Fred Hutch
Cancer Center

What’s the Weight?

Estimating Controlled Outcome Differences in Complex Surveys for Health Disparities Research

Stephen Salerno
Fred Hutchinson Cancer Center

JSM, Portland, OR
August 4, 2024

1 Background & Motivation

2 Proposed Method

3 Simulations

4 Race & Telomere Length

5 Conclusions

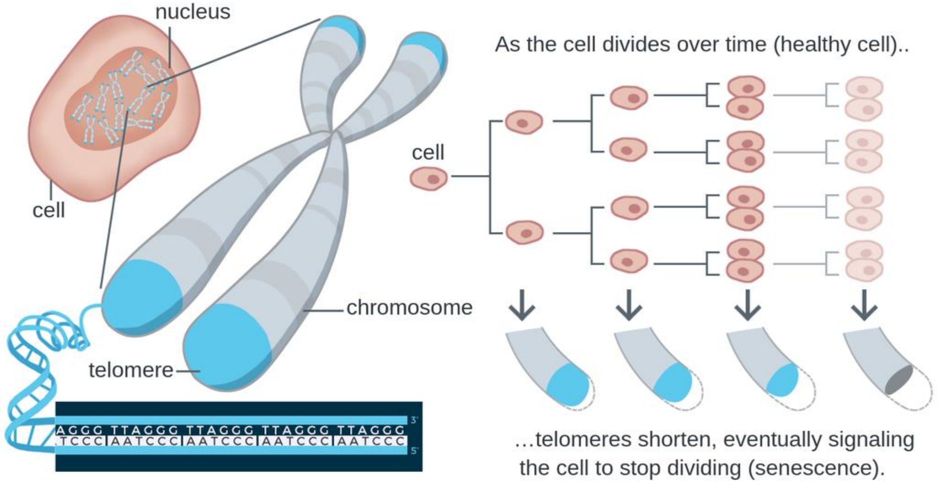
Our Motivation

Telomere Length and its Relationship with Race and SES

- ***Regions of DNA*** at the ends of chromosomes that protect against cell death
- ***Shortening*** associated w/ ***cardiometabolic outcomes***
- Affected by age, sex, ***race/ethnicity,*** genetics, SES, environment, psychosocial stress, …
- ***Longer*** telomeres in Black individuals (paradox)

  BUT

- ***Comparable length*** in populations w/ similar SES


nucleus
cell
cell
chromosome
telomere
AGGG TTAGGG TTAGGG TTAGGG TTAGGG
TCCC AATCCC AATCCC AATCCC AATCCC
3'
5'
As the cell divides over time (healthy cell)..
…telomeres shorten, eventually signaling the cell to stop dividing (senescence).

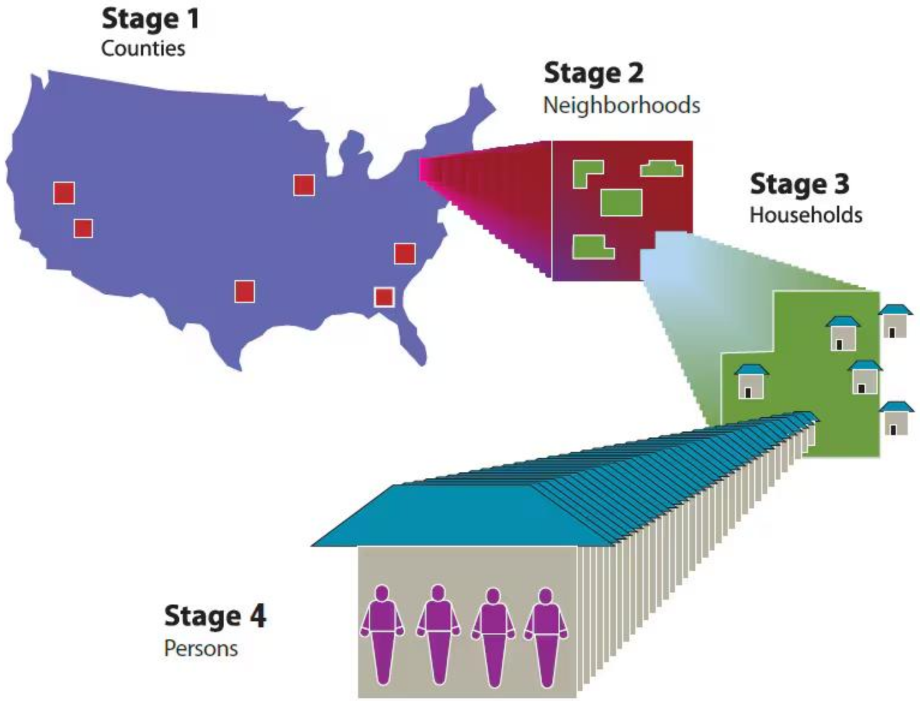
Credit: theory.labster.com/telomere-length

“If we could hypothetically ***balance*** SES between Black and White individuals in a ***nationally representative sample***, would we still see significant Black/White ***differences*** in telomere length?”

