When “Representative” Surveys Fail: Can a Non-ignorable Missingness Mechanism Explain Bias in Estimates of COVID-19 Vaccine Uptake?

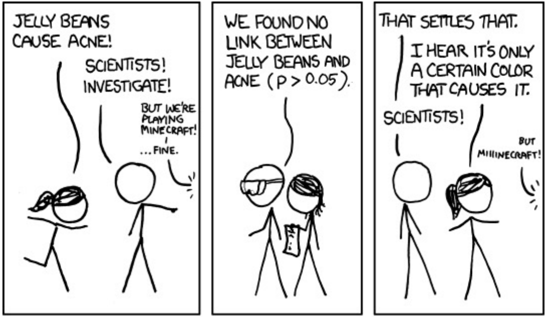
Rebecca Andridge

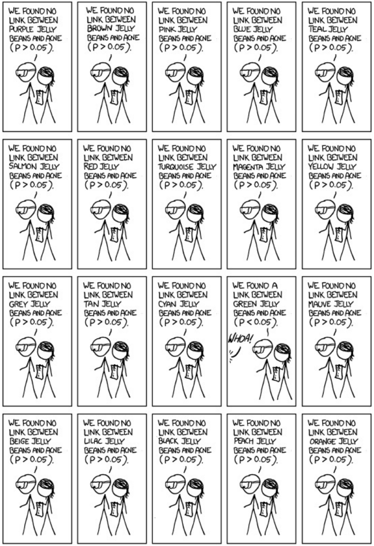
The Ohio State University College of Public Health

March 13, 2024

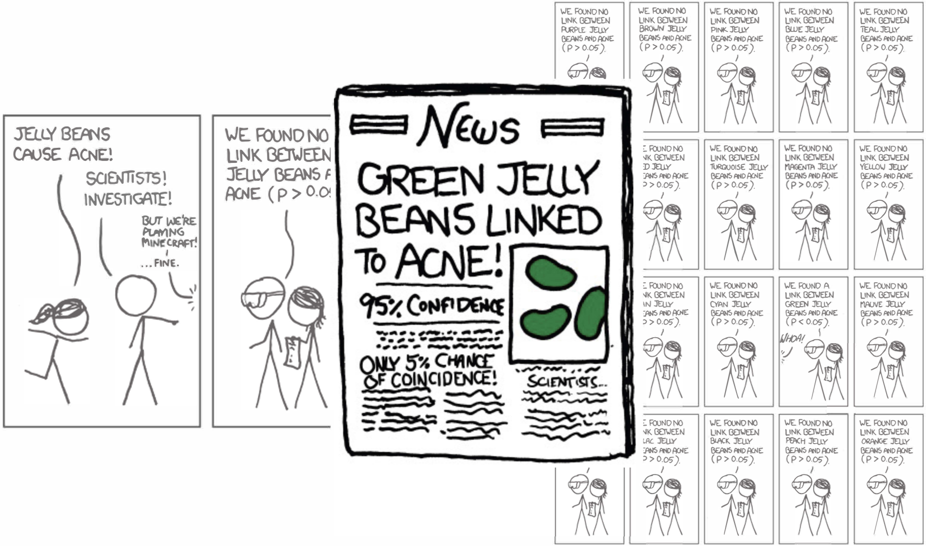
Outline

"Big Data": Friend or Foe?

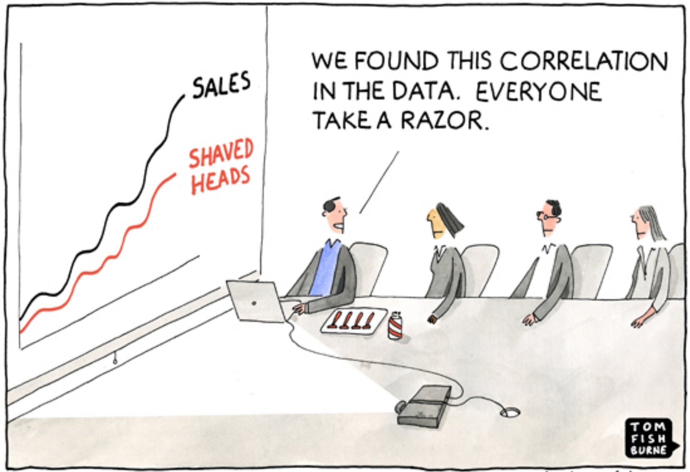




"Big Data": Friend or Foe?



“Big Data”: Friend or Foe?



"Big Data": Friend or Foe?

"Big Data": Friend or Foe?

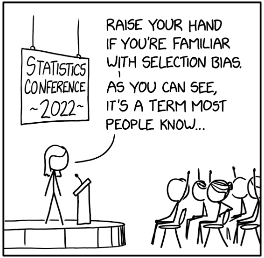
Problems most people immediately think of:

- Big sample size $\rightarrow$ small p-values
- Multiple testing
- "Spurious correlations"

“Big Data”: Friend or Foe?

Problems most people immediately think of:

- Big sample size → small p-values
- Multiple testing
- “Spurious correlations”

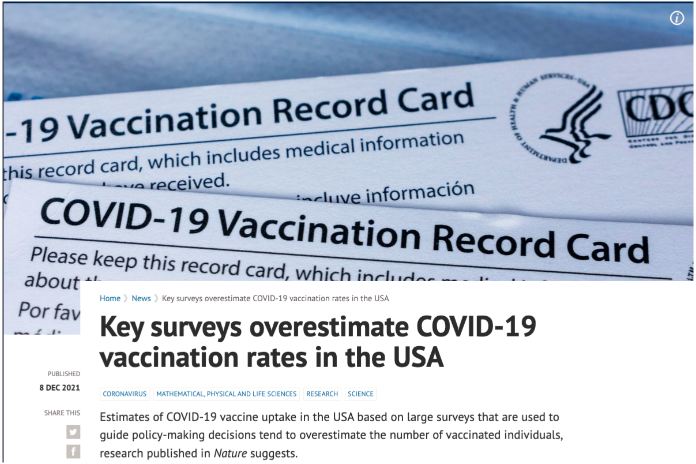


Another major issue: **Selection bias**

Also a problem for “Big Surveys” with **low response rates**

- “Big Data” = Non-probability samples → Selection bias
- “Big Surveys” = Probability samples → Nonresponse bias

“Big (COVID) Surveys” = “Big Miss”...



