Improving External Validity of Association Estimation Using Kernel Weighting Approach

Lingxiao Wang

Joint Program in Survey Methodology, University of Maryland, College Park, MD, U.S.A
National Cancer Institute, DCEG, Biostatistics Branch, Rockville, MD, U.S.A

March, 11, 2019

Overview

Introduction

Estimates of Association Using Unrepresentative Sample?

$$\begin{aligned} &\text{Superpopulation:} && y \sim f(y|x;\beta) \\ &\text{Fit:} && g(\mu) = \beta x, \quad \mu = E(y) \end{aligned}$$

- **? Small bias**
  - Confounders are controlled in the outcome model (Pizzi et al., 2011; Richiardi et al., 2013)
- **? Large bias**
  - Limited availability of confounders

Introduction

Biased Naive Sample Estimates of Association

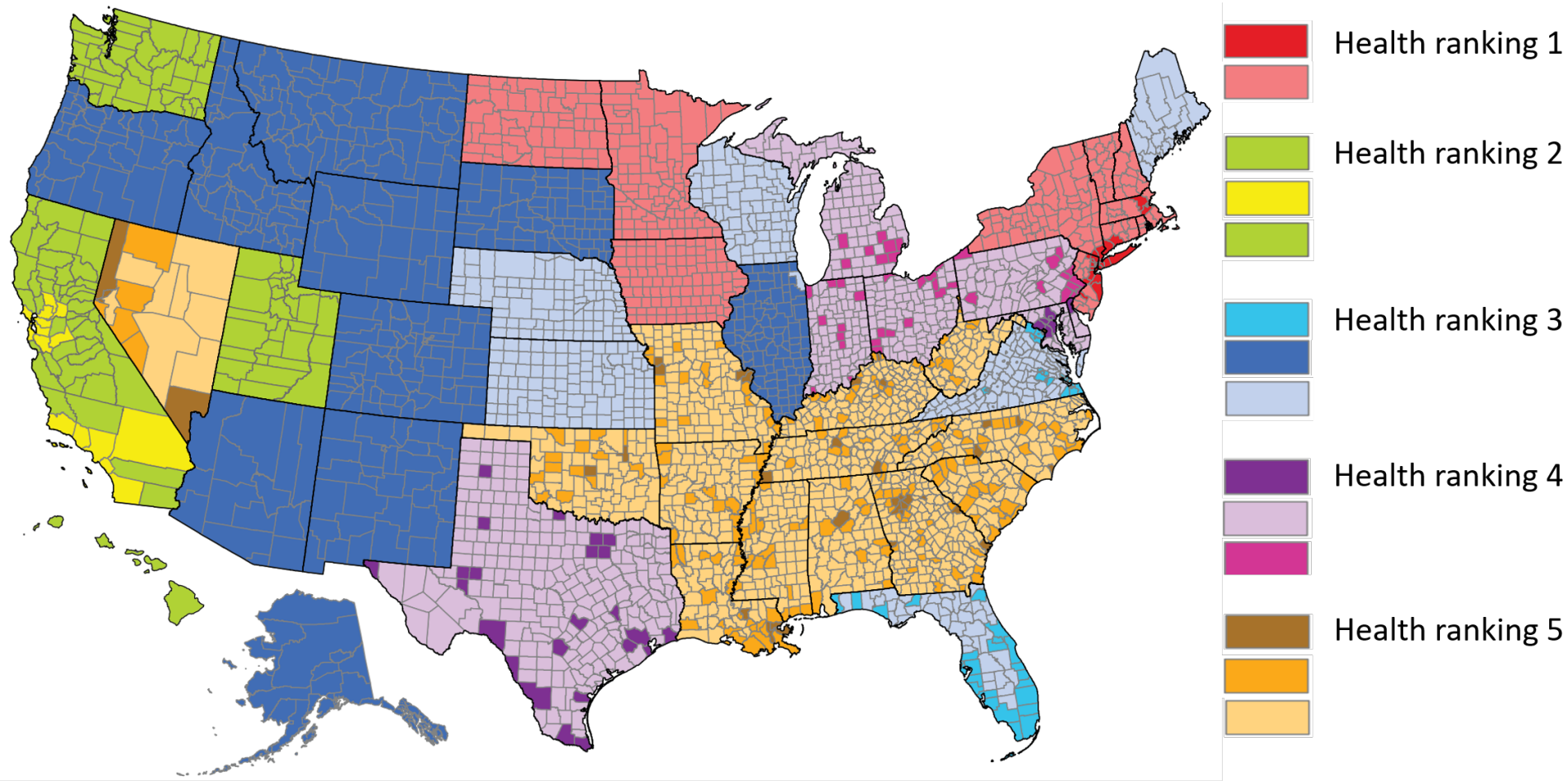
In probability-based survey samples...

Informative sample selection (Fuller, 1999)

- Distributions of outcome variable in the sample and population are different given covariates in the **true** outcome model

$$f(y|x;\beta,s) \neq f(y|x;\beta,P), \qquad \begin{array}{ll} \text{Biased} & \hat{\beta}_0 \\ \text{Unbiased} & \hat{\beta}_w \end{array}$$

Example NHANES (2011-2014)



Source: Johnson et al. (2014)

Death rate, infant mortality, % high blood pressure, % overweight/obese, % smokers

Biased Naive Sample Estimates of Association

In probability-based survey samples...

