMI and height models) were the auxiliary variables  $A$  used to compute  $X$ . Although the theory presented in this paper assumes that  $Y$  and  $X$  are bivariate normal, we still consider linear probability models for the three binary indicators to assess the performance of the methodology when the  $Y$  variable is clearly non-normal. The “true”

values of the coefficients in each model arise from a fully design-based analysis of the HRS data, incorporating the complex sampling features (including weights) in estimation and variance estimation.

For the fully Bayesian approach to the analysis of the MUBNS indices, we assumed a  $\text{UNIFORM}(0,1)$  prior for  $\phi$  and non-informative Jeffreys' priors for the remaining parameters. We examined the correlation of the medians of the posterior draws of the MUBNS indices for each coefficient with their estimated biases, computed as the differences between the unweighted GfG coefficients and the survey-weighted estimates of the HRS coefficients. We also examined the ability of 95% credible intervals for MUBNS to cover these estimated biases, and whether the intervals suggested a non-zero bias. Recall that the MUBNS index is based on the difference in a coefficient between the selected and non-selected cases. These analyses therefore assume a very small sampling fraction for the GfG cases, in which case the bias of the coefficient for the selected cases would be equal to the difference represented by the MUBNS index.

**Table 1:** Estimates of coefficients in simple linear regression models for two continuous variables (height and BMI) and three binary variables (ever smoke more than 100 cigarettes, college degree, and diabetes) as a function of the PGSs, from GfG (unweighted) and HRS (survey-weighted), in addition to posterior medians and 95% Bayesian credible intervals for the MUBNS index for each coefficient.

	GfG Coef. (SE)	HRS Coef. (SE)	Actual Est. Bias	Median of MUBNS posterior distribution	95% Credible interval for MUBNS	Cor(X,Y Z)
<i>Height</i>						0.733
Intercept	66.07 (0.09)	67.08 (0.09)	-1.01	-2.87	[-2.08, -3.90]	
PGS slope	0.82 (0.09)	0.80 (0.15)	0.02	0.40	[-0.35, 1.24]	
<i>Diabetes</i>						0.324
Intercept	0.16 (0.01)	0.20 (0.01)	-0.04	0.02	[-0.03, 0.13]	
PGS slope	0.03 (0.01)	0.07 (0.01)	-0.04	-0.04	[-0.14, -0.01]	
<i>Ever Smoke</i>						0.219
Intercept	0.62 (0.01)	0.58 (0.02)	0.04	0.00	[-0.20, 0.17]	
PGS slope	0.05 (0.01)	0.01 (0.02)	0.05	0.09	[0.01, 0.46]	
<i>BMI</i>						0.218
Intercept	29.65 (0.16)	29.24 (0.18)	0.41	1.84	[0.37, 9.82]	
PGS slope	1.69 (0.16)	1.79 (0.13)	-0.10	4.95	[0.65, 27.49]	

<i>College Degree</i>						0.192
Intercept	0.51 (0.01)	0.37 (0.02)	0.14	0.28	[0.06, 1.40]	
PGS slope	-0.12 (0.01)	0.09 (0.01)	-0.20	-0.44	[-2.93, -0.05]	

Overall, we see that the estimates of bias in the intercepts and the PGS slopes based on the GfG data (when treating the HRS estimates as truth) are generally small, suggesting that selection bias in the GfG sample is not severe in the cases of these five models. We also note relatively small ( $< 0.3$ ) conditional correlations of  $X$  with  $Y$  (when conditioning on the PGSs) for three of the five models, suggesting limited unique information in the additional auxiliary variables  $A$  considered for these three models. We note that while the credible intervals cover the actual differences in estimated coefficients between the GfG and the HRS in 7 out of 10 cases, this high coverage may be partly due to the wide intervals for the three models associated with the smallest conditional correlations (consistent with our simulation study).