Home 〉 News 〉 Key surveys overestimate COVID-19 vaccination rates in the USA

Key surveys overestimate COVID-19 vaccination rates in the USA

PUBLISHED
8 DEC 2021

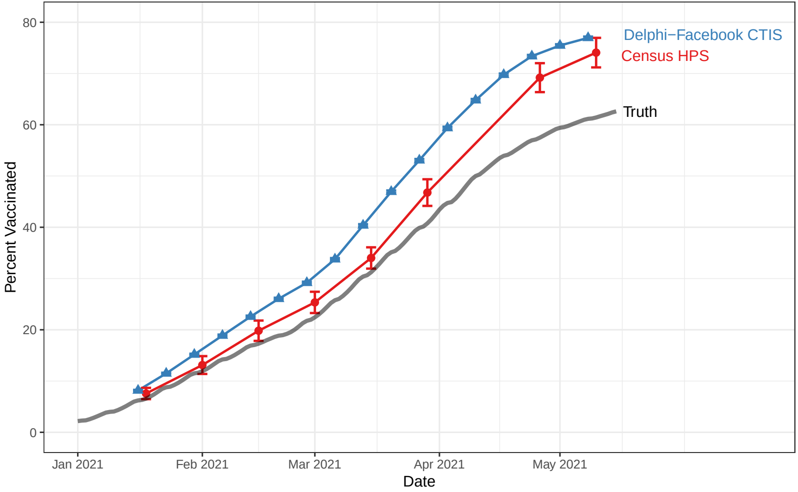
CORONAVIRUS | MATHEMATICAL, PHYSICAL AND LIFE SCIENCES | RESEARCH | SCIENCE

SHARE THIS

Estimates of COVID-19 vaccine uptake in the USA based on large surveys that are used to guide policy-making decisions tend to overestimate the number of vaccinated individuals, research published in *Nature* suggests.

Published study: Bradley et al. 2021, Nature

(Over-)Estimation of COVID-19 Vaccine Uptake



"Big Data Paradox: The bigger the data, the surer we fool ourselves" (Meng 2018, p.702)

Problem Statement

Goal: Estimate population proportion from probability samples with very low response rates (effectively non-probability samples)
$\rightarrow$ *Proportion having at least one dose of COVID-19 vaccine*

Problem Statement

Goal: Estimate population proportion from probability samples with very low response rates (effectively non-probability samples)

$\rightarrow$ *Proportion having at least one dose of COVID-19 vaccine*

Problem: Potential for bias due to non-ignorable nonresponse

- Ignorable: probability of survey participation depends on *observed characteristics*
- Non-ignorable: probability of survey participation depends at least in part on *unobserved characteristics*

$\rightarrow$ *Participation might depend on your vaccine status*

Problem Statement

Goal: Estimate population proportion from probability samples with very low response rates (effectively non-probability samples)
$\rightarrow$ *Proportion having at least one dose of COVID-19 vaccine*

Problem: Potential for bias due to non-ignorable nonresponse

- Ignorable: probability of survey participation depends on *observed characteristics*
- Non-ignorable: probability of survey participation depends at least in part on *unobserved characteristics*

$\rightarrow$ *Participation might depend on your vaccine status*

Approach: Use the **Proxy Pattern-Mixture Model (PPMM)** to assess potential nonresponse/selection bias in proportion estimates (Andridge and Little 2020; Andridge et al. 2019)
$\rightarrow$ *Sensitivity analysis allowing survey participation to depend on vaccine status*

Outline

Census Household Pulse Survey (HPS)[*]

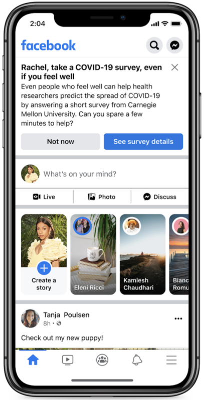
- Launched April 23, 2020; still ongoing
- Collaboration between 8+ agencies
- Online survey (Qualtrics)
- Repeated cross-sectional probability samples
- Sampling frame: Census Bureau Master Address File *where at least one email address or cell phone known*
- 1- and then 2-week waves
- n=68,000-80,000 respondents per wave [Jan-May 2021]

Q: *Have you received a COVID-19 vaccine?* {Yes, No}

[*] https://www.census.gov/data/experimental-data-products/household-pulse-survey.html

Delphi-Facebook COVID-19 Trends and Impacts Survey (CTIS)*

- Launched April 6, 2020; Ended June 25, 2022
- Both U.S. and Global samples
- Online survey (Qualtrics)
- Repeated cross-sectional probability samples
- Sampling frame: Facebook users 18+ who were active on the platform in the last month
- Daily samples (pooled into weekly waves)
- n=160,000-290,000 respondents per wave [Jan-May 2021]



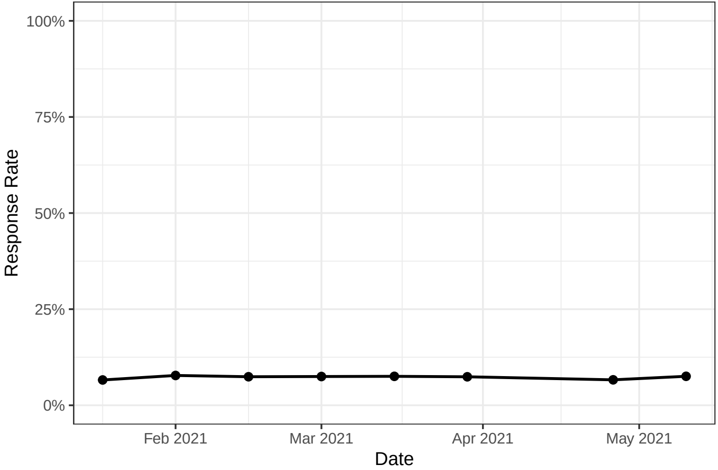
Q: *Have you had a COVID-19 vaccination?* {Yes, No, I don't know}