Non-informative sample selection

- Outcome variable is independent from the design variables (sample selection) given the covariates in the true outcome model.
  - Correctly specified outcome model

$$f(y|x;\beta,s) = f(y|x;\beta,P)$$

Unbiased $\hat{\beta}_0$
Unbiased $\hat{\beta}_w$

  - Mis-specified outcome model (Korn & Graubard, 1999)

Fit: $y \sim f^{'}(y|x;B)$

$$f^{'}(y|x;B,s) \neq f^{'}(y|x;B,P)$$

Biased $\hat{B}_0$
Unbiased $\hat{B}_w$

Example NMIHS (1998)

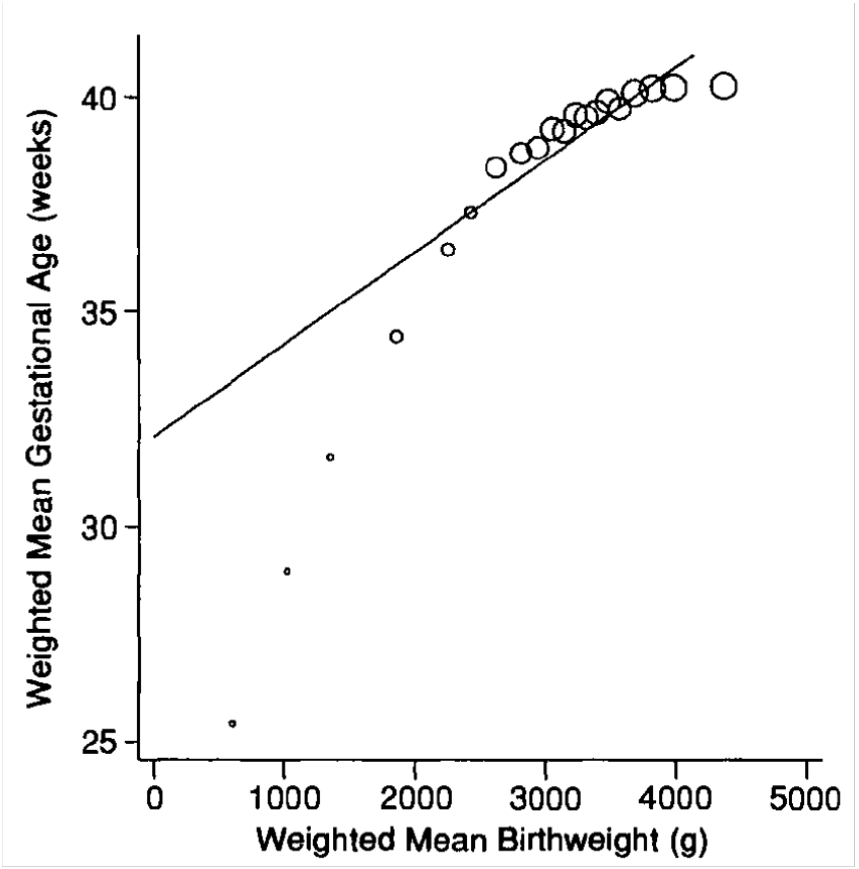
1998 National Maternal and Infant Health Survey (NMIHS)

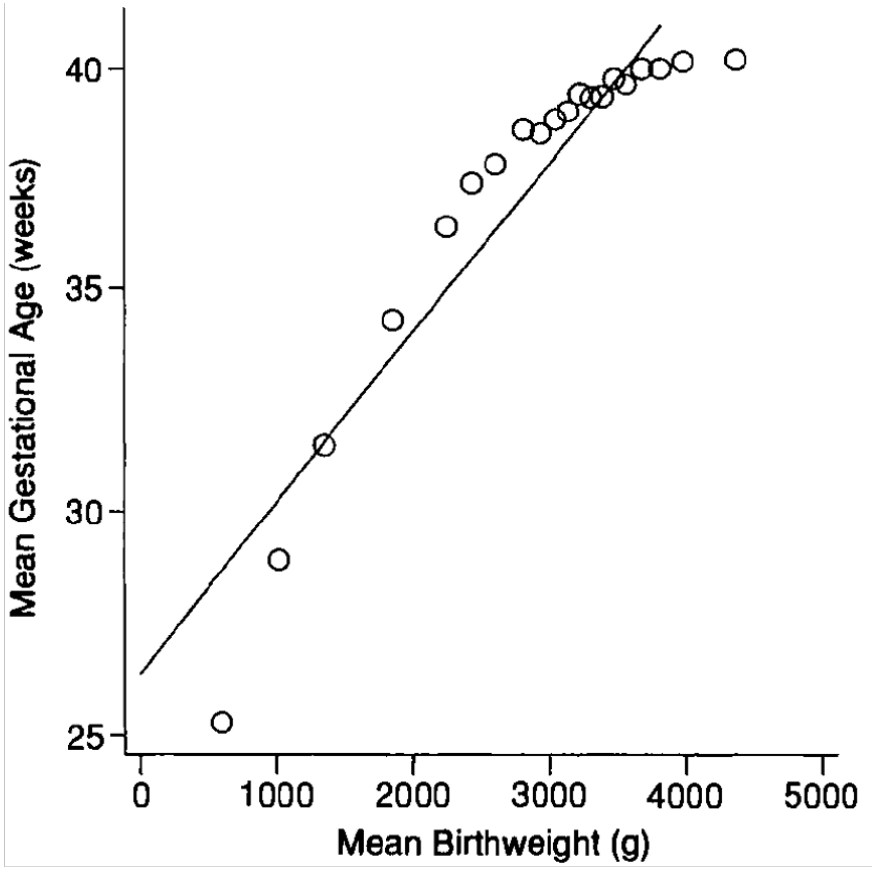
- Stratified simple random sample
- Strata formed using **states**, mothe's **race** and baby's **birth weight**.

Race and birthweight	Sample size	Inverse of probability of selection	Average weight[2]
Total	13,417	...	...
Black			
Less than 1,500 grams	1,296	14	21.8822611
1,500–2,499 grams	194	55	84.24777
2,500 grams or more	4,948	113	161.8481219
Other than black			
Less than 1,500 grams	951	29	41.9928622
1,500–2,499 grams	938	160	214.8166309
2,500 grams or more	4,090	720	923.3830125

Source: Sanderson et al. (1998)

Example NMIHS (1998)- Cont'd





Source: Korn & Graubard (1999)

$\hat{B}_w$: estimates a population quantity.

