Applying Non-ignorable Missing Data Methods to U.S. Election Polling Data

Rebecca Andridge

The Ohio State University College of Public Health

May 26, 2022

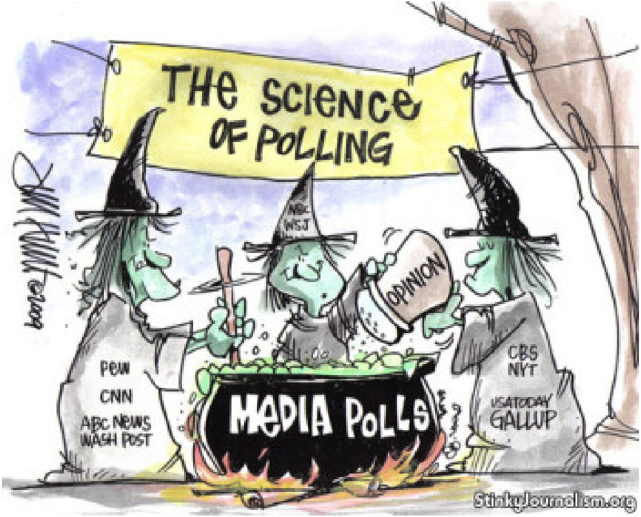
Joint work with Brady West (University of Michigan)

Based on prior work with Rod Little, Phil Boonstra, and Fernanda Alvarado-Leiton (University of Michigan)

Outline

Pre-election polling has had some negative press lately. . .

Pre-election polling has had some negative press lately. . .



Pre-election polling has had some negative press lately. . .

The Problem with Pre-Election Polls

- 2020 U.S. presidential polls had highest error in 40 years – a "failure"
- Many issues from 2016 do not appear to be the problem
    - Late deciders / Changes in voting intention – not an issue in 2020 (early voting helped)
    - Failing to account for educational differences when reweighting for nonresponse/noncoverage – done for most state-level 2020 polls

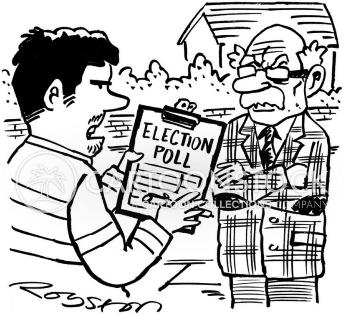
The Problem with Pre-Election Polls

- 2020 U.S. presidential polls had highest error in 40 years – a “failure”
- Many issues from 2016 do not appear to be the problem
  - Late deciders / Changes in voting intention – not an issue in 2020 (early voting helped)
  - Failing to account for educational differences when reweighting for nonresponse/noncoverage – done for most state-level 2020 polls
- Typical polls, though probability samples, have very low response rates (e.g., 4.5-6.5%)
- Weighting adjustments assume selection/response is at random, conditional on the variables used to compute the weights

The Problem with Pre-Election Polls

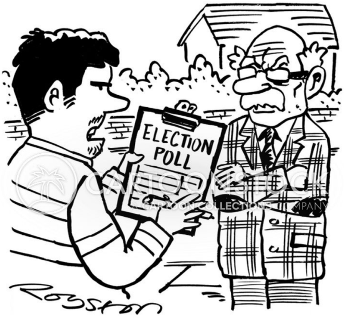
- 2020 U.S. presidential polls had highest error in 40 years – a "failure"
- Many issues from 2016 do not appear to be the problem
  - Late deciders / Changes in voting intention – not an issue in 2020 (early voting helped)
  - Failing to account for educational differences when reweighting for nonresponse/noncoverage – done for most state-level 2020 polls
- Typical polls, though probability samples, have very low response rates (e.g., 4.5-6.5%)
- Weighting adjustments assume selection/response is at random, conditional on the variables used to compute the weights
- But. . . in 2020 might Trump supporters have been likely to answer a pre-election poll, even conditional on demographic characteristics?

The Problem with Pre-Election Polls



"Should I put **genuinely** undecided or **damned if I'm telling you** undecided?"

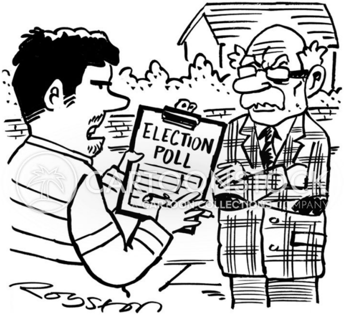
The Problem with Pre-Election Polls



"Should I put **genuinely** undecided or **damned if I'm telling you** undecided?"

AAPOR Task Force report on 2020 Pre-Election Polling conclusion:
"The Democrats/Republicans who responded had different opinions than those who did not (within-party nonresponse)" (AAPOR 2020, p.71)

The Problem with Pre-Election Polls



"Should I put **genuinely** undecided or **damned if I'm telling you** undecided?"

AAPOR Task Force report on 2020 Pre-Election Polling conclusion:
"The Democrats/Republicans who responded had different opinions than those who did not (within-party nonresponse)" (AAPOR 2020, p.71)

Non-ignorable missing data / sample selection!

Problem Statement

Goal: Estimate population proportion from non-probability sample (or probability sample with low response rate)
$\rightarrow$ *Proportion voting for Trump*

Problem Statement

Goal: Estimate population proportion from non-probability sample (or probability sample with low response rate)
$\rightarrow$ *Proportion voting for Trump*

Problem: Potential for selection bias due to non-ignorable selection/nonresponse mechanisms

- Ignorable: probability of selection depends on *observed characteristics*
- Non-ignorable: probability of selection depends on *unobserved characteristics*

$\rightarrow$ *Response to poll might depend on candidate preference*

Problem Statement

Goal: Estimate population proportion from non-probability sample (or probability sample with low response rate)
$\rightarrow$ *Proportion voting for Trump*

Problem: Potential for selection bias due to non-ignorable selection/nonresponse mechanisms

- Ignorable: probability of selection depends on *observed characteristics*
- Non-ignorable: probability of selection depends on *unobserved characteristics*

$\rightarrow$ *Response to poll might depend on candidate preference*

Approach: Use a model-based **index of selection bias**, $MUBP(\phi)$, that allows assessment of potential selection bias in proportion estimates (Andridge et al. 2019)
$\rightarrow$ *Sensitivity analysis allowing non-ignorable selection*

Definitions/Notation

Notation:

- $Y = (y_1, \ldots, y_N)$ = survey data for each unit in pop. $i = 1, \ldots, N$
  - $Y = (Y_{inc}, Y_{exc})$ for units **inc**luded, **exc**luded from sample
- Z = set of fully observed auxiliary or design variables (known for units both in and out of the sample)
- $S = (S_1, \ldots, S_N)$ = selection indicator

Definitions/Notation

Notation:

- $Y = (y_1, \ldots, y_N)$ = survey data for each unit in pop. $i = 1, \ldots, N$
  - $Y = (Y_{inc}, Y_{exc})$ for units **inc**luded, **exc**luded from sample
- Z = set of fully observed auxiliary or design variables (known for units both in and out of the sample)
- $S = (S_1, \ldots, S_N)$ = selection indicator

Joint distribution:

$$f_{Y,S}(Y, S|Z, \theta, \xi) = \overbrace{f_Y(Y|Z, \theta)}^{\text{inference target}} \underbrace{f_{S|Y}(S|Y, Z, \xi)}_{\text{selection mechanism}}$$

Definitions/Notation

Probability sampling = "extremely" **ignorable** selection

- Selection may depend on Z but not Y (Y_{inc} or Y_{exc})
- Inclusion in sample is independent of Y and any unobserved variables
- $f_{S|Y}(S|Y, Z, \xi) = f_{S|Y}(S|Z)$ (no ξ!)
- Thus inference for θ can *ignore* distribution of S...