---

* https://delphi.cmu.edu/covid19/ctis/

Big Surveys, Small Response Rates

Census HPS Response Rates*



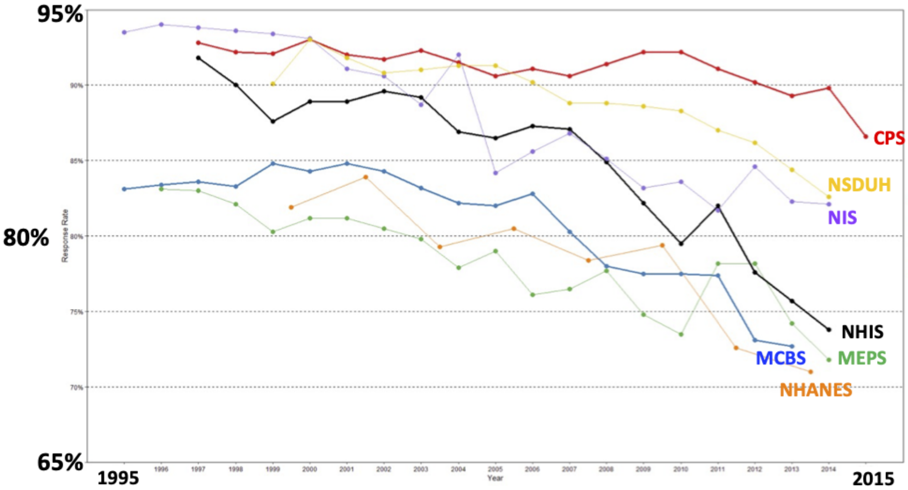
* Percent who responded out of all sampled persons

Big Surveys, Small Response Rates

Delphi-Facebook Cooperation Rates*

*Percent who responded out of all who saw survey invite (logged into FB)

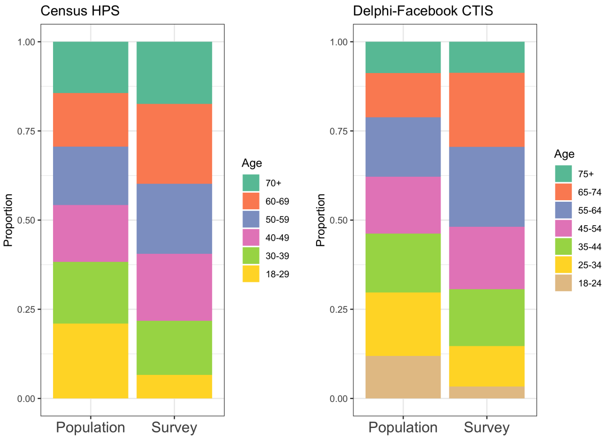
Compare to Traditional "Big Survey" Response Rates



Czajka and Beyler 2016

COVID Surveys: Respondents don't resemble Population

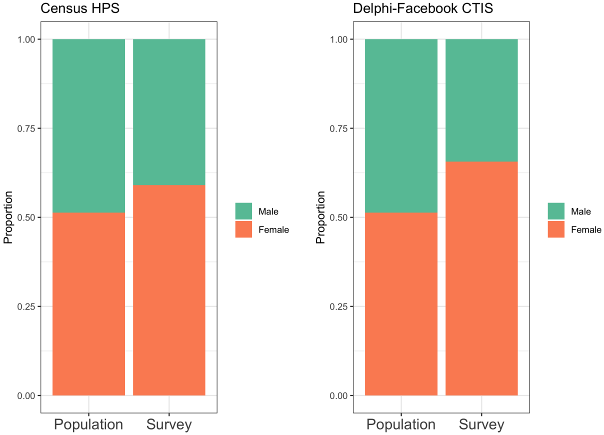
Age*



*Demographics shown for last wave analyzed of each survey

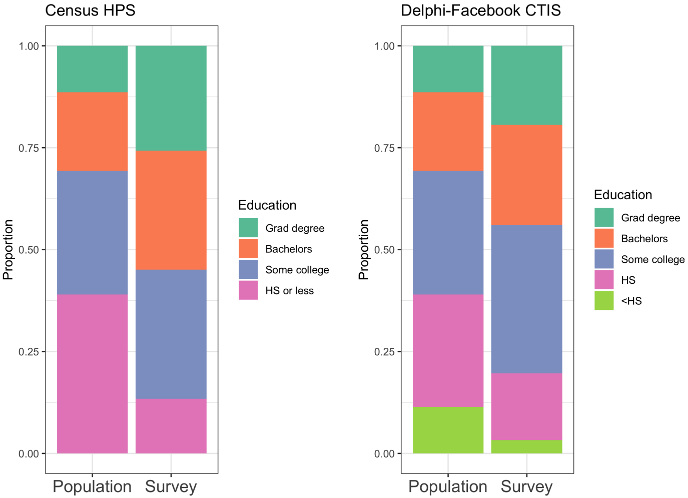
COVID Surveys: Respondents don't resemble Population