The Pearson correlation of the posterior medians for the MUBNS indices and the bias estimates in Table 1 was 0.56, suggesting that these medians are useful indicators of potential bias, and could be used to order the coefficients in terms of their potential bias. Four GfG estimates present the strongest evidence of selection bias: the intercept in the model for height (corresponding to the expected height for the mean PGS), the PGS slope in the model for diabetes, and both the intercept and PGS slope in the model for the college degree indicator. The Bayesian credible intervals for the MUBNS indices provide correct evidence of a non-zero negative bias in the intercept and zero bias in the PGS slope in the height model. This underscores the importance of informative auxiliary variables for the performance of the indices; note the wide intervals for the MUBNS indices that result from the low conditional correlation in the BMI model. The credible interval for the MUBNS index for the PGS slope in the diabetes model also correctly covers and provides evidence of the non-zero negative bias in the estimate of this slope. Finally, the MUBNS intervals for the

college degree model also provide correct evidence of non-zero positive and negative selection bias in the intercept and slope, respectively, despite the relatively small conditional correlation of  $X$  with  $Y$ .

We remind readers that we only needed sufficient statistics for the non-selected cases (estimated based on the HRS data) to compute the Bayesian intervals, and that data from a large probability sample (e.g., HRS) could in general be employed to generate estimates of these quantities in other applications. Example code used for the calculations in this application is available at <https://github.com/bradytwest/IndicesOfNISB>.

#### ***4.2 Past-year Employment for Smartphone Users with Less Than High School Education in the National Survey of Family Growth***

Little et al. (2019) evaluated the potential non-ignorable selection bias in selected estimates of means based on self-identified smartphone users in the National Survey of Family Growth (NSFG), including the estimated mean number of months worked in the past 12 months for females and males. These authors assumed that smartphone users were a non-probability sample selected from a hypothetical population defined by the full NSFG sample, given that smartphones are now making it easier to participate in opt-in online surveys (e.g., Revilla, 2017). This subsample, however, was a larger fraction of the overall NSFG “population” than would be characteristic of most non-probability samples, given the high penetration of mobile devices in the U.S. (Blumberg and Luke, 2018). A large body of research has established positive correlations between education, current employment status, and income (e.g., Morgan and David, 1963; Muller, 2002). In addition, research suggests that individuals with lower education may be more responsive to surveys inviting sampled persons to participate with some monetary incentive promised in return (e.g., Petrolia and Bhattacharjee, 2009;



Ryu, Couper and Marans, 2005). We therefore focused this application of our proposed measures on smartphone users with less than high school education as a hypothetical non-probability sample with a smaller sampling fraction than reported by Little et al. (2019). We again treated the NSFG sample as the overall population, enabling calculation of the sampling fraction and therefore MUB (as opposed to MUBNS) indices.

Specifically, we sought to fit a linear regression model predicting the number of months worked in the past year as a function of gender (male / female) and age (15-18, 19-29, or 30-49), given the importance of these socio-demographic subgroups in employment research (Mandel and Semyonov, 2014). We fit this linear regression model in the “non-probability sample” defined by smartphone users with less than high school education in the NSFG ( $n = 2,977$ ), and then computed the MUB indices and their intervals for the coefficients estimated from this subsample. We were able to compute the same coefficients for the “non-selected cases” in the remainder of the NSFG sample ( $n = 16,823$ ), enabling validation of the computed MUB indices. Our auxiliary variables in this application included race/ethnicity (non-Hispanic White, non-Hispanic Black, Hispanic, Other), marital status (married, divorced/widowed/separated), household income (<\$19,999, \$20,000-\$59,999, \$60,000+), region of the United States (Midwest, Northeast, South, and West), current employment status (working / not working), and presence of children under the age of 16 in the household (yes, no). These auxiliary variables were not of primary interest like gender and age, but were still thought to be predictive of current employment status when adjusting for gender and age.

Table 2 presents the estimated coefficients in the model fitted to the non-probability sample, the same estimated coefficients in the model fitted to the full NSFG sample, the median of the MUB posterior distribution for each coefficient, and a 95% credible interval for MUB.

**Table 2:** Estimates of coefficients in simple linear regression models for the number of months worked in the past year as a function of gender and age, for both the non-probability sample defined by NSFG respondents who are smartphone users with less than high school education and the full NSFG sample (or “population”), in addition to posterior medians and 95% Bayesian credible intervals for the MUB index for each coefficient.