$\hat{B}_0$: fits the unweighted sample better, **but will change depending on sample design!**

Introduction

Representative Survey Sample for Estimating Association

- **Informative sample-selection**

  Weights always need to be incorporated for design consistent estimates of association

- **Non-informative sample-selection**

  - Correct outcome model

    Naive estimates of association are consistent

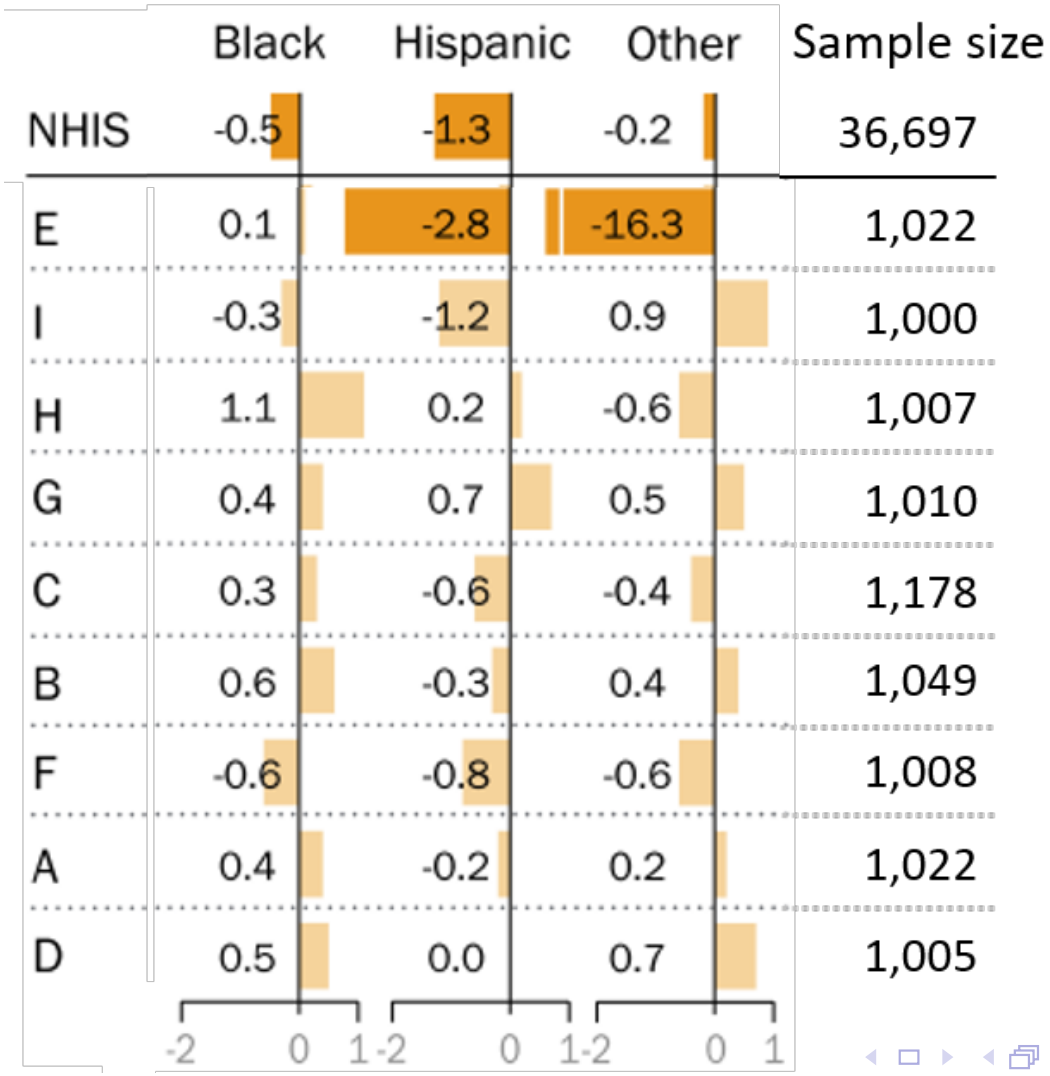
  - Mis-specified outcome model

    Naive estimates: biased for population quantities, unreproducible
    Weighted estimates: consistent

Estimating Association Using Non-probability Samples?

Non-probability Samples? (Source: Kennedy et al., 2016)

logit{P(smoking daily)}∼age+education+region+sex+race



Introduction

Unrepresentative Non-probability Samples

Naive Sample Estimates of Association

- Bias
- Variance/ Type I error of hypothesis tests
- Limited Literature

Goal

- Improving external validity using weighting methods?
- Reduce bias?
- Valid variance estimates?

Method – A Quick Review

Three Propensity-Score-Based Weighting Methods

Method	Samples	Propensity score	Kernel function
IPSW	$s_c \cup U$	Estimating participation rates	—
PSAS	$s_c \cup s_s$	Measure of similarity	Stratified uniform
KW			Gaussian, triangular etc.

$$w_j^{KW} = \sum_{i \in s_s} k_{ij} \cdot w_i$$

$k_{ij} = \frac{K\left(d\left(x_i^{(s)}, x_j^{(c)}\right)/h\right)}{\sum_{j \in s_c} K\left(d\left(x_i^{(s)}, x_j^{(c)}\right)/h\right)}$ for $j \in s_c,\ i \in s_s$: proportionally distributes w_i to j

Consistent estimate of $\bar{Y}$.

Consistent estimate of association.

Simulations

Aims

In regression analyses,

1. Unrepresentative sample introduces bias?
2. Weighting methods reduce bias?
3. Variance?

Scenarios

1. Informative sample selection
   - Correct outcome model, correct propensity score model
2. Non-informative sample selection
   - Correct outcome model
   - Misspecified outcome model

     Correct & misspecified propensity score models

Simulations

Finite Population Generation

1. $M = 3,000$ clusters with size=3,000 (population size $N = 9 \times 10^6$)
2. Generate population variables
   - Race/ethnicity, age, sex, income, and urban/rural (2015 ACS)
   - Continuous exposure Env

Simulations – Informative Design

Finite Population Generation

- **Disease status** $y \sim Bernoulli(\mu)$

$$\mu = \{1 + \exp(-\beta_0 - \beta_1 age - \beta_2 Env)\}^{-1}$$

- Continuous variable $z = y - u, u \sim N(0, 0.01)$

Assemble the Survey Sample and Cohort

Two-stage PPS design with measures of size calculated from

- $q_i = \exp\{\gamma_0 + \gamma_1 Env_i + \gamma_2 z_i \cdot Env_i\}$ for cohort selection
- $s_i = \exp\{\frac{1}{2}(-\gamma_0 - \gamma_1 Env_i - \gamma_2 z_i \cdot Env_i)\}$ for survey sample selection

True propensity score model $logit\,(p) \sim Env + z_i \cdot Env$

Outcome model $logit\,(P\{y = 1\}) = \beta_0 + \beta_1 age + \beta_2 Env$

Ignoring the weights will introduce bias to estimate of β_2!