Definitions/Notation

Probability sampling = "extremely" **ignorable** selection

- Selection may depend on Z but not Y (Y_{inc} or Y_{exc})
- Inclusion in sample is independent of Y and any unobserved variables
- $f_{S|Y}(S|Y, Z, \xi) = f_{S|Y}(S|Z)$ (no ξ!)
- Thus inference for θ can *ignore* distribution of S...

  if there is no nonresponse!

Definitions/Notation

Probability sampling = "extremely" **ignorable** selection

- Selection may depend on Z but not Y (Y_{inc} or Y_{exc})
- Inclusion in sample is independent of Y and any unobserved variables
- $f_{S|Y}(S|Y,Z,\xi) = f_{S|Y}(S|Z)$ (no ξ!)
- Thus inference for θ can *ignore* distribution of S...

  if there is no nonresponse!

Non-probability sampling[1] = might be **non-ignorable** selection

- Selection may depend on Y_{exc}, i.e., something unobserved
- $f_{S|Y}(S|Y,Z,\xi)$ *necessary* for inference about θ
- Hard (impossible?) to model S – can we quantify the potential **selection bias** arising from ignoring the selection mechanism?

[1] or probability sample with nonresponse

Previous Work

Some methods exist for attempting to assess a sample's representativeness (and thus hint at **selection bias**)

Previous Work

Some methods exist for attempting to assess a sample's representativeness (and thus hint at **selection bias**)

- **R-indicator** – function of response propensities; agnostic about the survey variables of interest (Schouten et al. 2009)

Previous Work

Some methods exist for attempting to assess a sample's representativeness (and thus hint at **selection bias**)

- **R-indicator** – function of response propensities; agnostic about the survey variables of interest (Schouten et al. 2009)
- **H1 indicator** – based on survey variables of interest, but assumes ignorable selection mechanism (Särndal and Lundstrom 2010)
  - Assumes $f_{S|Y}(S|Y, Z, \xi) = f_{S|Y}(S|Y_{inc}, Z, \xi)$
  - Not as "extremely" ignorable as probability sampling, but still ignorable
  - Do not need to specify distribution for S for inference about θ

Previous Work

Some methods exist for attempting to assess a sample's representativeness (and thus hint at **selection bias**)

- **R-indicator** – function of response propensities; agnostic about the survey variables of interest (Schouten et al. 2009)
- **H1 indicator** – based on survey variables of interest, but assumes ignorable selection mechanism (Särndal and Lundstrom 2010)
  - Assumes $f_{S|Y}(S|Y,Z,\xi) = f_{S|Y}(S|Y_{inc},Z,\xi)$
  - Not as "extremely" ignorable as probability sampling, but still ignorable
  - Do not need to specify distribution for S for inference about θ
- **SMUB(ϕ)** – newly proposed index allowing for non-ignorable selection; provides range of potential selection bias for estimating means (continuous Y) (Little et al. 2020)

Previous Work

Some methods exist for attempting to assess a sample's representativeness (and thus hint at **selection bias**)

- **R-indicator** – function of response propensities; agnostic about the survey variables of interest (Schouten et al. 2009)
- **H1 indicator** – based on survey variables of interest, but assumes ignorable selection mechanism (Särndal and Lundstrom 2010)
  - Assumes $f_{S|Y}(S|Y, Z, \xi) = f_{S|Y}(S|Y_{inc}, Z, \xi)$
  - Not as "extremely" ignorable as probability sampling, but still ignorable
  - Do not need to specify distribution for S for inference about θ
- **SMUB(ϕ)** – newly proposed index allowing for non-ignorable selection; provides range of potential selection bias for estimating means (continuous Y) (Little et al. 2020)

SMUB(ϕ) close to what we want – but for proportions

Outline

Illustrative Example: National Survey of Family Growth

- (Fake) Population = entire NSFG sample $(N = 19,800)$
- Selected sample = all smartphone users $(n = 15,923)$
  - Note high selection fraction (≈80%) – atypical for non-prob sample
- Outcome of interest = Never married (by gender[2])

[2] Note: NSFG only captures gender as a binary variable

Illustrative Example: National Survey of Family Growth

- (Fake) Population = entire NSFG sample $(N = 19,800)$
- Selected sample = all smartphone users $(n = 15,923)$
  - Note high selection fraction (≈80%) – atypical for non-prob sample
- Outcome of interest = Never married (by gender[2])
- We know the true selection bias in this artificial example

	Females	**Males**
Population proportion	0.468	0.566
Selected sample proportion	0.466	0.555
True bias	−0.002	−0.011

[2] Note: NSFG only captures gender as a binary variable

Illustrative Example: National Survey of Family Growth

- (Fake) Population = entire NSFG sample $(N = 19,800)$
- Selected sample = all smartphone users $(n = 15,923)$
  - Note high selection fraction (≈80%) – atypical for non-prob sample
- Outcome of interest = Never married (by gender[2])
- We know the true selection bias in this artificial example

	Females	**Males**
Population proportion	0.468	0.566
Selected sample proportion	0.466	0.555
True bias	−0.002	−0.011
Manski bounds* of bias	(−0.098, 0.085)	(−0.094, 0.118)

**assume all non-selected are 1s, all non-selected are 0s*

[2] Note: NSFG only captures gender as a binary variable

Illustrative Example: National Survey of Family Growth

- (Fake) Population = entire NSFG sample $(N = 19,800)$
- Selected sample = all smartphone users $(n = 15,923)$
  - Note high selection fraction (≈80%) – atypical for non-prob sample
- Outcome of interest = Never married (by gender[2])
- We know the true selection bias in this artificial example

	Females	**Males**
Population proportion	0.468	0.566
Selected sample proportion	0.466	0.555
True bias	−0.002	−0.011
Manski bounds* of bias	(−0.098, 0.085)	(−0.094, 0.118)

**assume all non-selected are 1s, all non-selected are 0s*

- Can we do better than the Manski bounds?