	Smartphone Users with Less Than HS Educ.: Coef. (SE)	Full NSFG “Population”: Coef. (SE)	Estimated Bias	Median of MUB posterior distribution	95% Credible interval for MUB
Intercept	1.06 (0.13)	2.09 (0.09)	-1.03	-1.20	[-1.87, -0.75]
Male	1.34 (0.16)	1.01 (0.07)	0.33	0.44	[0.16, 0.85]
Age 19-29	5.33 (0.20)	5.64 (0.10)	-0.31	-0.16	[-0.63, 0.24]
Age 30-49	5.75 (0.18)	6.43 (0.09)	-0.68	-0.20	[-0.67, 0.14]

Compared to the population estimates based on the full NSFG sample, the estimates from the hypothetical non-probability sample suggest significantly lower mean past-year employment for younger females (the intercept term in each model). In addition, we see evidence of a larger estimated gap in the mean between males and females based on the non-probability sample, and smaller gaps between age groups 19-29 and 30-49 compared to those who are 15-18. The conditional correlation of the auxiliary proxy defined by  $X$  with number of months worked in the past year in this example was 0.692, which was nearly as high as that found for the height variable in the Genes for Good application. In this context, the posterior MUB medians had a high correlation with the actual differences in the coefficients between the selected and non-selected cases, and the 95% credible intervals for the MUB indices covered or nearly covered the actual differences in the coefficients between the non-probability sample and the full population without having extreme widths.

## 5. Discussion

We have addressed an important gap in the literature by developing model-based indices of selection bias for regression coefficients estimated from non-probability samples, and evaluated the utility of these indices in different settings. Simulation studies and applications

of the proposed measures to real data sets suggest that the indices are effective when informative auxiliary variables are available, especially the Bayesian version of the approach that takes into account uncertainty in the regression parameters for selected cases and non-selected cases. As Little et al. (2019) noted, quantifying non-ignorable selection bias for survey means and proportions may not be possible without access to informative auxiliary variables for the larger population. The same caveat applies to assessing selection bias in regression coefficients, with “informative” now meaning predictive of the outcome after conditioning on the covariates in the substantive model. Without such auxiliary variables, no method is likely to be effective.

Collectively, our simulation study and our applications provide important recommendations for practice when applying these indices to assess potential selection bias in regression coefficients estimated from non-probability samples:

1. Identify good auxiliary predictors of the outcome variable in the model of interest, such that the correlation between a linear predictor of the outcome based on these auxiliary predictors and the outcome itself is moderate to high after conditioning on the primary predictor(s) of interest (e.g., the models for height and number of months worked in the past year in our applications);
2. If this conditional correlation is moderate to high, apply a fully Bayesian approach to form a credible interval for the measure of selection bias, namely MUB if the selection fraction is non-negligible and is known or can be estimated, and MUBNS otherwise; and
3. If this conditional correlation is low, the Bayesian credible intervals for the selection bias may become wide, reflecting the limited information available in the auxiliary variables used to form  $X$ .

We have provided code for computing the proposed MUBNS and MUB indices and forming both types of intervals at <https://github.com/bradytwest/IndicesOfNISB>.

An important limitation of our method is that it requires individual records or summary measures, in the population or in a large random sample of the population, for auxiliary variables that are at least moderately predictive of the outcome  $Y$ , after adjusting for the covariates in the target regression model. However, we believe that such variables are necessary for any credible method for measuring selection bias. Public-use data files from large survey programs employing national probability samples, such as the HRS, provide good potential sources of this type of information.

This work has important implications for other studies in a variety of disciplines that are employing so-called big data, large volunteer samples, or convenience samples to make statements about relationships between variables in target populations, especially concerning genetics and genomics. In these situations, investigators do not have control over the selection mechanism that is generating the data. The indices proposed here can be used to assess the potential for selection bias in the estimated regression coefficients in such settings.