Simulations – Informative Design

$\hat{\beta}_2$ in Correct Outcome Model Using True PS Model

	Method			
	Naive	**IPSW**	**PSAS**	**KW**
Rel Bias(%)	-64.95	1.40	-34.68	0.18
Emp Var(10^2)	1.13	2.40	1.37	1.69
VarRatio(TL)	–	0.49	0.65	0.79
(JK)	–	1.01	1.07	1.06
MSE(10^2)	96.1	2.45	28.4	1.69

Simulations – Non-informative Design

Finite Population Generation

- **Disease status** $y \sim Bernoulli(\mu)$

$$\mu = \{1 + \exp(-\beta_0 - \beta_1 age - \beta_2 Env - \beta_3 age \cdot Env)\}^{-1}$$

- Continuous variable $z = \mu - u, u \sim N(0, 0.01)$

Assemble the Survey Sample and Cohort

Two-stage PPS design with measures of size calculated from

- $q_i = \exp\{\gamma_0 + \gamma_1 age_i + \gamma_2 z_i \cdot age_i\}$ for cohort selection
- $s_i = \exp\{\frac{1}{2}(-\gamma_0 - \gamma_1 age_i - \gamma_2 z_i \cdot age_i)\}$ for survey sample selection

True propensity score model $logit\,(p) \sim age + z_i \cdot age$

Outcome model $logit\,(P\{y = 1\}) = \beta_0 + \beta_1 age + \beta_2 Env + \beta_3 age \cdot Env$

Ignoring the weights will NOT introduce bias to estimates of β_1, β_2, or β_3!

Simulations – Non-informative Design

$\hat{\beta}$ in Correct Outcome Model Using True PS Model

Method	Relative Bias(%)			MSE (10^3)		
	β_1	β_2	β_3	β_1	β_2	β_3
Naive	0.46	-2.67	0.72	5.20	49.1	2.70
IPSW	1.60	-1.19	0.83	9.00	71.7	4.30
PSAS	1.30	-1.44	0.76	7.90	61.8	3.60
KW	1.54	-1.36	0.86	8.80	71.1	4.20

Simulations – Non-informative Design

True outcome model

$$logit\left(P\{y=1\}\right) = \beta_0 + \beta_1 age + \beta_2 Env + \beta_3 age \cdot Env$$

Underfitted outcome model

$$logit\left(P\{y=1\}\right) = \beta_0 + \beta_2 Env$$

Ignoring the weights will introduce bias to estimate of β_2!

	Method			
	Naive	**IPSW**	**PSAS**	**KW**
Rel Bias(%)	35.16	1.02	11.2	-0.13
Emp Var$\left(10^3\right)$	5.30	8.33	6.72	6.77
VarRatio(TL)	–	0.70	0.73	0.83
(JK)	–	1.00	1.20	1.05
MSE$\left(10^3\right)$	213.7	8.50	27.7	6.78

Mis-specified Propensity Score Model

True model M1(T)

$$logit(p) \sim age + z_i \cdot age$$

Overfitted model M2(O)

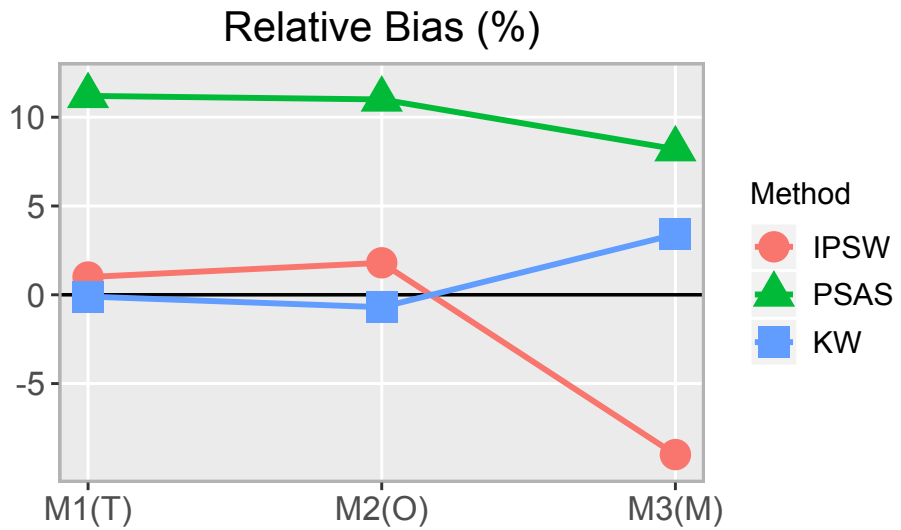
$$logit(p) \sim age + z_i \cdot age + Env + race$$

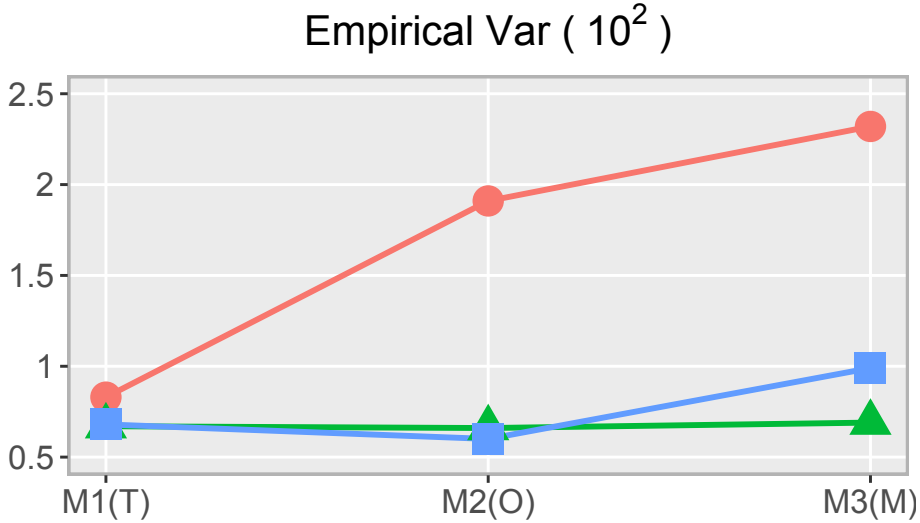
Mixed model M3(M)