[2] Note: NSFG only captures gender as a binary variable

Available Data

- Assume we have microdata for selected cases:
  - Y = binary variable of interest = never married
  - Z = auxiliary variables = age, race, education, etc.

Available Data

- Assume we have microdata for selected cases:
  - Y = binary variable of interest = never married
  - Z = auxiliary variables = age, race, education, etc.
- Assume we have summary statistics on Z for non-selected cases
  - Mean (vector) and Variance (matrix) of Z
  - In practice, could come from Census, large probability sample, etc.
  - If instead we have summary statistics of Z for population, could "back-out" the non-selected mean/variance
  - If we don't have variance, could assume it's the same as among selected cases

Outline

Index of Selection Bias: $MUBP(\phi)$

Measure of Unadjusted Bias for a Proportion, MUBP(ϕ)

- Extension of **SMUB(ϕ)** of Little et al. (2020) (for means) to binary Y (proportions) (Andridge et al. 2019)
  - Based on pattern-mixture models
  - Makes explicit assumption(s) about distribution of S
  - Provides sensitivity analysis to assess range of bias under different assumptions about S

Index of Selection Bias: $MUBP(\phi)$

Measure of Unadjusted Bias for a Proportion, MUBP(ϕ)

- Extension of **SMUB(ϕ)** of Little et al. (2020) (for means) to binary Y (proportions) (Andridge et al. 2019)
  - Based on pattern-mixture models
  - Makes explicit assumption(s) about distribution of S
  - Provides sensitivity analysis to assess range of bias under different assumptions about S
- Basic idea:
  - We can measure the degree of selection bias present in Z

Index of Selection Bias: $MUBP(\phi)$

Measure of Unadjusted Bias for a Proportion, MUBP(ϕ)

- Extension of **SMUB(ϕ)** of Little et al. (2020) (for means) to binary Y (proportions) (Andridge et al. 2019)
  - Based on pattern-mixture models
  - Makes explicit assumption(s) about distribution of S
  - Provides sensitivity analysis to assess range of bias under different assumptions about S
- Basic idea:
  - We can measure the degree of selection bias present in Z
  - If Y is correlated with Z, then this tells you something about the potential selection bias in Y

Index of Selection Bias: $MUBP(\phi)$

Measure of Unadjusted Bias for a Proportion, MUBP(ϕ)

- Extension of **SMUB(ϕ)** of Little et al. (2020) (for means) to binary Y (proportions) (Andridge et al. 2019)
  - Based on pattern-mixture models
  - Makes explicit assumption(s) about distribution of S
  - Provides sensitivity analysis to assess range of bias under different assumptions about S
- Basic idea:
  - We can measure the degree of selection bias present in Z
  - If Y is correlated with Z, then this tells you something about the potential selection bias in Y
  - Use pattern-mixture models to explicitly model non-ignorable selection (i.e., selection dependent on Y)

$MUBP(\phi)$: Theory

- Y = binary variable of interest, only available for selected sample
  - Woman (Man) has never been married
- Z = auxiliary variables, available for selected cases and in aggregate for non-selected sample
  - Age, race, education, marital status, region, income, kids in HH

$MUBP(\phi)$: Theory

- Y = binary variable of interest, only available for selected sample
  - Woman (Man) has never been married
- Z = auxiliary variables, available for selected cases and in aggregate for non-selected sample
  - Age, race, education, marital status, region, income, kids in HH
- U = underlying normally distributed unobserved latent variable
  - $Y = 1$ when $U > 0$

$MUBP(\phi)$: Theory

- Y = binary variable of interest, only available for selected sample
  - Woman (Man) has never been married
- Z = auxiliary variables, available for selected cases and in aggregate for non-selected sample
  - Age, race, education, marital status, region, income, kids in HH
- U = underlying normally distributed unobserved latent variable
  - $Y = 1$ when $U > 0$
- X = "proxy" for Y
  - Constructed from probit regression of Y on Z for selected cases (linear predictor from the regression)
  - Available for selected cases and in aggregate for non-selected sample

$MUBP(\phi)$: Theory

- Y = binary variable of interest, only available for selected sample
  - Woman (Man) has never been married
- Z = auxiliary variables, available for selected cases and in aggregate for non-selected sample
  - Age, race, education, marital status, region, income, kids in HH
- U = underlying normally distributed unobserved latent variable
  - $Y = 1$ when $U > 0$
- X = "proxy" for Y
  - Constructed from probit regression of Y on Z for selected cases (linear predictor from the regression)
  - Available for selected cases and in aggregate for non-selected sample
- S = selection indicator (i.e., $S = 1$ for smartphone users)
- V = other covariates, independent of Y and X (may be related to S)

$MUBP(\phi)$: Theory

- Assume a proxy pattern-mixture model [3] for U and X given S:

$$(U, X|S = j) \sim N_2 \left(\begin{bmatrix} \mu_u^{(j)} \\ \mu_x^{(j)} \end{bmatrix}, \begin{bmatrix} \sigma_{uu}^{(j)} & \rho_{ux}^{(j)} \sqrt{\sigma_{uu}^{(j)} \sigma_{xx}^{(j)}} \\ \rho_{ux}^{(j)} \sqrt{\sigma_{uu}^{(j)} \sigma_{xx}^{(j)}} & \sigma_{xx}^{(j)} \end{bmatrix} \right)$$

$$S \sim \mathrm{Bernoulli}(\pi)$$

[3] Andridge and Little 2011,2020

$MUBP(\phi)$: Theory

- Assume a proxy pattern-mixture model [3] for U and X given S:

$$(U, X|S = j) \sim N_2\left(\begin{bmatrix}\mu_u^{(j)}\\ \mu_x^{(j)}\end{bmatrix}, \begin{bmatrix}\sigma_{uu}^{(j)} & \rho_{ux}^{(j)}\sqrt{\sigma_{uu}^{(j)}\sigma_{xx}^{(j)}}\\ \rho_{ux}^{(j)}\sqrt{\sigma_{uu}^{(j)}\sigma_{xx}^{(j)}} & \sigma_{xx}^{(j)}\end{bmatrix}\right)$$

$$S \sim \mathrm{Bernoulli}(\pi)$$

- To identify this model, assume selection into the sample is a function of V and a linear combination of X and U:

$$\Pr(S = 1|U, X, V) = f((1-\phi)X^* + \phi U, V)$$

  - $\phi \in [0, 1]$ is a sensitivity parameter (no info in data about it)
  - $X^* = X\sqrt{\sigma_{uu}^{(1)}/\sigma_{xx}^{(1)}}$ = rescaled proxy X