We employed a simple formulation of the polygenic score in our first application. Although commonly used in modern genetic research, PGSs have been criticized both from methodological and ethical angles. From a methodological perspective, missing heritability (differences in explained variability of disease occurrence between PGS and family studies) is a major limitation of the approach (Dudbridge, 2016; Wray et al., 2013). Even when including multiple genetic variables, the predictive power of PGSs is still very low and outperformed by simpler methods like family history (Dudbridge, 2016; Khoury, Janssens,

and Ransohoff, 2013). Furthermore, SNPs included in the PGSs are often chosen using discovery thresholds based on  $p$ -values, which are known for their far-reaching limitations (Dudbridge, 2016; Maher, 2015; Wray et al., 2013), and final PGSs are obtained using somewhat arbitrary weighting of the SNPs (Maher, 2015). Another major critique of PGSs is the lack of representation of subjects with non-European ancestry (Lewis & Vassos, 2017; Torkamani, Wineinger & Topol, 2018). European ancestry subjects make up about 79% of all subjects in genetic studies, while this group represents 16% of the world's population. This disparity is expected to exacerbate existing health access disparities, given that methods are being developed for a population that already has better access to health services (Martin et al., 2019). The measures described in the present study will enable researchers to gauge potential selection biases in studies involving PGSs as predictors of other health outcomes.

Finally, future work in this area needs to extend the developments in this study to generalized linear models (e.g., logistic regression). This would benefit studies where the dependent variables are not necessarily continuous and/or normally distributed.

## References

- Andridge, R.R., West, B.T., Little, R.J.A, Boonstra, P.S., & Alvarado-Leiton, F. (2019). Indices of Non-Ignorable Selection Bias for Proportions Estimated from Non-Probability Samples. *Journal of the Royal Statistical Society, Series C*. DOI: 10.1111/rssc.12371.
- Andridge, R. R., & Little, R. J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27(2), 153-180.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., ... & Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1(2), 90-143.

Belsky, D.W., & Israel, S. (2014). Integrating genetics and social science: Genetic risk scores. *Biodemography and Social Biology*, 60(2), 137-155.

Blumberg, S., Luke, J. (2018). Wireless Substitution: Early Release of Estimates From the National Health Interview Survey, January–June 2018. Accessed at <https://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201812.pdf>.

Boonstra, P.S., Andridge, R.R., West, B.T., Little, R.J.A., & Alvarado-Leiton, F. (*Under Review*, 2019). A simulation study of diagnostics for bias in non-probability samples. *Submitted to Journal of Official Statistics*, July 2019.

Brick, J.M., & Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *ANNALS of the American Academy of Political and Social Science*, 645(1), 36-59.

de Leeuw, E., Hox, J. & Luiten, A. (2018). International Nonresponse Trends across Countries and Years: An analysis of 36 years of Labour Force Survey data. *Survey Insights: Methods from the Field*. Retrieved from <https://surveyinsights.org/?p=10452>.

Dudbridge, F. (2016). Polygenic epidemiology. *Genetic epidemiology*, 40(4), 268-272.

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

Han, J. W., Zheng, H. F., Cui, Y., Sun, L. D., Ye, D. Q., Hu, Z., ... Zhang, X.J. (2009). Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nature Genetics*, 41(11), 1234-1239.

Houlston, R. S., Cheadle, J., Dobbins, S. E., Tenesa, A., Jones, A. M., Howarth, K., ... Tomlinson, I.P.M. (2010). Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nature Genetics*, 42(11), 973-979.

International Schizophrenia Consortium. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256), 748-752.

Kapoor, M., Chou, Y. L., Edenberg, H. J., Foroud, T., Martin, N. G., Madden, P. A. F., ... Agrawal, A. (2016). Genome-wide polygenic scores for age at onset of alcohol dependence and association with alcohol-related measures. *Translational psychiatry*, 6(3), e761.

Khoury, M. J., Janssens, A. C. J., & Ransohoff, D. F. (2013). How can polygenic inheritance be used in population screening for common diseases?. *Genetics in Medicine*, 15(6), 437-443.

Lewis, C. M., & Vassos, E. (2017). Prospects for using risk scores in polygenic medicine. *Genome medicine*, 9, 96.