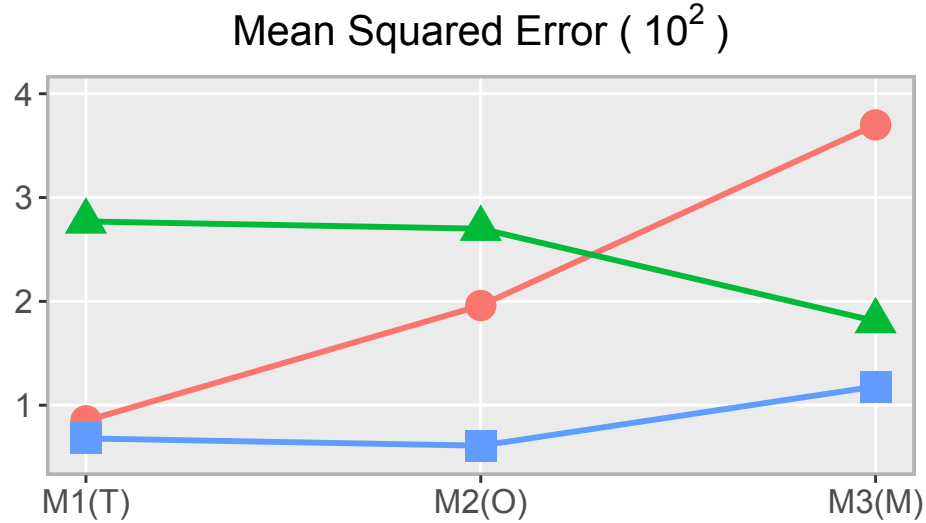
$$logit(p) \sim age + Env + Env \cdot age$$

Simulations – Non-informative Design

$\hat{\beta}_2$ in Mis-specified Outcome Model Using Difference PS Models







Data Example: NHANES III & NHIS 1994

Data Materials

- **Aim**

  Estimate association between obese and mortality among adults in U.S.

- **Data**

  1. The Third US National Health and Nutrition Examination Survey (NHANES III)

     From 1988-1994, adults ($n_c = 20,050$)
     Combining interviews and physical examinations.
     A nationwide probability sample
     $\widehat{N} = 187,647,206$. 49 strata, 2 PSU's per stratum.

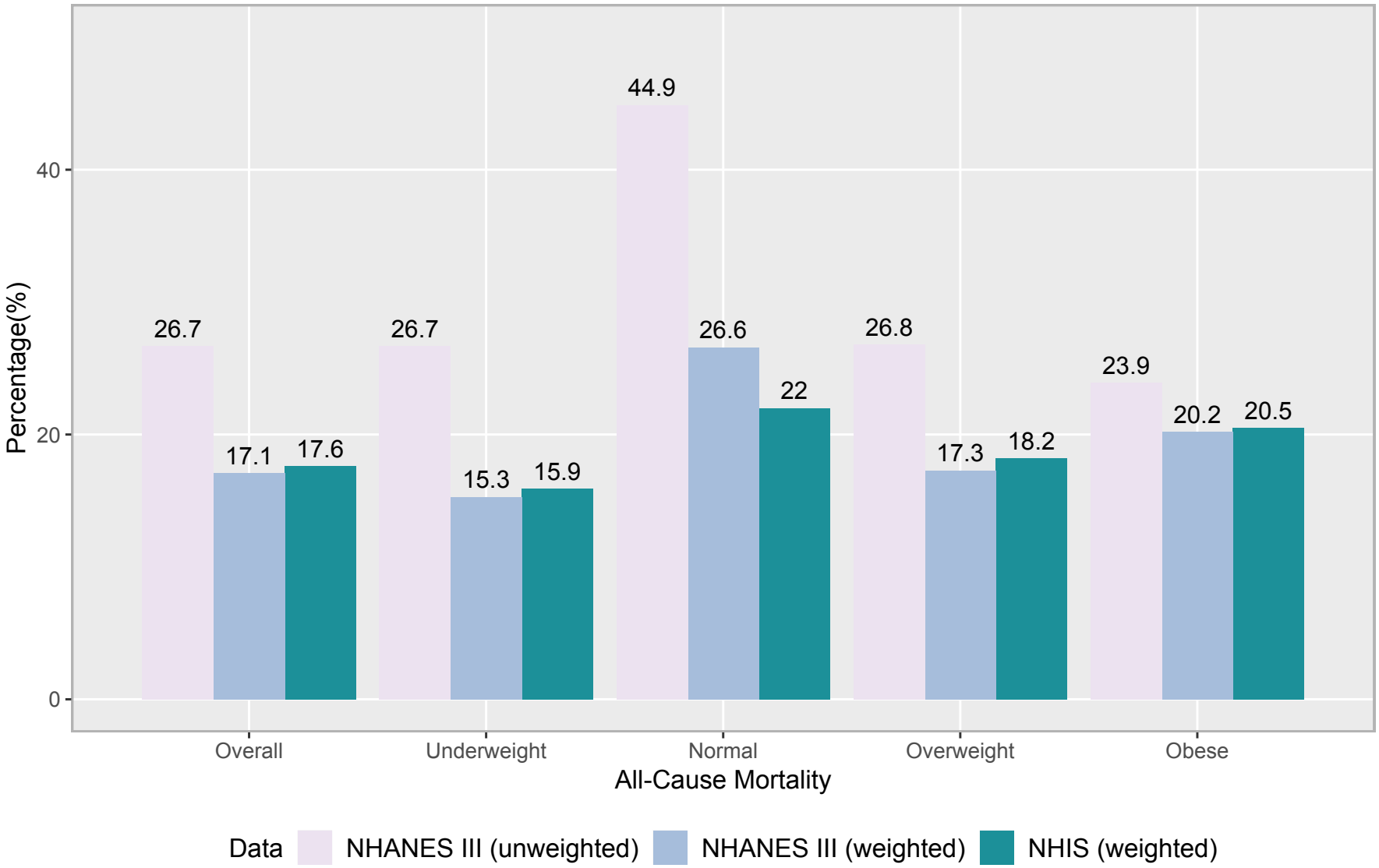
  2. 1994 US National Health Interview Survey (NHIS)

     A cross-sectional household interview survey of the civilian noninstitutionalized US population. ($n_s = 19,738$)
     $\widehat{N} = 189,608,549$. 124 strata, 2 PSU's per stratum.