[3] Andridge and Little 2011,2020

$MUBP(\phi)$: Theory

$$(U, X|S = j) \sim N_2\left(\begin{bmatrix}\mu_u^{(j)}\\ \mu_x^{(j)}\end{bmatrix}, \begin{bmatrix}\sigma_{uu}^{(j)} & \rho_{ux}^{(j)}\sqrt{\sigma_{uu}^{(j)}\sigma_{xx}^{(j)}}\\ \rho_{ux}^{(j)}\sqrt{\sigma_{uu}^{(j)}\sigma_{xx}^{(j)}} & \sigma_{xx}^{(j)}\end{bmatrix}\right)$$

$$S \sim \text{Bernoulli}(\pi)$$

- WLOG set $\sigma_{uu}^{(1)} = 1$ (latent variable scale)

$MUBP(\phi)$: Theory

$$(U, X|S = j) \sim N_2 \left(\begin{bmatrix} \mu_u^{(j)} \\ \mu_x^{(j)} \end{bmatrix}, \begin{bmatrix} \sigma_{uu}^{(j)} & \rho_{ux}^{(j)} \sqrt{\sigma_{uu}^{(j)} \sigma_{xx}^{(j)}} \\ \rho_{ux}^{(j)} \sqrt{\sigma_{uu}^{(j)} \sigma_{xx}^{(j)}} & \sigma_{xx}^{(j)} \end{bmatrix} \right)$$

$$S \sim \text{Bernoulli}(\pi)$$

- WLOG set $\sigma_{uu}^{(1)} = 1$ (latent variable scale)
- Marginal mean of Y is target of inference:

$$\mu_y = \Pr(Y = 1) = \Pr(U > 0) = \pi \underbrace{\Phi\left(\mu_u^{(1)}\right)}_{\text{sel. prop.}} + (1 - \pi) \underbrace{\Phi\left(\mu_u^{(0)} / \sqrt{\sigma_{uu}^{(0)}}\right)}_{\text{non-sel. prop.}}$$

$MUBP(\phi)$: Theory

$$(U, X|S=j) \sim N_2\left(\begin{bmatrix}\mu_u^{(j)} \\ \mu_x^{(j)}\end{bmatrix}, \begin{bmatrix}\sigma_{uu}^{(j)} & \rho_{ux}^{(j)}\sqrt{\sigma_{uu}^{(j)}\sigma_{xx}^{(j)}} \\ \rho_{ux}^{(j)}\sqrt{\sigma_{uu}^{(j)}\sigma_{xx}^{(j)}} & \sigma_{xx}^{(j)}\end{bmatrix}\right)$$

$$S \sim \text{Bernoulli}(\pi)$$

- WLOG set $\sigma_{uu}^{(1)} = 1$ (latent variable scale)
- Marginal mean of Y is target of inference:

$$\mu_y = \Pr(Y=1) = \Pr(U>0) = \pi \underbrace{\Phi\left(\mu_u^{(1)}\right)}_{\text{sel. prop.}} + (1-\pi) \underbrace{\Phi\left(\mu_u^{(0)}/\sqrt{\sigma_{uu}^{(0)}}\right)}_{\text{non-sel. prop.}}$$

- Key parameter: $\rho_{ux}^{(j)}$ = biserial correlation of binary Y and X
  - Quantifies how related Y and X (Z) are
  - Can estimate $\rho_{ux}^{(1)}$ using selected sample

$MUBP(\phi)$: Theory

- Non-identifiable parameters of pattern-mixture model $\left\{\mu_u^{(0)}, \sigma_{uu}^{(0)}, \rho_{ux}^{(0)}\right\}$ are just identified by selection mechanism assumption

$$\Pr(S = 1|U, X, V) = f((1-\phi)X^* + \phi U, V)$$

$MUBP(\phi)$: Theory

- Non-identifiable parameters of pattern-mixture model $\left\{\mu_u^{(0)}, \sigma_{uu}^{(0)}, \rho_{ux}^{(0)}\right\}$ are just identified by selection mechanism assumption

$$\Pr(S = 1|U, X, V) = f((1 - \phi)X^* + \phi U, V)$$

- Selected value of sensitivity parameter ϕ determines selection mechanism:
  - $\phi = 0 \rightarrow \Pr(S = 1|U, X, V) = f(X^*, V)$
    - **Ignorable selection**
    - Only depends on observed X and V (not U or Y)

$MUBP(\phi)$: Theory

- Non-identifiable parameters of pattern-mixture model $\left\{\mu_u^{(0)}, \sigma_{uu}^{(0)}, \rho_{ux}^{(0)}\right\}$ are just identified by selection mechanism assumption

$$\Pr(S = 1 | U, X, V) = f((1-\phi)X^* + \phi U, V)$$

- Selected value of sensitivity parameter ϕ determines selection mechanism:
  - $\phi = 0 \rightarrow \Pr(S = 1 | U, X, V) = f(X^*, V)$
    - **Ignorable selection**
    - Only depends on observed X and V (not U or Y)
  - $\phi = 1 \rightarrow \Pr(S = 1 | U, X, V) = f(U, V)$
    - **"Extremely" Non-ignorable selection**
    - Depends entirely on unobserved U (and thus Y) and V (not X)

$MUBP(\phi)$: Theory

- Non-identifiable parameters of pattern-mixture model $\left\{\mu_u^{(0)}, \sigma_{uu}^{(0)}, \rho_{ux}^{(0)}\right\}$ are just identified by selection mechanism assumption

$$\Pr(S = 1|U, X, V) = f((1-\phi)X^* + \phi U, V)$$

- Selected value of sensitivity parameter ϕ determines selection mechanism:
  - $\phi = 0 \rightarrow \Pr(S = 1|U, X, V) = f(X^*, V)$
    - **Ignorable selection**
    - Only depends on observed X and V (not U or Y)
  - $\phi = 1 \rightarrow \Pr(S = 1|U, X, V) = f(U, V)$
    - **"Extremely" Non-ignorable selection**
    - Depends entirely on unobserved U (and thus Y) and V (not X)
  - $0 < \phi < 1 \rightarrow \Pr(S = 1|U, X, V) = f((1-\phi)X^* + \phi U, V)$
    - **Non-ignorable selection**
    - Depends (at least) partially on unobserved U (and thus Y) and V

$MUBP(\phi)$: Theory

- For a specified ϕ we can estimate μ_y and compare to selected sample proportion $\hat{\mu}_y^{(1)}$ to obtain a **Measure of Unadjusted Selection Bias for a Proportion**:

$$\boldsymbol{MUBP(\phi)} = \hat{\mu}_y^{(1)} - \hat{\mu}_y^{(\phi)}$$

where $\hat{\mu}_y$ depends on chosen ϕ

$MUBP(\phi)$: Theory

- For a specified ϕ we can estimate μ_y and compare to selected sample proportion $\hat{\mu}_y^{(1)}$ to obtain a **Measure of Unadjusted Selection Bias for a Proportion**:

$$\boldsymbol{MUBP(\phi)} = \hat{\mu}_y^{(1)} - \hat{\mu}_y^{(\phi)}$$

  where $\hat{\mu}_y$ depends on chosen ϕ

- In a nutshell:
  1. Choose a selection mechanism by specifying ϕ
  2. Estimate overall proportion $\hat{\mu}_y^{(\phi)}$ based on pattern-mixture model
  3. Estimate selection bias (MUBP) as difference between this and the selected sample proportion