Our Data

National Health and Nutrition Examination Survey

- ***Nationally representative survey*** by the CDC
- ***Rich data*** from interviews, physical examinations, laboratory tests, …
- Stratified, clustered ***complex design***:
  - ***Primary sampling units*** (counties)
  - Drawn from ***demographic-specific strata***
  - Oversamples ***non-Hispanic Black*** participants and ***≤ 130% poverty limit***



Credit: cdc.gov/nchs/nhanes/

Our Problem

Confounding + Selection Bias + Design

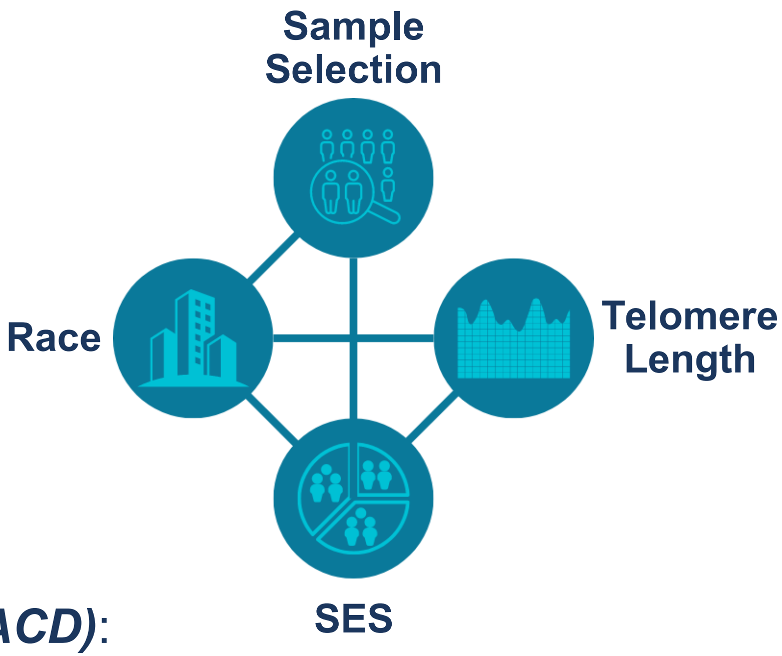
- ***Observational data*** often limited by ***confounding, covariate imbalance, lack of representation***
- ***Generalizing*** results while accounting for ***confounding*** difficult due to ***complex survey designs***
- This question is ***statistically*** challenging because:
  - Characteristic of interest ***(race)*** is ***correlated*** with ***SES***
  - ***Both factors*** influence the probability of ***being sampled***

Our Approach

Notation and Target of Inference

Survey of *n* participants from a ***super population*** of *N* individuals:

- *A*: ***Groups*** of Interest (race; 1 = Black, 0 = White)
- *X*: ***Confounders*** (SES)
- *Y*: ***Outcome*** (log telomere length)
- *S*: Sample ***Selection Indicator***



Want to estimate the ***population average controlled difference (ACD)***:

$$\text{ACD} = E_X[E(Y \mid A = 1, X) - E(Y \mid A = 0, X)]$$

Our Approach

Identification Formulas

- ***Questions***: Do you survey weight the propensity model? How to weight the outcome?
- ***Answer:*** Depends on factorization of ***selection*** (S = 1) and ***group membership*** (A = a) probability:

$$(1) \quad \mathbb{E}_X\left[\mathbb{E}[Y \mid A = a, S = 1, X] \cdot \frac{\Pr(S = 1)}{\Pr(S = 1 \mid X)} \mid S = 1\right]$$

$$(2) \quad \mathbb{E}_X\left[\frac{AY}{\Pr(A = a \mid X)} \cdot \frac{\Pr(S = 1)}{\Pr(S = 1 \mid A = a, X)} \mid S = 1\right]$$

$$(3) \quad \mathbb{E}_X\left[\frac{AY}{\Pr(A = a \mid S = 1, X)} \cdot \frac{\Pr(S = 1)}{\Pr(S = 1 \mid X)} \mid S = 1\right]$$

Estimate via ***g-formula*** (1) or ***inverse probability weighting*** (2, 3)

Either we ***weight*** our ***propensity score*** and specifically take ***selection given A = a***

Or we fit a ***within-sample*** propensity score and ***marginalize A out*** of the selection probability

Assumptions

ACD versus PATE

To estimate the ACD, ***we assume***:

1. ***Positivity***: $\Pr(A = a \mid X = x) > 0 \ \forall a \in \mathcal{A}$ and every x s.t. $f_X(x) > 0$
2. ***Selection Positivity***: $\Pr(S = 1 \mid A = a, X = x) > 0$ for every a, x s.t. $f_{A,X}(a, x) > 0$
3. ***Weak Selection Exchangeability***: $E[Y \mid A = a, X] = E[Y \mid A = a, S = 1, X]$.

Note: Can target ***population potential outcome means,*** $E[Y^a]$, with ***stronger assumptions***:

$\overset{*}{3}$. ***Weak Selection Exchangeability***: $E[Y^a \mid A = a, X] = E[Y^a \mid A = a, S = 1, X]$