Gender*



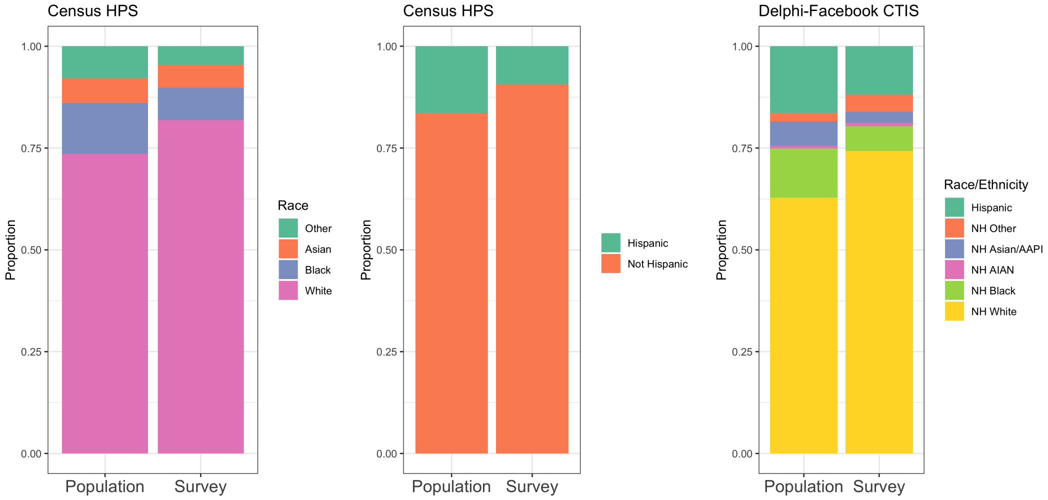
*Limitation: gender used as a binary variable

Education


Census HPS
1.00
0.75
0.50
0.25
0.00
Proportion
Population
Survey
Education
Grad degree
Bachelors
Some college
HS or less
Delphi-Facebook CTIS
1.00
0.75
0.50
0.25
0.00
Proportion
Population
Survey
Education
Grad degree
Bachelors
Some college
HS
<HS

COVID Surveys: Respondents don't resemble Population

Race and Ethnicity



Solution: Nonresponse Weighting Adjustments

- Adjust sample weights to make respondents “look like” population
  - Upweight male, younger, lower education, non-white

Solution: Nonresponse Weighting Adjustments

- Adjust sample weights to make respondents "look like" population
  - Upweight male, younger, lower education, non-white
- Both surveys did this, but with limited demographic information:
  - Census HPS: age, gender[1], race/ethnicity, education, state
  - Delphi-Facebook: age, gender[2]
  - Population data sources: American Community Survey, Current Population Survey

[1] Limitation: gender collected as a binary variable

[2] Limitation: collected gender with >2 categories but have to weight to a source that has gender as a binary variable

Solution: Nonresponse Weighting Adjustments

- Adjust sample weights to make respondents "look like" population
  - Upweight male, younger, lower education, non-white
- Both surveys did this, but with limited demographic information:
  - Census HPS: age, gender[1], race/ethnicity, education, state
  - Delphi-Facebook: age, gender[2]
  - Population data sources: American Community Survey, Current Population Survey
- Weighting makes respondents look like the population **with respect to the weighting variables**
- Assumes that two people of the same (age, gender, race/ethnicity, education) or (age, gender) are **interchangeable**, one who participated and one who did not

[1] Limitation: gender collected as a binary variable

[2] Limitation: collected gender with >2 categories but have to weight to a source that has gender as a binary variable

Solution: Nonresponse Weighting Adjustments

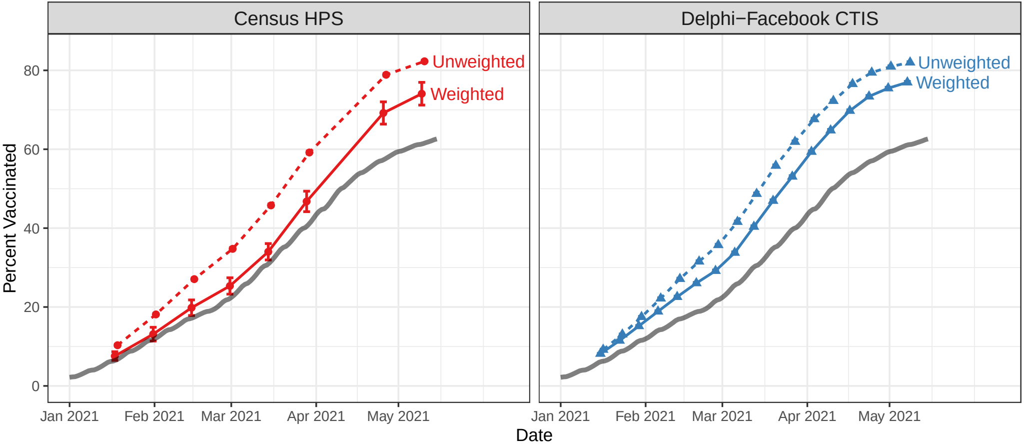
- Adjust sample weights to make respondents "look like" population
  - Upweight male, younger, lower education, non-white
- Both surveys did this, but with limited demographic information:
  - Census HPS: age, gender[1], race/ethnicity, education, state
  - Delphi-Facebook: age, gender[2]
  - Population data sources: American Community Survey, Current Population Survey
- Weighting makes respondents look like the population **with respect to the weighting variables**
- Assumes that two people of the same (age, gender, race/ethnicity, education) or (age, gender) are **interchangeable**, one who participated and one who did not

Do we believe this assumption? In the context of COVID?

[1] Limitation: gender collected as a binary variable

[2] Limitation: collected gender with >2 categories but have to weight to a source that has gender as a binary variable

Weighting Helped Somewhat. . . But Not Enough!



Weighted estimates closer to truth, but still biased

Let's see if the PPMM can do better!