$MUBP(\phi)$: Theory

Formula is messy, but gives insight into how the $MUBP(\phi)$ index works:

$$\boldsymbol{MUBP(\phi)} = \hat{\mu}_y^{(1)} - \left[\hat{\pi}\Phi\left(\hat{\mu}_u^{(1)}\right) + (1-\hat{\pi})\Phi\left(\hat{\mu}_u^{(0)}/\sqrt{\hat{\sigma}_{uu}^{(0)}}\right)\right]$$

where

$$\hat{\mu}_u^{(0)} = \hat{\mu}_u^{(1)} + \left(\frac{\phi + (1-\phi)\hat{\rho}_{ux}^{(1)}}{\phi\hat{\rho}_{ux}^{(1)} + (1-\phi)}\right)\left(\frac{\hat{\mu}_x^{(0)} - \hat{\mu}_x^{(1)}}{\sqrt{\hat{\sigma}_{xx}^{(1)}}}\right)$$

$$\hat{\sigma}_{uu}^{(0)} = 1 + \left(\frac{\phi + (1-\phi)\hat{\rho}_{ux}^{(1)}}{\phi\hat{\rho}_{ux}^{(1)} + (1-\phi)}\right)^2\left(\frac{\hat{\sigma}_{xx}^{(0)} - \hat{\sigma}_{xx}^{(1)}}{\hat{\sigma}_{xx}^{(1)}}\right)$$

$\hat{\pi}$ = estimated selection fraction

Biserial correlation *in selected sample* ($\hat{\rho}_{ux}^{(1)}$) a very important component

Estimation

"Modified" Maximum Likelihood (MML) estimation:

- $\hat{\pi}$ = selection fraction
- $\left\{\hat{\mu}_x^{(1)}, \hat{\sigma}_{xx}^{(1)}, \hat{\mu}_x^{(0)}, \hat{\sigma}_{xx}^{(0)}\right\}$ = standard ML estimates (e.g., $\hat{\mu}_x^{(1)} = \bar{x}_{inc}$)
- $\hat{\rho}_{ux}^{(1)}$ = biserial correlation estimated via two-step method (Olsson et al. 1982)
- $\hat{\mu}_u^{(1)} = \Phi^{-1}(\hat{\mu}_y^{(1)}) = \Phi^{-1}(\bar{y}_{inc})$ = from two-step method
- Suggested sensitivity analysis: $\phi = \{0, 0.5, 1\}$

Estimation

"Modified" Maximum Likelihood (MML) estimation:

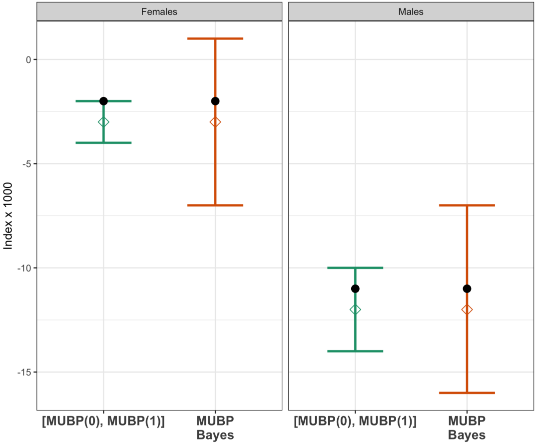
- $\hat{\pi}$ = selection fraction
- $\left\{\hat{\mu}_x^{(1)}, \hat{\sigma}_{xx}^{(1)}, \hat{\mu}_x^{(0)}, \hat{\sigma}_{xx}^{(0)}\right\}$ = standard ML estimates (e.g., $\hat{\mu}_x^{(1)} = \bar{x}_{inc}$)
- $\hat{\rho}_{ux}^{(1)}$ = biserial correlation estimated via two-step method (Olsson et al. 1982)
- $\hat{\mu}_u^{(1)} = \Phi^{-1}(\hat{\mu}_y^{(1)}) = \Phi^{-1}(\bar{y}_{inc})$ = from two-step method
- Suggested sensitivity analysis: $\phi = \{0, 0.5, 1\}$

Bayesian approach:

- Non-informative priors for identified parameters
- Incorporates uncertainty in the probit regression model for $Y|Z$ that creates X
- No info in data about ϕ, so take $\phi \sim \text{Uniform}(0, 1)$ (other priors are possible)

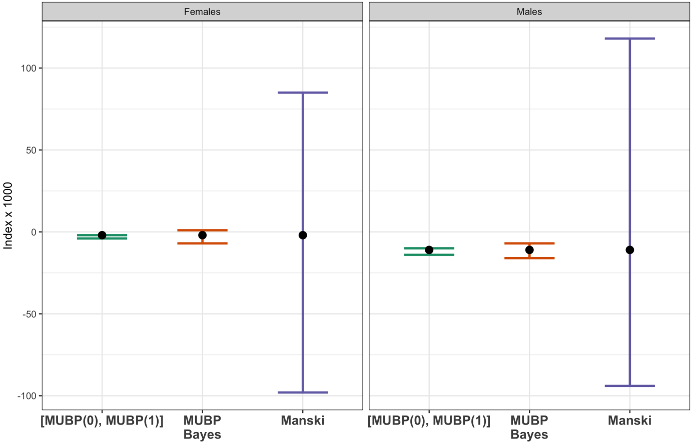
Outline

Proportion Never Married



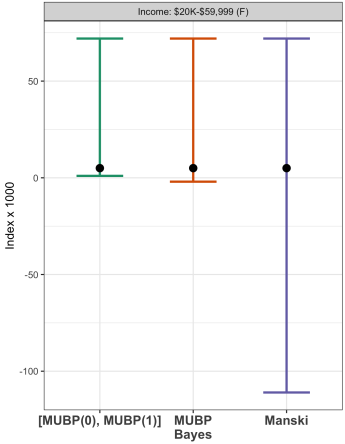
- True bias shown as black dot; $MUBP(0.5)$ shown as colored diamond
- Bayes 95% credible intervals longer than MML – but still short!