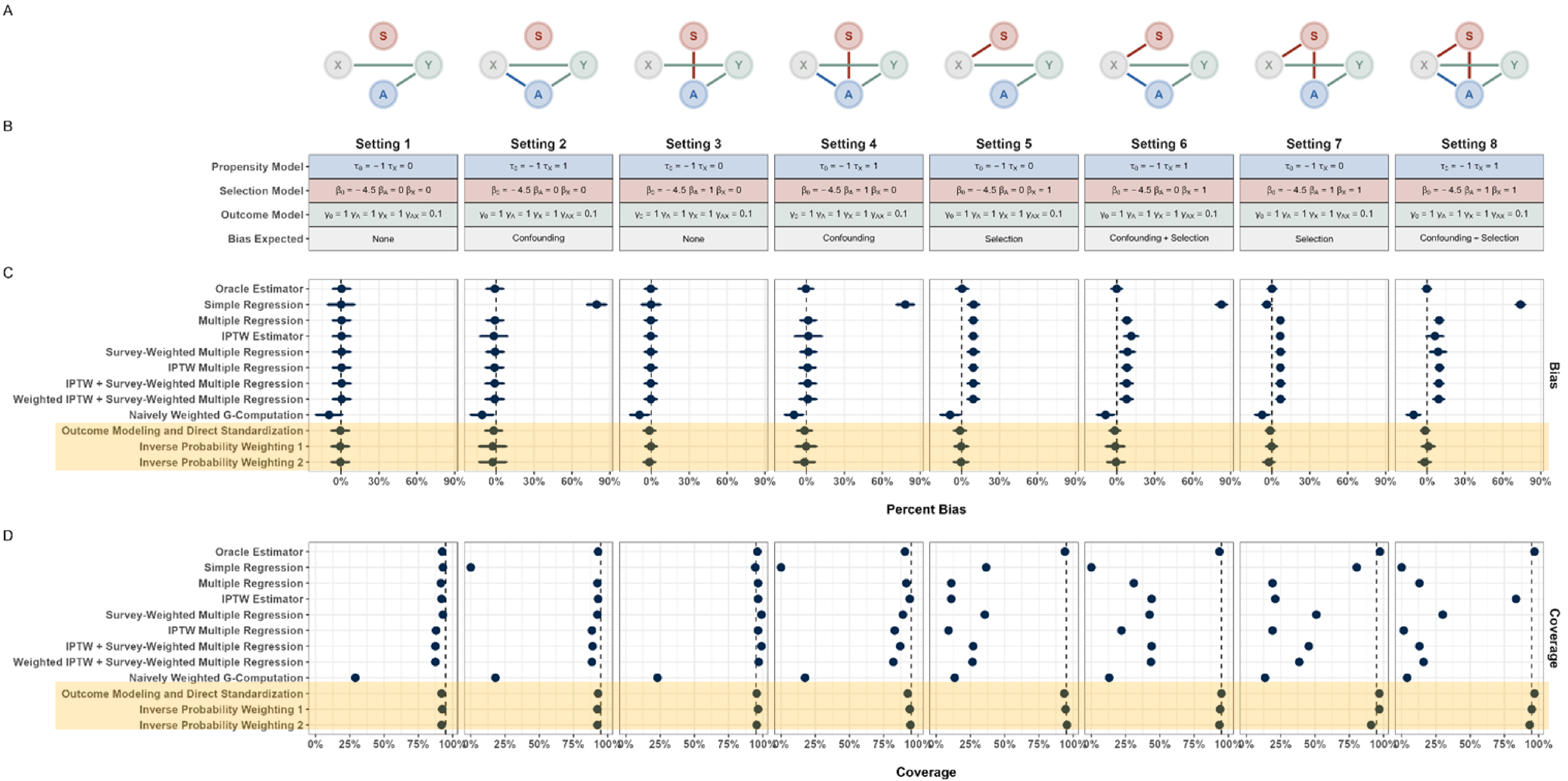
4. ***Stable Unit Treatment Value Assumption*** (SUTVA; No Interference + Consistency)
5. ***Weak Exchangeability***: $E[Y^a \mid X] = E[Y^a \mid A = a, X]$

Comparison of Methods

When to Use Each Approach

- Existing methods that do not account for ***confounding and selection*** will be ***biased***
- Weights are ***not the same*** as traditional g-computation or IPTW, we derive ***new estimators***
  - (1) and (3) require selection be ***marginalized*** over *A* and a ***within-sample*** propensity score
  - (2) requires ***group-specific selection probabilities*** and ***survey-weighted*** propensity score
- G-computation is ***most efficient*** if correctly specified, but IPWs ***more robust***
- In practice, even if sampling weights are given, may not know the ***true sampling mechanism***
  - Can model the survey weights via ***beta or simplex regression***

Simulation Results



	Setting 1	Setting 2	Setting 3	Setting 4	Setting 5	Setting 6	Setting 7	Setting 8
Propensity Model	$\tau_0 = -1\ \tau_X = 0$	$\tau_0 = -1\ \tau_X = 1$	$\tau_0 = -1\ \tau_X = 0$	$\tau_0 = -1\ \tau_X = 1$	$\tau_0 = -1\ \tau_X = 0$	$\tau_0 = -1\ \tau_X = 1$	$\tau_0 = -1\ \tau_X = 0$	$\tau_0 = -1\ \tau_X = 1$
Selection Model	$\beta_0 = -4.5\ \beta_A = 0\ \beta_X = 0$	$\beta_0 = -4.5\ \beta_A = 0\ \beta_X = 0$	$\beta_0 = -4.5\ \beta_A = 1\ \beta_X = 0$	$\beta_0 = -4.5\ \beta_A = 1\ \beta_X = 0$	$\beta_0 = -4.5\ \beta_A = 0\ \beta_X = 1$	$\beta_0 = -4.5\ \beta_A = 0\ \beta_X = 1$	$\beta_0 = -4.5\ \beta_A = 1\ \beta_X = 1$	$\beta_0 = -4.5\ \beta_A = 1\ \beta_X = 1$
Outcome Model	$\gamma_0 = 1\ \gamma_A = 1\ \gamma_X = 1\ \gamma_{AX} = 0.1$	$\gamma_0 = 1\ \gamma_A = 1\ \gamma_X = 1\ \gamma_{AX} = 0.1$	$\gamma_0 = 1\ \gamma_A = 1\ \gamma_X = 1\ \gamma_{AX} = 0.1$	$\gamma_0 = 1\ \gamma_A = 1\ \gamma_X = 1\ \gamma_{AX} = 0.1$	$\gamma_0 = 1\ \gamma_A = 1\ \gamma_X = 1\ \gamma_{AX} = 0.1$	$\gamma_0 = 1\ \gamma_A = 1\ \gamma_X = 1\ \gamma_{AX} = 0.1$	$\gamma_0 = 1\ \gamma_A = 1\ \gamma_X = 1\ \gamma_{AX} = 0.1$	$\gamma_0 = 1\ \gamma_A = 1\ \gamma_X = 1\ \gamma_{AX} = 0.1$
Bias Expected	None	Confounding	None	Confounding	Selection	Confounding + Selection	Selection	Confounding − Selection