- Lindgren, C. M., Heid, I. M., Randall, J. C., Lamina, C., Steinthorsdottir, V., Qi, L., ... Jackson, A. U. (2009). Correction: Genome-Wide Association Scan Meta-Analysis Identifies Three Loci Influencing Adiposity and Fat Distribution. *PLoS genetics*, 5(6), e1000508.
- Little, R.J.A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3), 471-483.
- Little, R.J.A., West, B.T., Boonstra, P., & Hu, J. (2019). Measures of the Degree of Departure from Ignorable Sample Selection. *Journal of Survey Statistics and Methodology*. DOI: 10.1093/jssam/smz023.
- Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., ... Speliotes, E.K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538), 197-206.
- Maher, B. S. (2015). Polygenic scores in epidemiology: risk prediction, etiology, and clinical utility. *Current epidemiology reports*, 2(4), 239-244.
- Mandel, H. & Semyonov, M. (2014). Gender pay gap and employment sector: Sources of earnings disparities in the United States, 1970-2010. *Demography*, 51(5), 1597-1618.
- Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51, 584-591.



- Morgan, J., & David, M. (1963). Education and income. *The Quarterly Journal of Economics*, 77(3), 423-437.
- Muller, A. (2002). Education, income inequality, and mortality: a multiple regression analysis. *British Medical Journal*, 324(7328), 23-25.
- Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., ... Singleton, A.B. (2014). Large scale meta analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature Genetics*, 46(9), 989-993.
- Neale, B. M., Medland, S. E., Ripke, S., Asherson, P., Franke, B., Lesch, K. P., ... Daly, M. (2010). Meta-Analysis of Genome-Wide Association Studies of Attention-Deficit / Hyperactivity Disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 49(9), 884-897.
- Nishimura, R., Wagner, J., & Elliott, M. (2016). Alternative indicators for the risk of non-response bias: a simulation study. *International Statistical Review*, 84(1), 43-62.
- Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., ... Oskarsson, S. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, 533(7604), 539.
- Petrolia, D. R., & Bhattacharjee, S. (2009). Revisiting incentive effects: evidence from a random-sample mail survey on consumer preferences for fuel ethanol. *Public Opinion Quarterly*, 73(3), 537-550.

Presser, S. & McCulloch, S. (2011). The Growth of Survey Research in the United States: Government-Sponsored Surveys, 1984–2004. *Social Science Research*, 40(4), 1019-1024.

Revilla, M. (2017). Analyzing survey characteristics, participation, and evaluation across 186 surveys in an online opt-in panel in Spain. *Methods, data, analyses*, 11(2), 28.

Rubin, D.B. (1976). Inference and Missing Data (w/ Discussion). *Biometrika*, 63(3), 581-592.

Ryu, E., Couper, M. P., & Marans, R. W. (2005). Survey incentives: Cash vs. in-kind; face-to-face vs. mail; response rate vs. nonresponse error. *International Journal of Public Opinion Research*, 18(1), 89-106.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 511(7510), 421-427.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the Representativeness of Survey Response. *Survey Methodology*, 35(1), 101-113.

Sklar, P., Ripke, S., Scott, L. J., Andreassen, O. A., Cichon, S., Craddock, N., ... Corvin, A. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature genetics*, 43(10), 977-983.

Stein, M. B., Ware, E. B., Mitchell, C., Chen, C. Y., Borja, S., Cai, T., ... Jain, S. (2017).

Genomewide association studies of suicide attempts in US soldiers. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 174(8), 786-797.

Torkamani, A., Wineinger, N. E., & Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9), 581-590.

Valliant, R. (2019). Comparing Alternatives for Estimation from Nonprobability Samples. *Journal of Survey Statistics and Methodology*. DOI: 10.1093/jssam/smz003.

Ware, E. B., Schmitz, L. L., Faul, J. D., Gard, A., Mitchell, C., Smith, J. A., & Kardina, S. L. (2017). Heterogeneity in polygenic scores for common human traits. *bioRxiv*, <https://www.biorxiv.org/content/early/2017/02/05/106062>.

Williams, D. & Brick, M. (2018). Trends in U.S. Face-to-Face Household Survey Nonresponse and Level of Effort. *Journal of Survey Statistics and Methodology*, 6(2), 186-211.

Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A., Dudbridge, F., & Middeldorp, C. M. (2014). Research review: polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, 55(10), 1068-1087.

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., & Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, 14(7), 507-515.