Proportion Never Married - with Manski Bounds



- Good predictors of Y: $\hat{\rho}_{ux}^{(1)} = 0.73$ (females), 0.82 (males)
- Much tighter bounds than Manski bounds (all 0s or all 1s)

Low Income - with Manski Bounds



- Weak predictors of Y: $\hat{\rho}_{ux}^{(1)} = 0.17$ (females)
- Very wide bounds $\rightarrow$ MUBP(1) = Manski bound (all 0s)

Outline

Reminder: “Failure” of Political Polling

- Recent high-profile “failure” of pre-election polls in the U.S.
- Polls are probability samples – but with low response rates
- Weighting adjustments assume selection is at random, conditional on the variables used to compute the weights

Reminder: “Failure” of Political Polling

- Recent high-profile “failure” of pre-election polls in the U.S.
- Polls are probability samples – but with low response rates
- Weighting adjustments assume selection is at random, conditional on the variables used to compute the weights
- But. . . might Trump supporters be less likely to answer a pre-election poll, even conditional on demographic characteristics?

Reminder: "Failure" of Political Polling

- Recent high-profile "failure" of pre-election polls in the U.S.
- Polls are probability samples – but with low response rates
- Weighting adjustments assume selection is at random, conditional on the variables used to compute the weights
- But. . . might Trump supporters be less likely to answer a pre-election poll, even conditional on demographic characteristics?
- $MUBP(\phi)$ could be used to adjust poll estimates to account for possible non-ignorable selection bias!

Data Source(s)

Proportion: Percentage voting for Trump

Sample: Publicly available data from seven different pre-election polls conducted in seven different states by ABC/Washington Post in 2020

- Random-digit dialing survey with low response rates (4.5-6.5%)
- Weighting adjustments to Census margins for age, gender (binary), education, race/ethnicity, party id

Truth: Official election outcomes in each state

Population: Likely voters

Data Source(s)

Proportion: Percentage voting for Trump

Sample: Publicly available data from seven different pre-election polls conducted in seven different states by ABC/Washington Post in 2020

- Random-digit dialing survey with low response rates (4.5-6.5%)
- Weighting adjustments to Census margins for age, gender (binary), education, race/ethnicity, party id

Truth: Official election outcomes in each state

Population: Likely voters

Tricky challenge: Finding population-level summary of “likely voter” characteristics (for non-selected cases)

Data Source for Non-Selected Sample (Likely Voters)

- Data sources considered:
  - 2020 Current Population Survey (CPS) voter supplement
  - 2020 American National Election Studies (ANES) pre-election survey
  - AP/NORC VoteCast 2020 data

Data Source for Non-Selected Sample (Likely Voters)

- Data sources considered:
  - 2020 Current Population Survey (CPS) voter supplement
  - 2020 American National Election Studies (ANES) pre-election survey
  - AP/NORC VoteCast 2020 data
- Ultimately, none were optimal
  - CPS, ANES – didn't have highly-relevant ideology/party preference
  - AP/NORC VoteCast – not actually available pre-election

Data Source for Non-Selected Sample (Likely Voters)

- Data sources considered:
  - 2020 Current Population Survey (CPS) voter supplement
  - 2020 American National Election Studies (ANES) pre-election survey
  - AP/NORC VoteCast 2020 data
- Ultimately, none were optimal
  - CPS, ANES – didn't have highly-relevant ideology/party preference
  - AP/NORC VoteCast – not actually available pre-election
- Decided to use **AP/NORC VoteCast**
  - Effectively doing a "post-mortem" on the poll results
  - Might non-ignorable selection/non-response (partially) explain the poor performance of the polls?

Data for MUBP Framework

- Y = indicator for voting for Trump
- Z = auxiliary data (Z) available in ABC/WP poll data: (binary) gender, age, education, race/ethnicity, political ideation, party identification
  - Strong predictors of Y – biserial correlations 0.80 to 0.86 among selected sample (poll respondents)

Data for MUBP Framework

- Y = indicator for voting for Trump
- Z = auxiliary data (Z) available in ABC/WP poll data: (binary) gender, age, education, race/ethnicity, political ideation, party identification
  - Strong predictors of Y – biserial correlations 0.80 to 0.86 among selected sample (poll respondents)
- Population-level estimates of mean Z from AP/NORC VoteCast data
  - Not without error – but we treat as if they were the "truth" (another paper!)

Data for MUBP Framework

- Y = indicator for voting for Trump
- Z = auxiliary data (Z) available in ABC/WP poll data: (binary) gender, age, education, race/ethnicity, political ideation, party identification
  - Strong predictors of Y – biserial correlations 0.80 to 0.86 among selected sample (poll respondents)
- Population-level estimates of mean Z from AP/NORC VoteCast data
  - Not without error – but we treat as if they were the "truth" (another paper!)
- Use unweighted ABC sample as the selected sample[4] and estimate $MUBP(\phi)$ with $\phi \sim$ Uniform(0,1)
- Produce MUBP-Adjusted estimates using $MUBP(\phi)$ to shift sample proportion

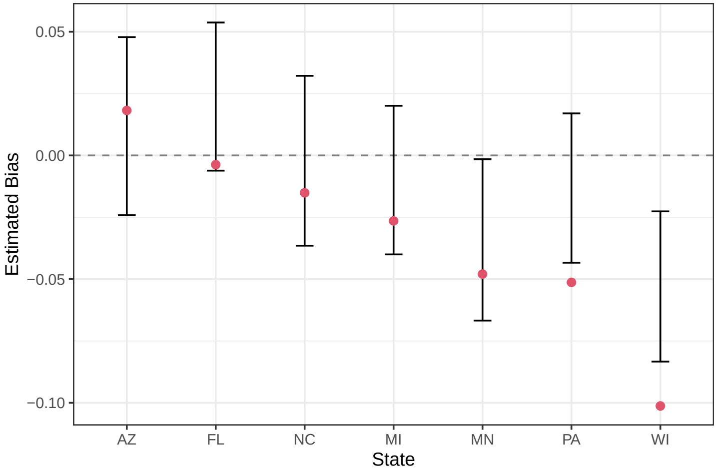
[4]ignoring sampling weights – treating as a non-probability sample

Data for MUBP Framework

- $Y =$ indicator for voting for Trump
- $Z =$ auxiliary data (Z) available in ABC/WP poll data: (binary) gender, age, education, race/ethnicity, political ideation, party identification
  - Strong predictors of Y – biserial correlations 0.80 to 0.86 among selected sample (poll respondents)
- Population-level estimates of mean Z from AP/NORC VoteCast data
  - Not without error – but we treat as if they were the "truth" (another paper!)
- Use unweighted ABC sample as the selected sample[4] and estimate $MUBP(\phi)$ with $\phi \sim$ Uniform(0,1)
- Produce MUBP-Adjusted estimates using $MUBP(\phi)$ to shift sample proportion
- Polls' selection fractions are teeny ($n \approx 1,000$ but $N =$ millions!) – Manski bounds are useless