A
S
X
Y
A
B
C
Oracle Estimator
Simple Regression
Multiple Regression
IPTW Estimator
Survey-Weighted Multiple Regression
IPTW Multiple Regression
IPTW + Survey-Weighted Multiple Regression
Weighted IPTW + Survey-Weighted Multiple Regression
Naively Weighted G-Computation
Outcome Modeling and Direct Standardization
Inverse Probability Weighting 1
Inverse Probability Weighting 2
0%
30%
60%
90%
Percent Bias
Bias
D
0%
25%
50%
75%
100%
Coverage
Coverage

Simulation Results

Our methods perform better in more complex settings

- **Settings 1 + 3: No Bias**

  Facilitates a fair comparison – All perform comparably w.r.t. bias, MSE, coverage

- **Settings 2 + 4 Confounding Bias**

  Simple linear regression performs poorly, all other methods that adjust for A and X perform well

- **Settings 5 + 7: Selection Bias**

  Approaches that do not survey weight have bias, poor coverage, other methods perform well

- **Settings 6 + 8: Confounding + Selection Bias**

  Assumed relationships in our data – our proposed estimators outperform all current approaches

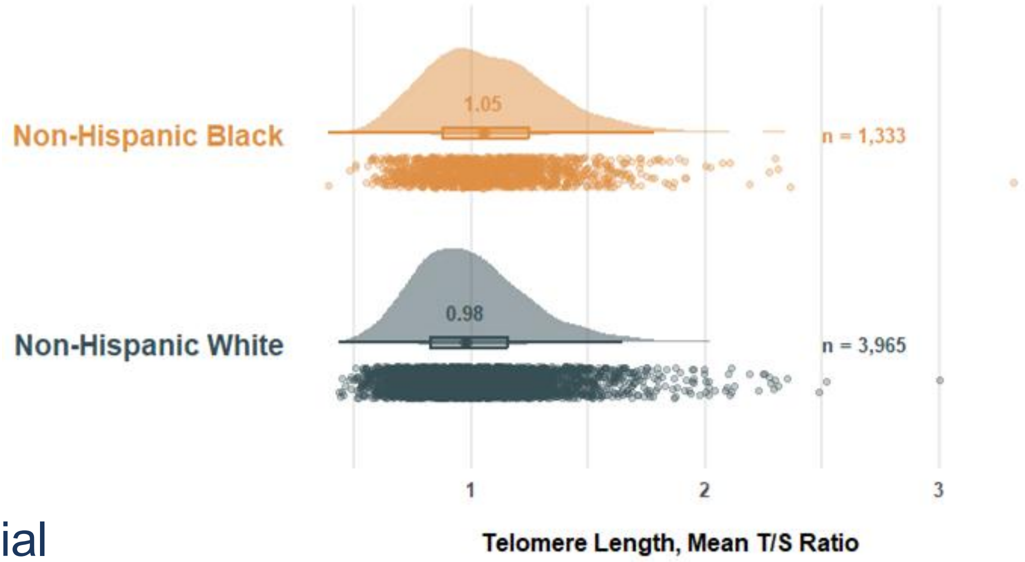
Our Data

National Health and Nutrition Examination Survey (NHANES)

- ***5,298 non-Hispanic Black/White*** adults from **1999 to 2002** with measured ***telomere length***
  - **12 socio-demographic indicators**:

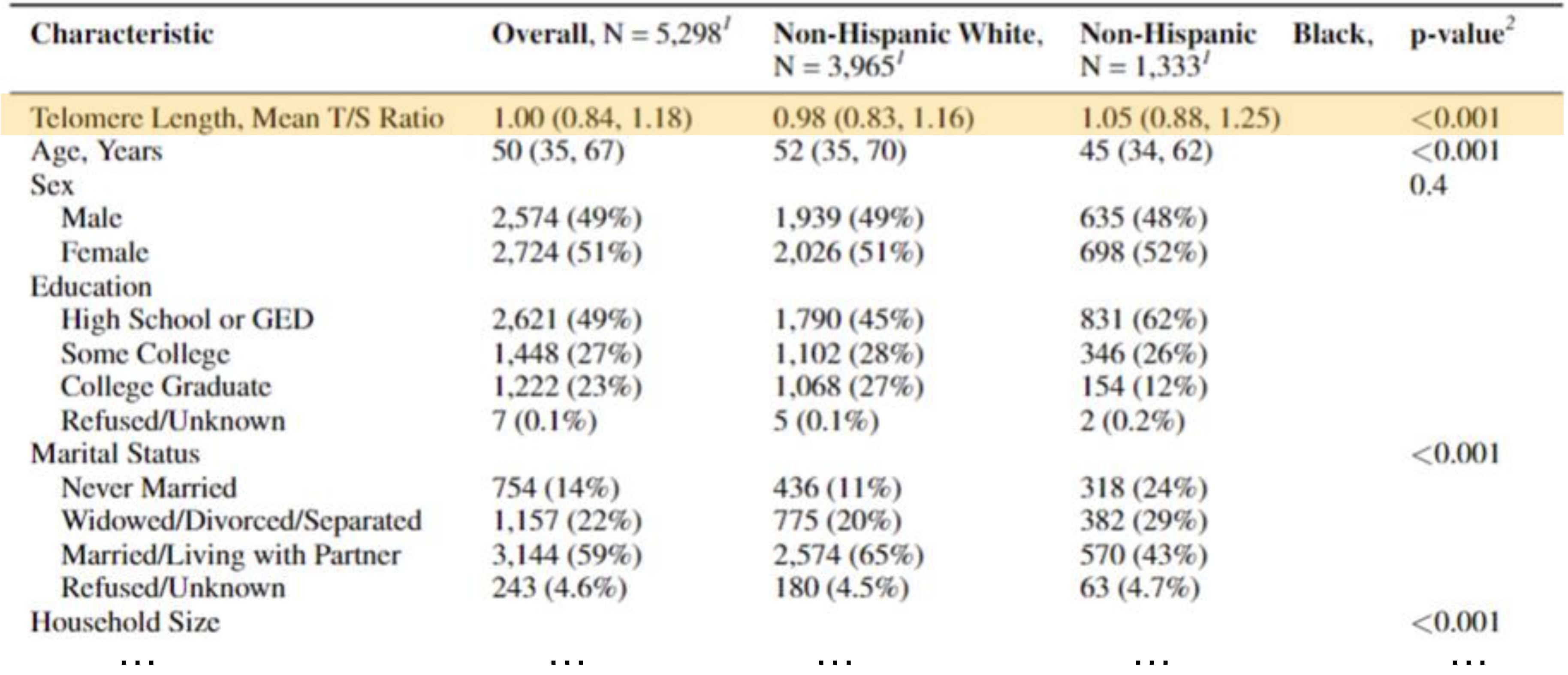
    Education, Marital Status, Household Size,
    Home Ownership, Home Type,
    Household Income, Poverty-Income Ratio,
    Employment Status, Occupational Category,
    Insurance Status, Food Security Status,
    WIC Utilization
  - ***8 precision covariates***:

    Age, Sex, White Blood Cell Count, 5-Part Differential



Descriptive Statistics

Univariate tests suggest Black/White differences across all socioeconomic indicators