  Note: Both datasets were linked to National Death Index (NDI) for mortality information.

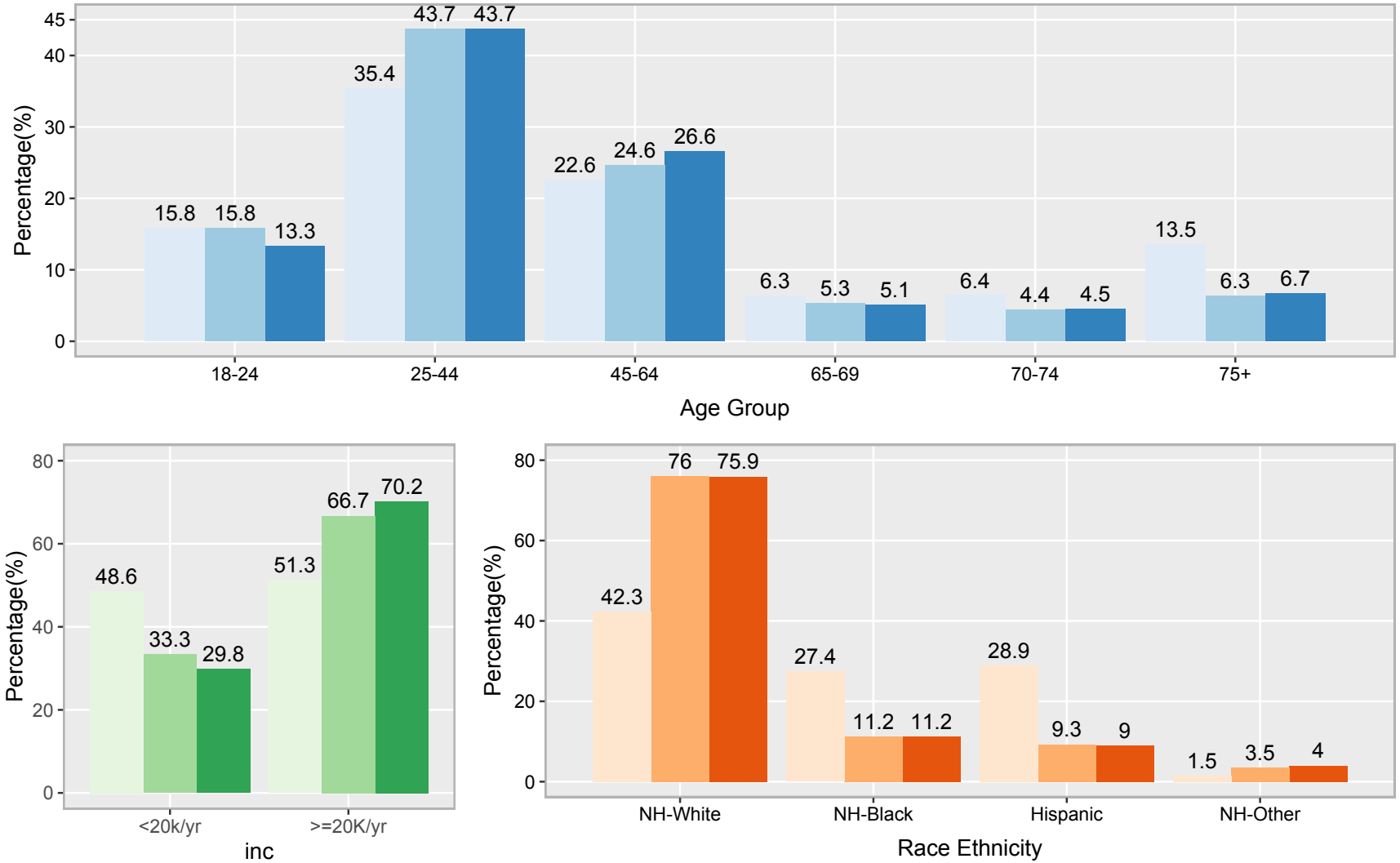
Real Data Example

15-Year Mortality (by BMI) in 1994 NHIS v.s. NHANES III



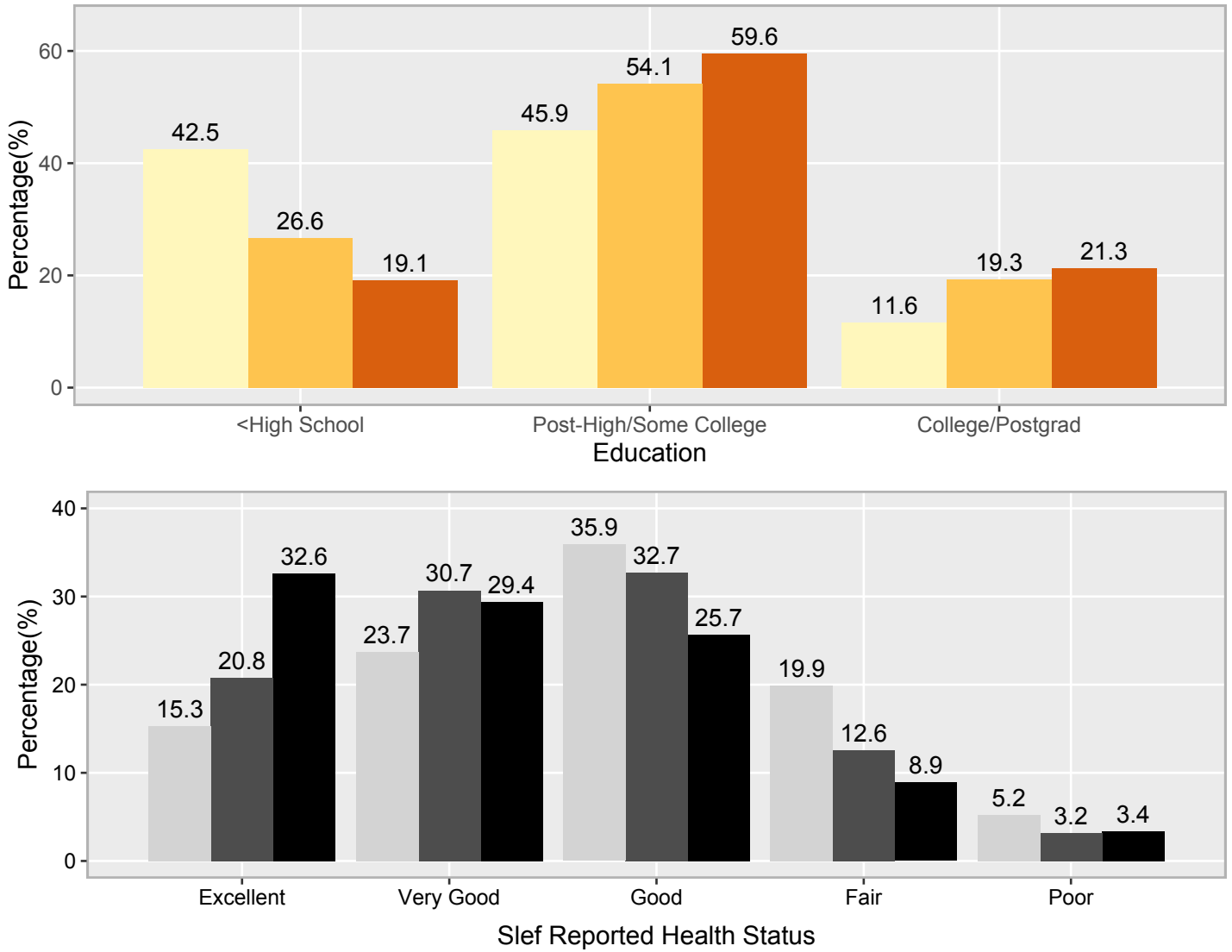
Real Data Example

Selected Variables in 1994 NHIS v.s. NHANES III



Real Data Example

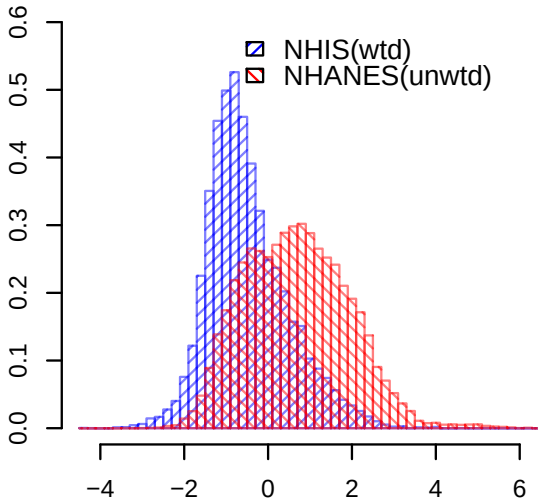