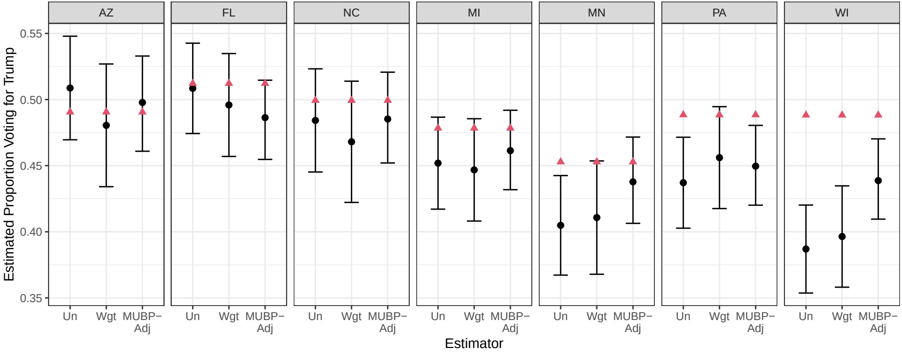
[4] ignoring sampling weights – treating as a non-probability sample

True Bias and MUBP Bayes intervals



Red circle = true bias

Comparison with ABC Poll Estimates



Red triangle = true proportion

Black circle = estimated proportions from ABC polls and $MUBP(\phi)$-adjusted

Results Summary

- MUBP correctly detected evidence of negative selection bias in MN and WI
- MUBP suggested negative bias in some other states (NC, MI), though 0 also in interval
- Huge polling miss in WI, and MUBP moved estimate in correct direction
- MUBP-adjustment often closer to truth than weighted estimate
- Credible intervals for MUBP-adjusted narrower than weighted
- MUBP did not suggest bias in PA, but there was negative bias

Key message: Need quality information on population margins for Z!

Outline

Summary and Related/Future Work

- MUBP(ϕ) provides a sensitivity analysis to assess the potential for non-ignorable selection bias
  - MUBP(0) – ignorable – could be "adjusted away"
  - MUBP(1) – non-ignorable – selection depends only on Y (through U)
  - MUBP(0.5) – could be used as a compromise "estimate" of the bias

Summary and Related/Future Work

- MUBP(ϕ) provides a sensitivity analysis to assess the potential for non-ignorable selection bias
  - MUBP(0) – ignorable – could be "adjusted away"
  - MUBP(1) – non-ignorable – selection depends only on Y (through U)
  - MUBP(0.5) – could be used as a compromise "estimate" of the bias
- Tailored to binary outcomes, an improvement over the normal-based (S)MUB of Little et al.
- Only requires summary statistics for covariates Z for non-selected

Summary and Related/Future Work

- MUBP(ϕ) provides a sensitivity analysis to assess the potential for non-ignorable selection bias
  - MUBP(0) – ignorable – could be "adjusted away"
  - MUBP(1) – non-ignorable – selection depends only on Y (through U)
  - MUBP(0.5) – could be used as a compromise "estimate" of the bias
- Tailored to binary outcomes, an improvement over the normal-based (S)MUB of Little et al.
- Only requires summary statistics for covariates Z for non-selected
- With weak predictive information, will return the natural Manski upper/lower bound

Summary and Related/Future Work

- MUBP(ϕ) provides a sensitivity analysis to assess the potential for non-ignorable selection bias
  - MUBP(0) – ignorable – could be "adjusted away"
  - MUBP(1) – non-ignorable – selection depends only on Y (through U)
  - MUBP(0.5) – could be used as a compromise "estimate" of the bias
- Tailored to binary outcomes, an improvement over the normal-based (S)MUB of Little et al.
- Only requires summary statistics for covariates Z for non-selected
- With weak predictive information, will return the natural Manski upper/lower bound
- Related work: Extension to estimation of selection bias for linear regression coefficients and probit regression coefficients (West et al., 2021)
- Future work: Extension to generalizability of randomized trials in the presence of unmeasured effect modifiers

Questions?

Thank you!
andridge.1@osu.edu

References

- Andridge, R.R. and Little, R.J.A. (2011). Proxy-pattern mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27, 153-180.
- Andridge, R.R., West, B.T., Little, R.J.A., Boonstra, P.S., and Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *JRSS-C (Applied Statistics)*, 68, 1465-1483.
- Andridge, R.R. and Little, R.J.A. (2020). Proxy pattern-mixture analysis for a binary survey variable subject to nonresponse. *Journal of Official Statistics*, 36;,703-728.
- Clinton, J., et al. (2020). "Task Force on 2020 Pre-Election Polling: An Evaluation of the 2020 General Election Polls." AAPOR. Available at `https://www.aapor.org/Education-Resources/Reports/2020-Pre-Election-Polling-An-Evaluation-of-the-202.aspx`
- Little, R.J.A., West, B.T., Boonstra, P.S., and Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, 8(5), 932-964.
- Nishimura, R., Wagner, J., and Elliott, M. (2016). Alternative indicators for the risk of non-response bias: A simulation study. *International Statistical Review*, 84(1), 43-62.
- Olsson, U., Drasgow, F. and Dorans, N. (1982). The polyserial correlation coefficient. *Psychometrika*, 47, 337-347.
- Rubin, D.B. (1976). Inference and missing data (with discussion). *Biometrika*, 63, 581-592.
- Särndal, C.-E., and S. Lundström (2010). Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, 36, 131–144.
- Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35(1), 101-113.
- West, B.T., and Andridge, R.R. (2022). An evaluation of 2020 pre-election polling estimates using new measures of non-ignorable selection bias. *Submitted*.
- West, B.T., Little, R.J.A., Andridge, R.R., Boonstra, P., Ware, E.B., Pandit, A., Alvarado-Leiton, F. (2021). Assessing selection bias in regression coefficients estimated from nonprobability samples with applications to genetics and demographic surveys. *Annals of Applied Statistics*, 15, 1556-1581.

Does Normal-based SMUB Work Well-Enough?