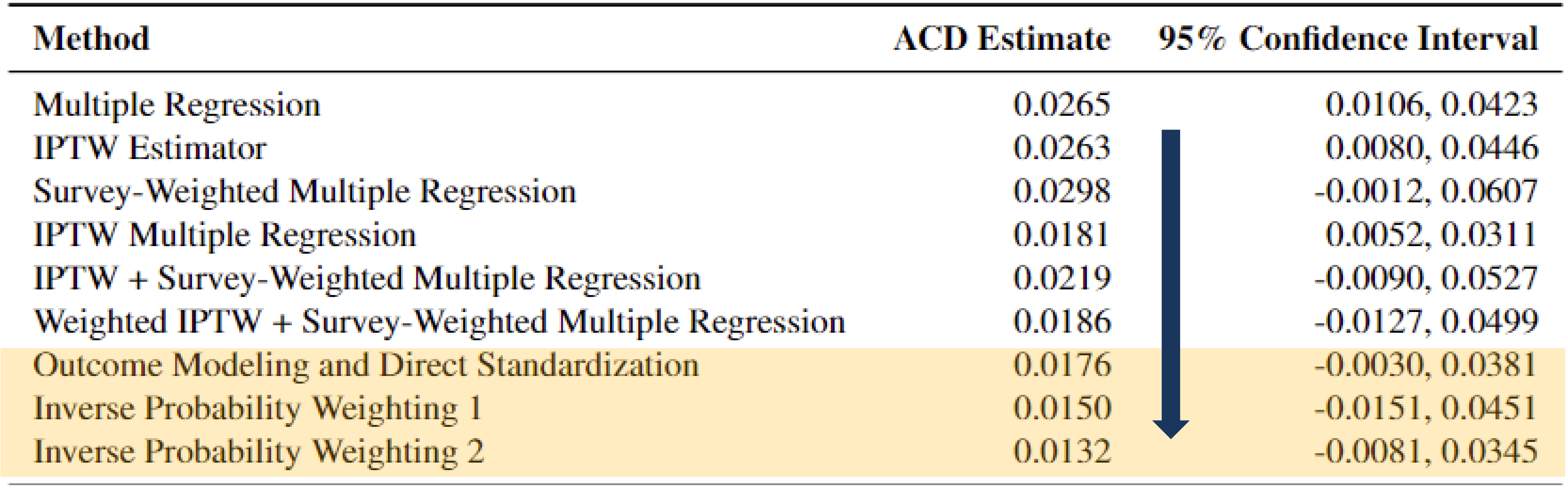


Characteristic	Overall, N = 5,298[1]	Non-Hispanic White, N = 3,965[1]	Non-Hispanic Black, N = 1,333[1]	p-value[2]
Telomere Length, Mean T/S Ratio	1.00 (0.84, 1.18)	0.98 (0.83, 1.16)	1.05 (0.88, 1.25)	<0.001
Age, Years	50 (35, 67)	52 (35, 70)	45 (34, 62)	<0.001
Sex				0.4
Male	2,574 (49%)	1,939 (49%)	635 (48%)	
Female	2,724 (51%)	2,026 (51%)	698 (52%)	
Education				
High School or GED	2,621 (49%)	1,790 (45%)	831 (62%)	
Some College	1,448 (27%)	1,102 (28%)	346 (26%)	
College Graduate	1,222 (23%)	1,068 (27%)	154 (12%)	
Refused/Unknown	7 (0.1%)	5 (0.1%)	2 (0.2%)	
Marital Status				<0.001
Never Married	754 (14%)	436 (11%)	318 (24%)	
Widowed/Divorced/Separated	1,157 (22%)	775 (20%)	382 (29%)	
Married/Living with Partner	3,144 (59%)	2,574 (65%)	570 (43%)	
Refused/Unknown	243 (4.6%)	180 (4.5%)	63 (4.7%)	
Household Size				<0.001
…	…	…	…	…

Our Study

Results

ACD estimates and corresponding 95% confidence intervals across various analytic approaches for the comparison of effect of race (non-Hispanic Black versus non-Hispanic White participants) on log-transformed telomere length among n = 5,270 NHANES participants (complete case -28 participants)



Method	ACD Estimate	95% Confidence Interval
Multiple Regression	0.0265	0.0106, 0.0423
IPTW Estimator	0.0263	0.0080, 0.0446
Survey-Weighted Multiple Regression	0.0298	-0.0012, 0.0607
IPTW Multiple Regression	0.0181	0.0052, 0.0311
IPTW + Survey-Weighted Multiple Regression	0.0219	-0.0090, 0.0527
Weighted IPTW + Survey-Weighted Multiple Regression	0.0186	-0.0127, 0.0499
Outcome Modeling and Direct Standardization	0.0176	-0.0030, 0.0381
Inverse Probability Weighting 1	0.0150	-0.0151, 0.0451
Inverse Probability Weighting 2	0.0132	-0.0081, 0.0345

Data Analysis Conclusions

From Comparison of Approaches

- ACD estimates ***attenuate*** as we make ***appropriate adjustments*** for confounding + selection bias:
    - ***Linear Regression***: ACD estimate of 0.0265 (95% CI: 0.0106-0.0423)
    - ***Proposed Approaches***: 0.0132 to 0.0176, failed to detect a statistically significant difference
- These results suggest there is a ***confounding relationship*** between ***SES and race***
- Methods which ***properly*** incorporated the NHANES ***stratified, clustered design*** tended to have more ***conservative standard error estimates*** than those which did not

Conclusions

Some thoughts on the approach and our results

- Approach for estimating ***controlled outcome differences*** when the group variable of interest and its confounders affect ***sample selection***
- Proposal ***minimizes bias*** and achieves ***correct inference*** compared to standard analysis methods
- Context of studying ***racial disparities*** presents these particularities in such a way that should be ***rigorously studied*** for ***best practice*** recommendations

Conclusions

Thoughts on Future Work

- Though we focus on ***complex survey designs,*** these concepts readily extend to other settings where ***observational data*** are collected via an underlying ***selection mechanism***
- Areas of interest for ***future work*** include:
    - ***Electronic health record*** data with unknown sampling probabilities
    - Expanded ***relationship diagrams*** or sampling designs
- Extending this framework to ***two-stage sampling*** or ***sequential designs***
- Expand on method with ***doubly robust, AIPW estimator***

Our Group

Stephen Salerno

Fred Hutchinson Cancer Center
Division of Public Health Sciences

Emily Roberts

University of Iowa
Department of Biostatistics

Belinda Needham

University of Michigan
Department of Epidemiology

Tyler McCormick

University of Washington
Department of Statistics
Department of Sociology

Bhramar Mukherjee

Yale University
Department of Biostatistics
Department of Epidemiology
Department of Statistics and Data Science

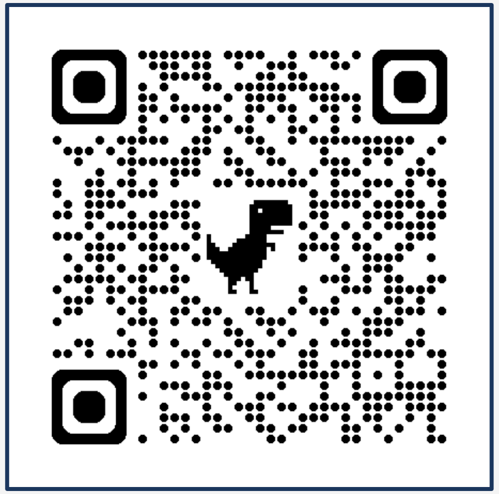
Xu Shi

University of Michigan
Department of Biostatistics

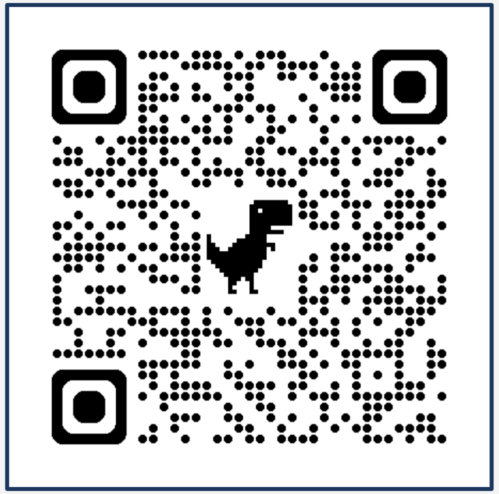
Fred Hutch
Cancer Center

Thank You!