Variable Harmonization??

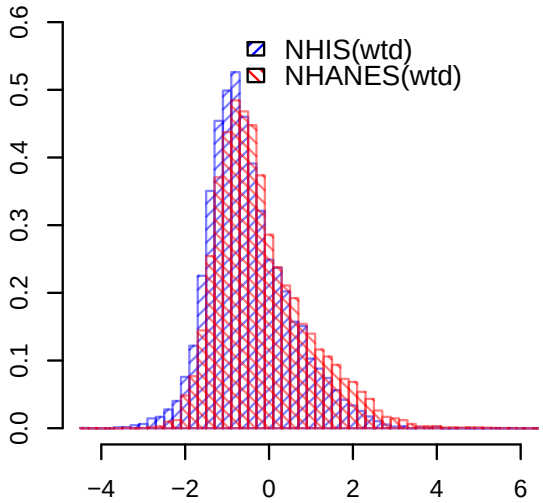


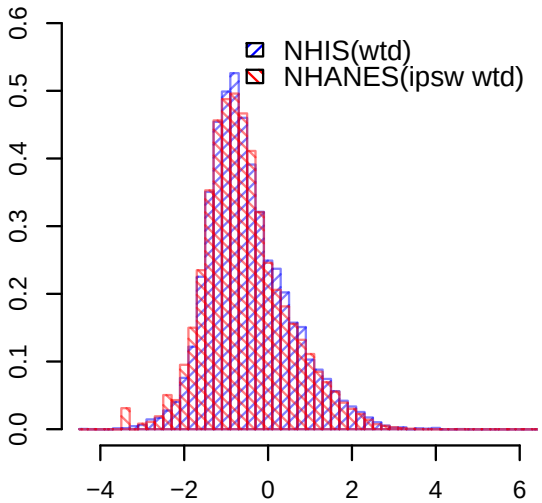
Real Data Example

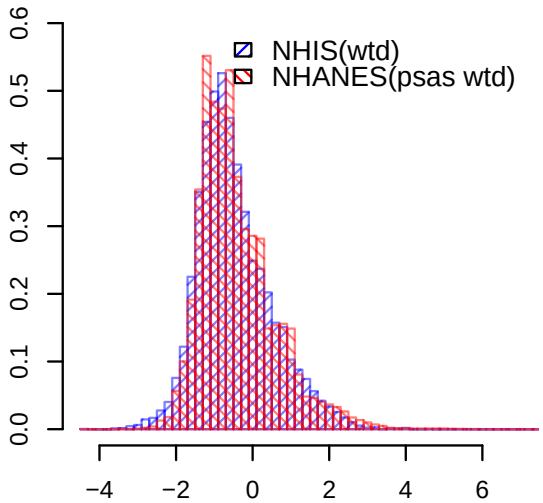
Predicted Propensity Scores on Logit Scale ($X\hat{\beta}$)

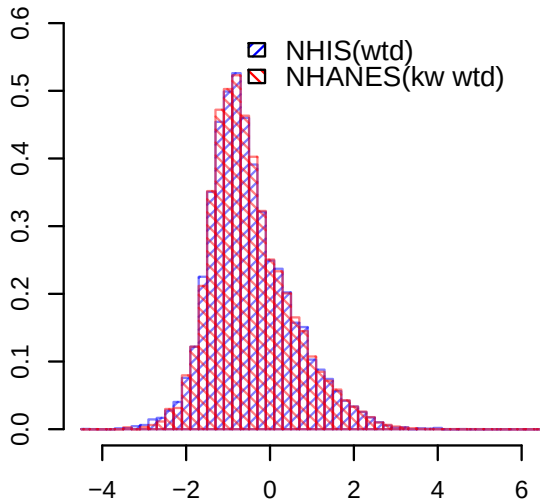
age, sex, race, region, educ, poverty, income, health, marital status, smoking, bmi