- SMUB(ϕ) much simpler than MUBP(ϕ)
  - Directly apply the proxy pattern-mixture model to Y and X instead of latent U and X
  - Relies on pearson correlation instead of biserial correlation
  - Unlike MUBP(ϕ), only need **means** from unselected cases (not variance)

$$SMUB(\phi) = \left(\frac{\phi + (1-\phi) r_{ux}^{(1)}}{\phi r_{yx}^{(1)} + (1-\phi)}\right) \left(\frac{\bar{x}^{(1)} - \bar{x}}{\sqrt{s_{xx}^{(1)}}}\right)$$

- Is there an advantage to proportion-based MUBP(ϕ) over means-based MUB(ϕ)?
  - To compare to MUBP(ϕ), we consider the unstandardized version, MUB(ϕ):

$$MUB(\phi) = \left(\frac{\phi + (1-\phi) r_{ux}^{(1)}}{\phi r_{yx}^{(1)} + (1-\phi)}\right) \frac{\sqrt{s_{yy}^{(1)}}}{\sqrt{s_{xx}^{(1)}}} \left(\bar{x}^{(1)} - \bar{x}\right)$$

Simulation Set-Up

Population Design

- Auxiliary variable: $z_i \sim N(0,1)$ for population size $N = 10,000$
- Latent variable: $u_i|z_i \sim N\left(\alpha_0 + \frac{\rho_{ux}}{\sqrt{(1-\rho_{ux}^2)}} z_i, 1\right)$
  - ρ_{ux} = biserial correlation for whole population (not selected sample)
  - α_0 chosen to obtain $E(Y) = \mu_Y$
- Binary outcome: $y_i = 1$ if $u_i > 0$ (and 0 otherwise)
- Varied $\rho_{ux} = \{0.2, 0.5, 0.8\}$, $\mu_Y = \{0.1, 0.3, 0.5\}$

Simulation Set-Up

Population Design

- Auxiliary variable: $z_i \sim N(0,1)$ for population size $N = 10,000$
- Latent variable: $u_i|z_i \sim N\left(\alpha_0 + \frac{\rho_{ux}}{\sqrt{(1-\rho_{ux}^2)}} z_i, 1\right)$
  - ρ_{ux} = biserial correlation for whole population (not selected sample)
  - α_0 chosen to obtain $E(Y) = \mu_Y$
- Binary outcome: $y_i = 1$ if $u_i > 0$ (and 0 otherwise)
- Varied $\rho_{ux} = \{0.2, 0.5, 0.8\}$, $\mu_Y = \{0.1, 0.3, 0.5\}$

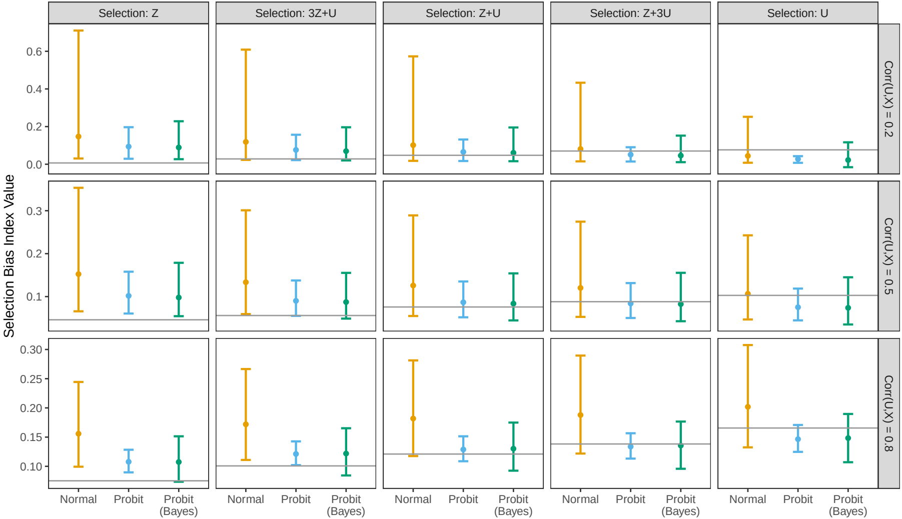
Selection Mechanisms

- Selection indicator S_i from logistic model:

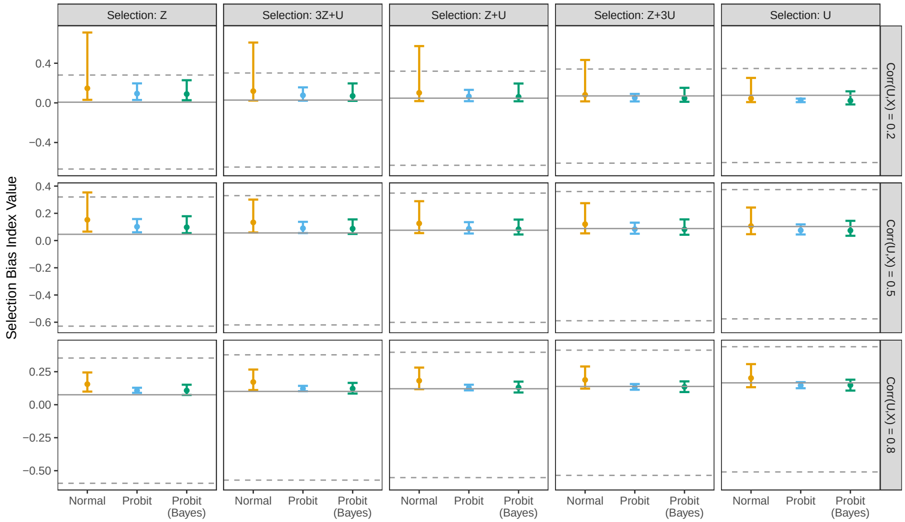
$$logit\{\Pr(s_i = 1|z_i, u_i)\} = \beta_0 + \beta_Z z_i + \beta_U u_i$$

- $\beta_U = 0$: Ignorable selection; $\beta_U > 0$: Non-ignorable
- β_0 chosen to give 5% selection fraction

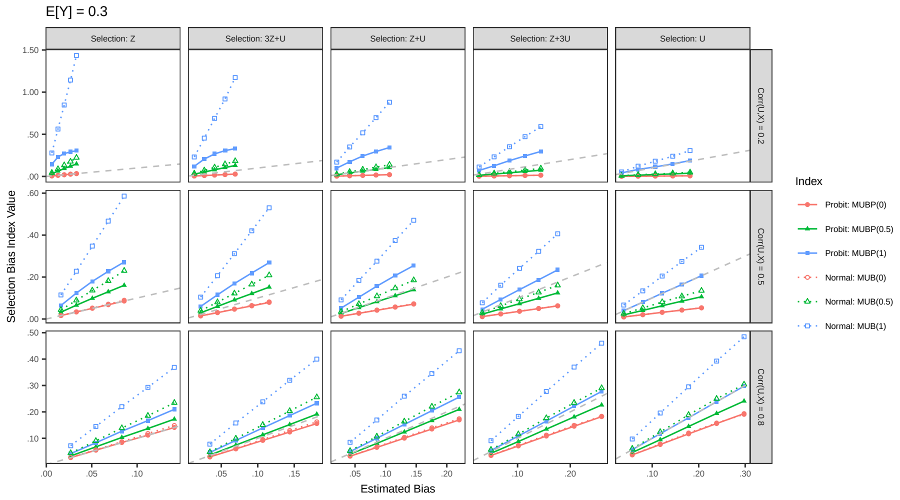
Simulation: One Replicate $(\mu_Y = 0.3)$



Simulation: One Replicate - w/Manski Bounds



Simulation: MUBP and MUB vs. True Estimated Bias



Simulation: Correlation of MUBP and MUB with Truth