Fred Hutch
Cancer Center



Paper



R Package

Land Acknowledgement

Fred Hutchinson Cancer Center acknowledges the Coast Salish peoples of this land, the land which touches the shared waters of all tribes and bands within the Duwamish, Puyallup, Suquamish, Tulalip and Muckleshoot nations.

Estimation

Outcome Modeling and Direct Standardization

- Define the ***functional*** μ(a) for ID1 and let $\hat{g}_a(X)$ be an estimator for $E[Y \mid A = a, S = 1, X]$
  - Can be estimated with standard ***parametric*** model, $g_a(X; \gamma)$ with finite-dimensional parameter γ
- Our ***outcome model-based estimator*** for μ(a) is given by

$$\hat{\mu}_{\mathrm{ID1}}(a) = \frac{1}{n} \sum_{i=1}^{n} \hat{g}_a(X_i) \frac{\hat{\pi}^0}{\hat{\pi}(X_i)}$$

where π's are estimators of the ***sample selection probabilities***

- If $g_a(X; \gamma)$ is correctly specified, then $\hat{g}_a(X; \gamma) \rightarrow_p E[Y \mid A = a, S = 1, X]$, and $\hat{\mu}_{ID1}(a) \rightarrow_p \mu(a)$ for $a \in A$

Estimation

Inverse Probability Weighting

- ID2A and ID2B rely on ***inverse probability weighting*** to control for confounding, where $\hat{e}_a(X)$ is the within-sample and $\hat{e}_a^w(X)$ is the survey-weighted ***propensity model*** estimator
- We can write out the IP-weighting estimator as either

$$\hat{\mu}_{\mathrm{ID2A}}(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(\mathcal{A} = a)\hat{\pi}^0 Y}{\hat{e}_a^w(X_i)\hat{\pi}(A_i, X_i)} \quad \text{or}$$

$$\hat{\mu}_{\mathrm{ID2B}}(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(\mathcal{A} = a)\hat{\pi}^0 Y}{\hat{e}_a(X_i)\hat{\pi}(X_i)}.$$

- Which are also consistent if the propensity and selection models are correctly specified

Causal Inference Target

Population Average Treatment Effect

- Can also target ***potential outcome means***, i.e., $E[Y^a]$ for $a \in A$ and the ***population average treatment effect*** (*PATE*), defined as the expected difference in counterfactual outcomes:

$$PATE = E[Y^1 - Y^0]$$

- ***In general***, $E[Y^1 - Y^0] \neq E[Y^1 - Y^0 \mid S=1]$; sample may not be ***representative*** of the larger population
- ***Goal***: Derive an ***identification formula*** for $E[Y^a]$, the potential outcome means
- ***Same as our main result*** for the ACD, but under stronger assumptions

Our Approach

Inference

- Inference can be carried out analytically via **Taylor expansion**, accounting for all variation, or via **numerical estimation via** the general theory from **M-estimation**
- Let θ(P) denote the **parameter vector** arising from the series of **estimating equations** and let ϕ_i denote the corresponding **influence function** for the ith individual
- Using **numerical optimization**, the **covariance matrix** of the **estimated parameters** is given by

$$\hat{V}(\hat{\theta}) = \sum_{i=1}^{n} \phi_i(y, \hat{\theta}, \mathcal{P}) \phi_i(y, \hat{\theta}, \mathcal{P})'$$