Outline

PPMM for Binary Outcomes

- Y = binary variable of interest, only available for respondents
  - Individual has received 1+ dose of vaccine
- Z = auxiliary variables, available for respondents and in aggregate for population $(\bar{Z})$
  - Age, gender, race/ethnicity, education (HPS)
- S = indicator for unit selected **and** responded

PPMM for Binary Outcomes

- Y = binary variable of interest, only available for respondents
  - Individual has received 1+ dose of vaccine
- Z = auxiliary variables, available for respondents and in aggregate for population $(\bar{Z})$
  - Age, gender, race/ethnicity, education (HPS)
- S = indicator for unit selected **and** responded
- U = underlying normally distributed unobserved latent variable
  - $Y = 1$ when $U > 0$

PPMM for Binary Outcomes

- Y = binary variable of interest, only available for respondents
  - Individual has received 1+ dose of vaccine
- Z = auxiliary variables, available for respondents and in aggregate for population $(\bar{Z})$
  - Age, gender, race/ethnicity, education (HPS)
- S = indicator for unit selected **and** responded
- U = underlying normally distributed unobserved latent variable
  - $Y = 1$ when $U > 0$
- X = "proxy" for Y, based on Z
  - Constructed from probit regression: $P(Y = 1|Z, S = 1) = \Phi(\alpha_0 + \alpha Z)$
  - Available at individual-level for selected/respondents: $X = \hat{\alpha}_0 + \hat{\alpha} Z$
  - Available in aggregate for rest of population: $\bar{X} = \hat{\alpha}_0 + \hat{\alpha}\bar{Z}$
  - *Proxy strength* = Biserial Corr$(Y, X|S = 1)$ = Corr$(U, X|S = 1)$

PPMM for Binary Outcomes

Basic idea:

- We can measure the degree of bias in the proxy X (known for population!)

PPMM for Binary Outcomes

Basic idea:

- We can measure the degree of bias in the proxy X (known for population!)
- If Y is correlated with X, then this tells you something about the *potential* bias in Y

PPMM for Binary Outcomes

Basic idea:

- We can measure the degree of bias in the proxy X (known for population!)
- If Y is correlated with X, then this tells you something about the *potential* bias in Y

General approach:

- Use pattern-mixture models to specify $f(Y, X, S) = f(Y, X|S)f(S)$
- Only $f(Y, X|S = 1)$ identifiable (and $f(X|S = 0)$)
- Make explicit, untestable assumption(s) about S to identify $f(Y, X|S = 0)$
- Creates sensitivity analysis to assess range of bias under different assumptions about S

PPMM for Binary Outcomes

Basic idea:

- We can measure the degree of bias in the proxy X (known for population!)
- If Y is correlated with X, then this tells you something about the *potential* bias in Y

General approach:

- Use pattern-mixture models to specify $f(Y, X, S) = f(Y, X|S)f(S)$
- Only $f(Y, X|S = 1)$ identifiable (and $f(X|S = 0)$)
- Make explicit, untestable assumption(s) about S to identify $f(Y, X|S = 0)$
- Creates sensitivity analysis to assess range of bias under different assumptions about S

Trick for convenience:

- Use latent U instead of binary Y

PPMM: Theory

- Assume a proxy pattern-mixture model* for U and X given S:

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

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

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

* Andridge and Little 2011, 2020

PPMM: Theory

- Assume a proxy pattern-mixture model* for U and X given S:

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

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

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

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

*Andridge and Little 2011, 2020

PPMM: Theory

- Assume a proxy pattern-mixture model* for U and X given S:

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

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

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

- Problem: unidentified parameters $= \left\{ \mu_u^{(0)}, \sigma_{uu}^{(0)}, \rho_{ux}^{(0)} \right\}$

* Andridge and Little 2011, 2020

PPMM: Theory

- Non-identifiable parameters $\left\{\mu_u^{(0)},\ \sigma_{uu}^{(0)},\ \rho_{ux}^{(0)}\right\}$ are just identified by assumption about selection/response mechanism:

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

  - $X^* = \frac{X}{\sqrt{\sigma_{xx}^{(1)}}}$ = rescaled proxy X
  - V = additional variables independent of X and U that may be associated with S
  - $\phi \in [0,1]$ is a sensitivity parameter (no info in data about it)

PPMM: Theory

- Non-identifiable parameters $\left\{\mu_u^{(0)},\, \sigma_{uu}^{(0)},\, \rho_{ux}^{(0)}\right\}$ are just identified by assumption about selection/response mechanism:

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

  - $X^* = \frac{X}{\sqrt{\sigma_{xx}^{(1)}}}$ = rescaled proxy X
  - V = additional variables independent of X and U that may be associated with S
  - $\phi \in [0, 1]$ is a sensitivity parameter (no info in data about it)

- Selected value of ϕ determines selection mechanism:
  - $\phi = 0 \rightarrow \Pr(S = 1|U, X, V) = f(X^*, V)$ **Ignorable selection**

PPMM: Theory

- Non-identifiable parameters $\left\{\mu_u^{(0)},\ \sigma_{uu}^{(0)},\ \rho_{ux}^{(0)}\right\}$ are just identified by assumption about selection/response mechanism:

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

  - $X^* = \frac{X}{\sqrt{\sigma_{xx}^{(1)}}}$ = rescaled proxy X
  - V = additional variables independent of X and U that may be associated with S
  - $\phi \in [0,1]$ is a sensitivity parameter (no info in data about it)

- Selected value of ϕ determines selection mechanism:
  - $\phi = 0 \rightarrow \Pr(S=1|U,X,V) = f(X^*, V)$ **Ignorable selection**
  - $\phi = 1 \rightarrow \Pr(S=1|U,X,V) = f(U, V)$ **"Extremely" Non-ignorable selection**

PPMM: Theory

- Non-identifiable parameters $\left\{\mu_u^{(0)},\ \sigma_{uu}^{(0)},\ \rho_{ux}^{(0)}\right\}$ are just identified by assumption about selection/response mechanism:

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

  - $X^* = \frac{X}{\sqrt{\sigma_{xx}^{(1)}}}$ = rescaled proxy X
  - V = additional variables independent of X and U that may be associated with S
  - $\phi \in [0,1]$ is a sensitivity parameter (no info in data about it)