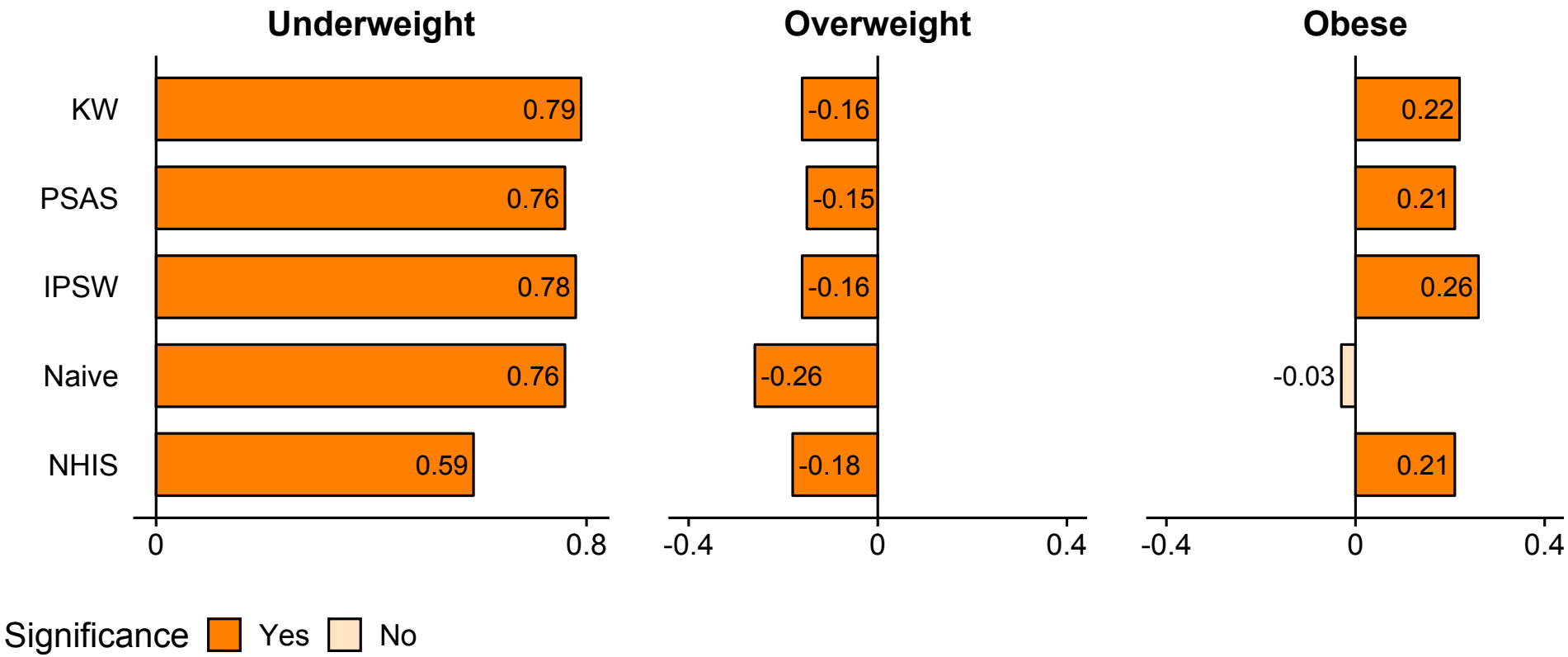




Real Data Example

Predicted Log Odds Ratios ($\hat{\beta}$)

15-year mortality $\sim$ age, sex, race, education, smoking status, bmi



Summary and Discussion

Summary

Naive sample estimates of Association can be biased

1. When the self-selection is informative
2. When the self-selection is non-informative but the outcome model is mis-specified

Type I error of hypothesis test can be changed

Weighting approaches

1. Reduce bias
2. Maintain Type I error

Summary and Discussion

Conclusion

Performance of KW, IPSW, and PSAS estimates

1. Unbiased KW and IPSW estimates under true propensity score model.
2. IPSW: more extreme weights, inflated variance.
3. PSAS: special case of KW, oversmoothed.
4. KW: smallest mean squared error.
5. TL underestimates variances. JK is recommended.

Summary and Discussion

Discussion

1. Analysis specific weights to further reduce bias?

   Weight optimization and smoothing
   Pfeffermann & Sverchkov (1999); Kim & Skinner (2013); Beaumont (2008)

   Reduce variance and not add bias
   - the weights are less informative
   - weights are correctly modeled
   - the outcome model is correctly specified

2. Machine learning methods for propensity score estimation

References

Pizzi, C., De Stavola, B., Merletti, F., Bellocco, R., dos Santos Silva, I., Pearce, N., and Richiardi, L. (2011).
Sample selection and validity of exposuredisease association estimates in cohort studies.
Journal of Epidemiology & Community Health, 65(5), 407-411.

Richiardi, L., Pizzi, C., and Pearce, N. (2013).
Commentary: Representativeness is usually not necessary and often should be avoided.
International journal of epidemiology, 42(4), 1018-1022.

Fuller, W. A. (2011).
Sampling statistics (Vol. 560).
John Wiley & Sons.

Korn, E. L., & Graubard, B. I. (1999).
Analysis of health surveys (Vol. 323).
John Wiley & Sons.

Johnson, C. L., Dohrmann, S. M., Burt, V. L.,& Mohadjer, L. K. (2014).
National health and nutrition examination survey: sample design, 2011-2014.
Vital and health statistics. Series 2, Data evaluation and methods research, (162), 1.

Sanderson, M.,& Gonzalez, J. F. (1998).
1988 National Maternal and Infant Health Survey: methods and response characteristics.
Vital and health statistics. Series 2, Data evaluation and methods research, (125), 1.

Questions?

Contact Information:

Lingxiao Wang

Ph.D. student

The Joint Program in Survey Methodology

University of Maryland, College Park, MD, U.S.A 20874

Email: wanglx89@umd.edu