A Note on Inference

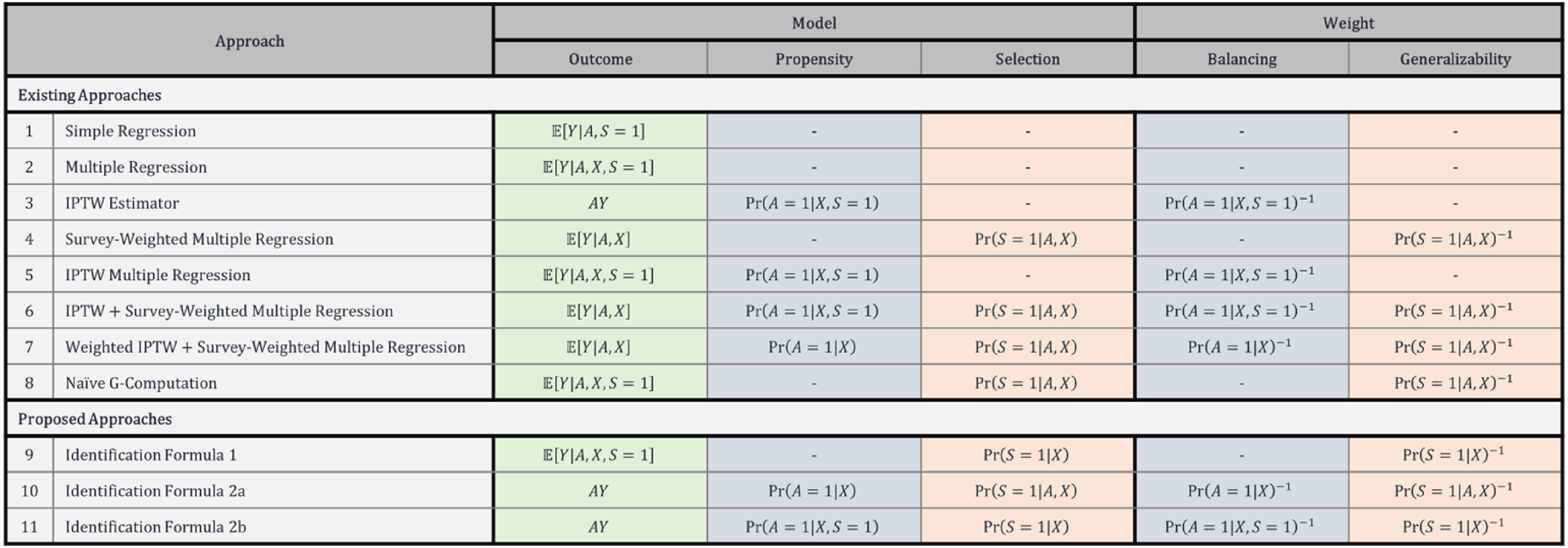
Modified Sandwich Variance Estimator

- While **bias** incurred from differential sampling probabilities is **accounted for** in estimation, inference is affected by **correlation** between individuals within strata and primary sampling units (PSUs)
- Modify **sandwich variance estimator** by summing contributions of the ith individual (i = 1, ..., nj), in the jth PSU (j = 1, ..., Jk), in the kth sampling stratum (k = 1, ..., K)
- Estimated **covariance matrix** for our **parameters** expressed in terms of the variability of the between PSU-level sums of the first order **Taylor approximations** within the sampling strata

$$\sum_{k=1}^{K} \frac{J_k}{J_k - 1} \sum_{j=1}^{J_k} \left\{\phi_{\cdot jk}\left(y, \hat{\theta}, \mathcal{P}\right) - \bar{\phi}_{\cdot\cdot k}\left(y, \hat{\theta}, \mathcal{P}\right)\right\} \left\{\phi_{\cdot jk}\left(y, \hat{\theta}, \mathcal{P}\right) - \bar{\phi}_{\cdot\cdot k}\left(y, \hat{\theta}, \mathcal{P}\right)\right\}'$$

Comparison of Methods

Existing Strategies for Analysis versus Our Approaches



Approach		Model			Weight	
		Outcome	Propensity	Selection	Balancing	Generalizability
Existing Approaches						
1	Simple Regression	$\mathbb{E}[Y\|A,S=1]$	-	-	-	-
2	Multiple Regression	$\mathbb{E}[Y\|A,X,S=1]$	-	-	-	-
3	IPTW Estimator	AY	$\Pr(A=1\|X,S=1)$	-	$\Pr(A=1\|X,S=1)^{-1}$	-
4	Survey-Weighted Multiple Regression	$\mathbb{E}[Y\|A,X]$	-	$\Pr(S=1\|A,X)$	-	$\Pr(S=1\|A,X)^{-1}$
5	IPTW Multiple Regression	$\mathbb{E}[Y\|A,X,S=1]$	$\Pr(A=1\|X,S=1)$	-	$\Pr(A=1\|X,S=1)^{-1}$	-
6	IPTW + Survey-Weighted Multiple Regression	$\mathbb{E}[Y\|A,X]$	$\Pr(A=1\|X,S=1)$	$\Pr(S=1\|A,X)$	$\Pr(A=1\|X,S=1)^{-1}$	$\Pr(S=1\|A,X)^{-1}$
7	Weighted IPTW + Survey-Weighted Multiple Regression	$\mathbb{E}[Y\|A,X]$	$\Pr(A=1\|X)$	$\Pr(S=1\|A,X)$	$\Pr(A=1\|X)^{-1}$	$\Pr(S=1\|A,X)^{-1}$
8	Naïve G-Computation	$\mathbb{E}[Y\|A,X,S=1]$	-	$\Pr(S=1\|A,X)$	-	$\Pr(S=1\|A,X)^{-1}$
Proposed Approaches						
9	Identification Formula 1	$\mathbb{E}[Y\|A,X,S=1]$	-	$\Pr(S=1\|X)$	-	$\Pr(S=1\|X)^{-1}$
10	Identification Formula 2a	AY	$\Pr(A=1\|X)$	$\Pr(S=1\|A,X)$	$\Pr(A=1\|X)^{-1}$	$\Pr(S=1\|A,X)^{-1}$
11	Identification Formula 2b	AY	$\Pr(A=1\|X,S=1)$	$\Pr(S=1\|X)$	$\Pr(A=1\|X,S=1)^{-1}$	$\Pr(S=1\|X)^{-1}$

Simulations

Setup

- **Covariate**: $X \sim N(1,1)$
- **Group Variable**: $A \mid X \sim Bin(N, p_A)$ where $p_A = logit^{-1}(\tau_0 + \tau_X X)$
- **Sampling Indicator**: $S \mid A, X \sim Bin(N, p_S)$ where $p_S = logit^{-1}(\beta_0 + \beta_A A + \beta_X X + \varepsilon_S)$ and $\varepsilon_S \sim N(0, 0.1)$
- We consider a **heterogeneous effect** with $\varepsilon_O \sim N(0, 1)$:

$$Y = \gamma_0 + \gamma_A A + \gamma_X X + \gamma_{AX} AX + \varepsilon_O$$

- Then take *nsims = 200* random samples based on our sampling indicator, *S*