- Selected value of ϕ determines selection mechanism:
  - $\phi = 0 \rightarrow \Pr(S=1|U,X,V) = f(X^*, V)$ **Ignorable selection**
  - $\phi = 1 \rightarrow \Pr(S=1|U,X,V) = f(U, V)$ **"Extremely" Non-ignorable selection**
  - $0 < \phi < 1 \rightarrow \Pr(S=1|U,X,V) = f((1-\phi)X^* + \phi U, V)$ **Non-ignorable selection**

PPMM: Theory

For a specified ϕ we can estimate μ_y:

$$\hat{\mu}_y = \hat{\pi} \underbrace{\Phi\left(\hat{\mu}_u^{(1)}\right)}_{\text{respondents}} + (1-\hat{\pi}) \underbrace{\Phi\left(\hat{\mu}_u^{(0)} / \sqrt{\hat{\sigma}_{uu}^{(0)}}\right)}_{\text{rest of pop.}}$$

where

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

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

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

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., $\hat{\mu}_x^{(1)} = \bar{x}_{resp}$)
- $\hat{\rho}_{ux}^{(1)}$ = biserial correlation estimated via two-step method (Olsson et al. 1982)
- $\hat{\mu}_u^{(1)} = \Phi^{-1}(\hat{\mu}_y^{(1)}) = \Phi^{-1}(\bar{y}_{resp})$ = 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., $\hat{\mu}_x^{(1)} = \bar{x}_{resp}$)
- $\hat{\rho}_{ux}^{(1)}$ = biserial correlation estimated via two-step method (Olsson et al. 1982)
- $\hat{\mu}_u^{(1)} = \Phi^{-1}(\hat{\mu}_y^{(1)}) = \Phi^{-1}(\bar{y}_{resp})$ = 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 $Y|Z, S = 1$ that creates X
- No info in data about ϕ, so take $\phi \sim \mathrm{Uniform}(0, 1)$
  (other priors are possible)

Outline

Available Data: COVID Surveys

Microdata for survey respondents $(S = 1)$:

- Y = vaccination status (received at least one dose)
  - Missing data treatment follows what the surveys did for reporting:
    - ★ Census HPS: If missing, assume "no"
    - ★ Delphi-Facebook CTIS: If missing, drop (≈6-7%)
- Z = auxiliary variables
  - Census HPS: age, gender, race, ethnicity, education
  - Delphi-Facebook CTIS: age, gender, race/ethnicity, education
  - Missing data treatment:
    - ★ Census HPS: No missing data (singly imputed by Census)
    - ★ Delphi-Facebook CTIS: If missing any, drop (≈15% additional)
- Sample sizes:
  - Census HPS: $n \approx$ 68,000-80,000 per wave
  - Delphi-Facebook CTIS: $n \approx$ 160,000-290,000 per week

Available Data: Population

Aggregate data $(\bar{Z})$ for rest of population $(S = 0)$:

- Source: 2019 American Community Survey
  - Weighted estimates from ACS treated as “known”
  - Same as using ACS totals for weight adjustments
- Technically, 2019 ACS gives $\bar{Z}$ for the full population, not just nonresponding – but selection fraction is tiny
  ($N \approx 250$ million, largest $n \approx 250$ thousand)

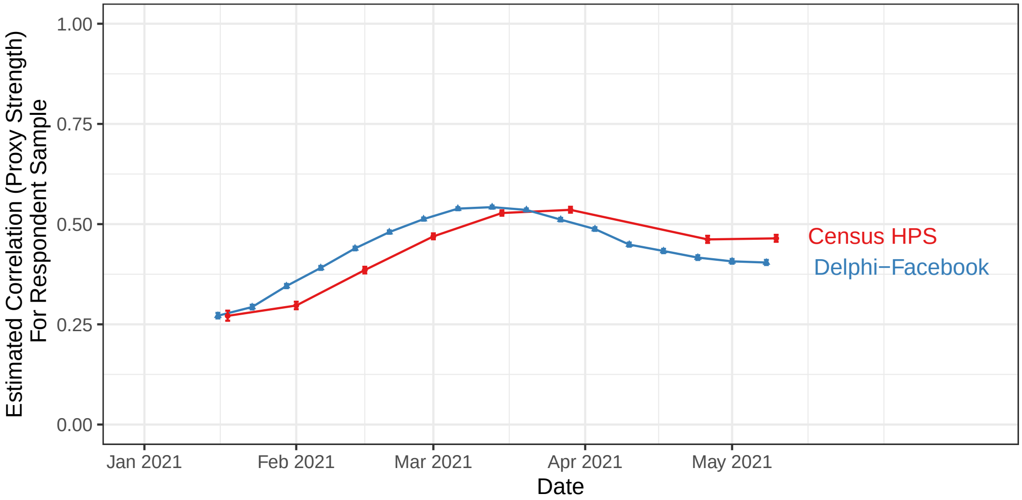
Population Truth:

- CDC benchmark numbers for vaccine uptake (retroactively corrected)

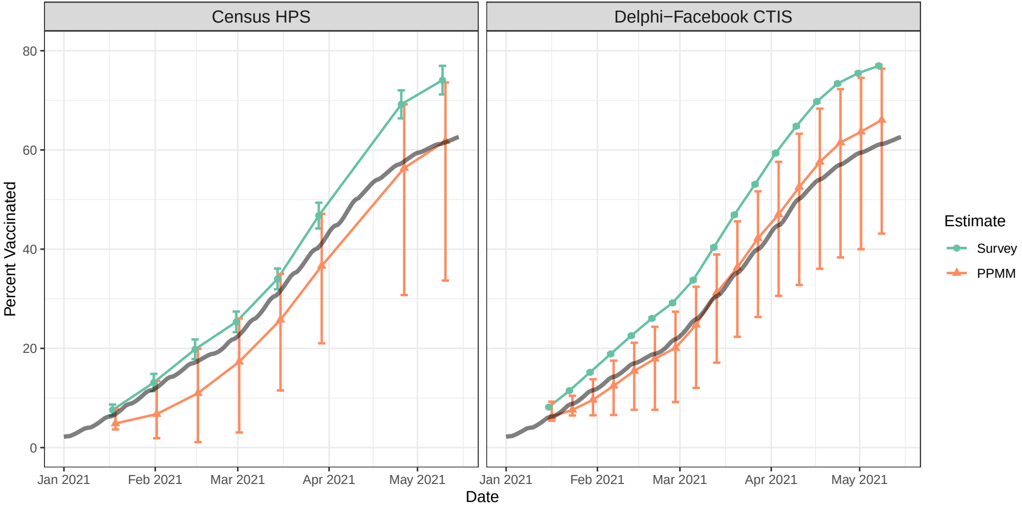
Estimation Details:

- Ignore sampling weights and treat as **non-probability samples**
- Bayesian approach with $\phi \sim$ Uniform$(0, 1)$

Percent Vaccinated: Proxy Strength



Percent Vaccinated: PPMM Estimates



Percent Vaccinated: Summary

- PPMM correctly detected direction of selection bias for both surveys in all waves/weeks
- PPMM with $\phi = 0.5$ remarkably close to truth for most CTIS weeks
- PPMM credible intervals cover the truth for both surveys in all waves/weeks
  - Direct survey estimates only covered truth twice (first two waves of Census HPS)
- PPMM credible intervals much wider than survey intervals despite large sample sizes
  - Reflects strength (weakness) of proxy model
  - Arguably a good feature: no "Big Data Paradox"!

Percent Vaccine Hesitant

Census HPS:

Once a vaccine to prevent COVID-19 is available to you, would you...

1. Definitely get a vaccine
2. Probably get a vaccine
3. Be unsure about getting a vaccine* [hesitant]
4. Probably NOT get a vaccine [hesitant]
5. Definitely NOT get a vaccine [hesitant]

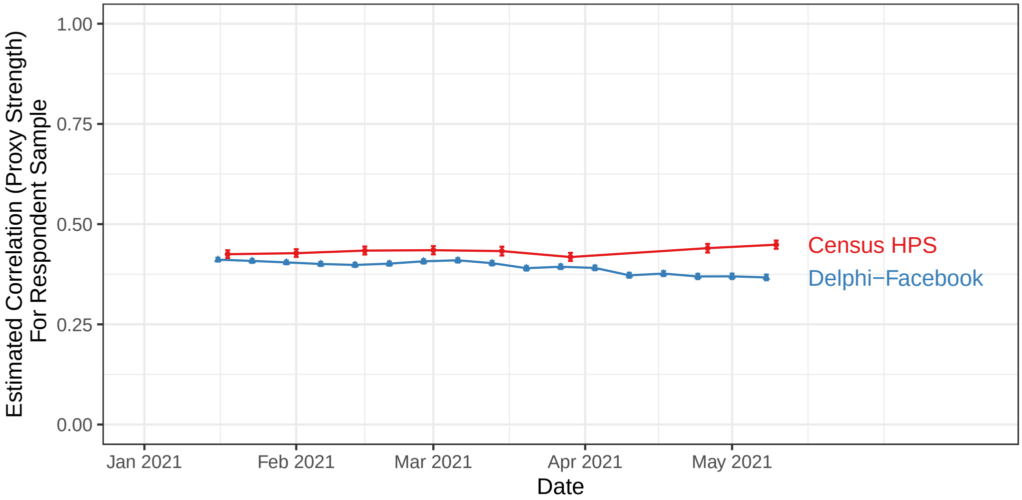
Delphi-Facebook CTIS:

If a vaccine to prevent COVID-19 were offered to you today, would you choose to get vaccinated?

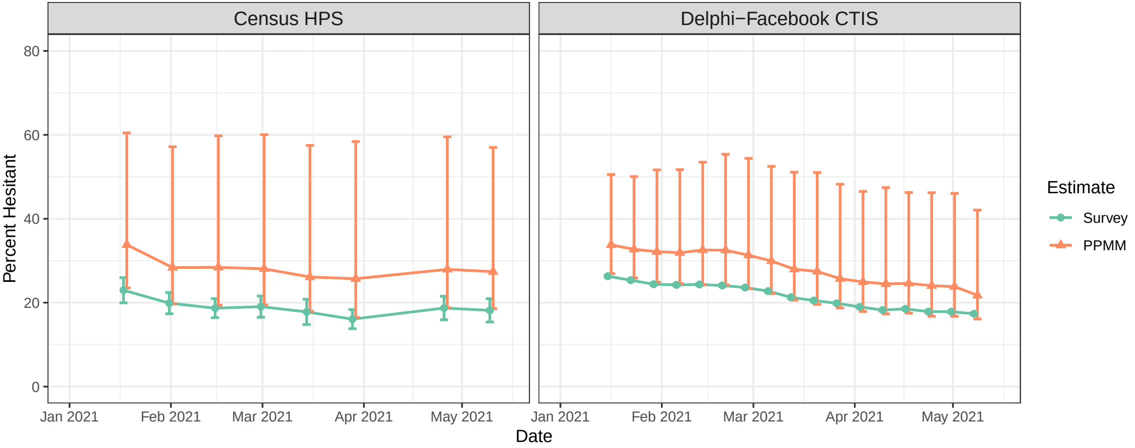
1. Yes, definitely
2. Yes, probably
3. No, probably not [hesitant]
4. No, definitely not [hesitant]

* option available starting in mid-April 2021

Percent Vaccine Hesitant: Proxy Strength



Percent Vaccine Hesitant: PPMM Estimates



$\phi = 0.5 \rightarrow$ hesitancy underestimated by $\approx$ 9% for HPS, $\approx$ 7% for CTIS

Outline

Summary and Related Work

- PPMM provides a sensitivity analysis to assess the potential for non-ignorable nonresponse/selection bias
  - $\phi = 0$ – ignorable – could be "adjusted away"
  - $\phi = 1$ – extreme non-ignorable: selection depends only on Y (via U)
  - $\phi = 0.5$ – could be used as a compromise "estimate" of the bias

Summary and Related Work

- PPMM provides a sensitivity analysis to assess the potential for non-ignorable nonresponse/selection bias
  - $\phi = 0$ – ignorable – could be "adjusted away"
  - $\phi = 1$ – extreme non-ignorable: selection depends only on Y (via U)
  - $\phi = 0.5$ – could be used as a compromise "estimate" of the bias
- Only requires summary statistics for covariates Z for non-selected
  - Same information as often used for weighting
  - Could be used during data collection to compare potential for bias across a range of Y
  - Easiest when population is well-defined and stable
    - Example when it's *not* easy: Pre-election polling!*
  - Key point: Need strong predictors of Y that are available at population-level

*West and Andridge 2023

Summary and Related Work

- PPMM provides a sensitivity analysis to assess the potential for non-ignorable nonresponse/selection bias
  - $\phi = 0$ – ignorable – could be "adjusted away"
  - $\phi = 1$ – extreme non-ignorable: selection depends only on Y (via U)
  - $\phi = 0.5$ – could be used as a compromise "estimate" of the bias
- Only requires summary statistics for covariates Z for non-selected
  - Same information as often used for weighting
  - Could be used during data collection to compare potential for bias across a range of Y
  - Easiest when population is well-defined and stable
    - ★ Example when it's *not* easy: Pre-election polling![*]
  - Key point: Need strong predictors of Y that are available at population-level
- PPMMs also available for estimating means (including deviations from normality) and linear and probit regression coefficients[†]

[*] West and Andridge 2023

[†] Andridge and Little 2011, Little et al. 2020, Andridge and Thompson 2015, Yang and Little 2021, West et al. 2021

Future Work / Extensions

Methods development:

- Using the PPMM to generate non-ignorable selection weights
- Extend PPMM for nominal responses
- Extend PPMM to multivariate outcomes
- Adapt PPMM for generalizability of randomized trials in the presence of unmeasured effect modifiers (current R03)

Additional applications:

- Apply PPMM to estimate *changes* in vaccine uptake (less biased?)
- Apply PPMM to variety of indicators to compare probability-based and opt-in online samples (AAPOR 2024 presentation)

Questions?

Thank you!
andridge.1@osu.edu

Full paper online ahead of print:

Andridge, R.R. (2024). Using proxy pattern-mixture models to explain bias in estimates of COVID-19 vaccine uptake from two large surveys. *Journal of the Royal Statistical Society – Series A*, https://doi.org/10.1093/jrsssa/qnae005.

References

- Andridge, R.R. (2024). Using proxy pattern-mixture models to explain bias in estimates of COVID-19 vaccine uptake from two large surveys. *Journal of the Royal Statistical Society – Series A*, Online ahead of print `https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnae005/7587622`.
- 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.
- Andridge, R.R. and Thompson, K.J. (2015). Assessing nonresponse bias in a business survey: Proxy pattern-mixture analysis for skewed data. *Annals of Applied Statistics*, 9(4), 2237-2265.
- Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X-L., Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600, 695-700.
- Czajka, J.L. and Beyler, A. (2016). *Background Paper: Declining response rates in federal surveys: Trends and implications.* Washington: Mathematica Policy Research. Available at: `https://aspe.hhs.gov/system/files/pdf/255531/Decliningresponserates.pdf`.
- 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.
- Meng, X.-L. (2018) Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics*, 12, 685-726.
- Olsson, U., Drasgow, F. and Dorans, N. (1982). The polyserial correlation coefficient. *Psychometrika*, 47, 337-347.
- West, B.T., and Andridge, R.R. (2023). An evaluation of 2020 pre-election polling estimates using new measures of non-ignorable selection bias. *Public Opinion Quarterly*, 87; 575-601.
- 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.
- Yang, Y., Little, R.J. (2021). Spline pattern-mixture models for missing data. *Journal of Data Science*, 19(1), 75-95.

BONUS SLIDE: How the PPMM Identification Works

Assumed model for U and X given S: $(U, X|S = j) \sim$ Bivariate Normal
Assumed response mechanism:

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

If $\phi = 0 \rightarrow$ response only depends on X (not U)

- Implies $[U|X, S = 0] = [U|X, S = 1]$
- Regression parameters for $[U|X, S = 0]$ are the same as for $S = 1$
- Standard regression estimator (e.g., under MAR assumption)

BONUS SLIDE: How the PPMM Identification Works

Assumed model for U and X given S: $(U, X|S = j) \sim$ Bivariate Normal
Assumed response mechanism:

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

If $\phi = 0 \rightarrow$ response only depends on X (not U)

- Implies $[U|X, S = 0] = [U|X, S = 1]$
- Regression parameters for $[U|X, S = 0]$ are the same as for $S = 1$
- Standard regression estimator (e.g., under MAR assumption)

If $\phi = 1 \rightarrow$ response only depends on U (not X)

- Implies $[X|U, S = 0] = [X|U, S = 1]$
- Regression parameters for $[X|U, S = 0]$ are the same as for $S = 1$
- "Inverse regression estimator"

BONUS SLIDE: How the PPMM Identification Works

Assumed model for U and X given S: $(U, X|S = j) \sim$ Bivariate Normal
Assumed response mechanism:

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

If $\phi = 0 \rightarrow$ response only depends on X (not U)

- Implies $[U|X, S = 0] = [U|X, S = 1]$
- Regression parameters for $[U|X, S = 0]$ are the same as for $S = 1$
- Standard regression estimator (e.g., under MAR assumption)

If $\phi = 1 \rightarrow$ response only depends on U (not X)

- Implies $[X|U, S = 0] = [X|U, S = 1]$
- Regression parameters for $[X|U, S = 0]$ are the same as for $S = 1$
- "Inverse regression estimator"

If $0 < \phi < 1$, let $W = (1 - \phi)X^* + \phi U$ and $[X|W, S = 0] = [X|W, S = 1]$