A Kernel Weighting Approach to Improve Population Representativeness for Estimating Prevalence of Risk-factors and Diseases

Yan Li

Joint Program in Survey Methodology, University of Maryland, College Park, MD, U.S.A

NISS workshop
on Using Surveys to Improve the Representativeness of Nonprobability Samples in Epidemiologic Studies

March 11, 2019

Overview

Volunteer-Based Cohorts versus Probability-Based Sample

	Volunteer cohort	Probability sample
Advantages	Less expensive	
	Quick	Representativeness
	Convenient	Inference for population
	More detailed, specific info	
	Large sample sizes	
Disadvantages	? Representativeness	?Cost
	? Biased estimates	

Volunteer-Based Cohorts versus Probability-Based Sample

	Volunteer cohort	**Probability sample**
Advantages	Less expensive Quick Convenient More detailed, specific info Large sample sizes	Representativeness Inference for population
Disadvantages	? Representativeness ? Biased estimates	?Cost

Volunteer Cohort

- ✓ Health Volunteer Effects (Pinsky et al. 2007)
- ✓ For example (Fry et al. 2017):

All-cause mortality rate in UK Biobank = Half of UK population

Using a cohort study to estimate population prevalence requires addressing the representativeness of the cohort!

Propensity-Score-Based Methods

In randomized trial study

✓ Match and balance the distributions of confounders (Rosenbaum & Rubin, 1983) to estimate treatment effect

In probability samples

✓ Estimate propensity of responding (Czajka et al., 1992) to adjust nonresponse bias

Can use the propensity score to improve the representativeness of cohort sample?

Existing Propensity-Score-Based Weighting Methods

- **Inverse of Propensity Score Weighting (IPSW)**

  Elliott (2013); Elliott et al. 2016; Chen et al., 2018; Kim & Wang 2018; etc

  Estimate the propensity for popualtion unit r included in the cohort

  For example, Valliant & Dever (2011)

$$log\frac{p(\boldsymbol{x_r})}{1-p(\boldsymbol{x_r})} = \alpha + \gamma^T \boldsymbol{x}_r \text{ for } r \in s_s \cup s_c$$

  where s_s is a survey sample and s_c is a cohort.

  The corresponding pseudo-weight is:

$$w_j^{IPSW} = \frac{1-\hat{p}(\boldsymbol{x}_j, \hat{\boldsymbol{\gamma}}_w)}{\hat{p}(\boldsymbol{x}_j, \hat{\boldsymbol{\gamma}}_w)} \text{ for } j \in s_c$$

Existing Propensity-Score-Based Weighting Methods

- **Inverse of Propensity Score Weighting (IPSW)**

  Properties:

  - Correct bias under the true propensity score model ✓
  - Sensitive to Model misspecification ?
  - Extreme pseudo-weights ?

Existing Propensity-Score-Based Weighting Methods

- **Propensity Score Adjustment by Subclassification (PSAS)**

  Lee & Valliant 2009

  The estimated PS is used to measure the similarity of the X distributions

  - Sort the combined sample by the estimated PS
  - Partition the sorted sample into K subclasses
  - Divide the sum(survey weights) by # of cohort units in each subclass

$$w_j^{PSAS} \text{ for } j \in s_c$$

Existing Propensity-Score-Based Weighting Methods

- **Propensity Score Adjustment by Subclassification (PSAS)**

  Assume: All cohort units with subclasses represent the same # of population units

  Properties:
  - Variance ✓
  - Bias ?
  - Number of classes ?

Research Goal

Propose a new propensity-score-based weighting approach

- ✓ Variance reduction
- ✓ Bias reduction
- ✓ No ad-hoc subclassification
- ✓ Appropriate variance estimation for weighted estimates

Notation

Study Population: U with size N

Survey Sample (s_s)	Cohort (s_c)
H strata $\rightarrow a_h$ PSUs in stratum h $\rightarrow$ individuals	a study centers $\rightarrow$ individuals

Combined sample: $s_s \cup s_c$ ($H + 1$ strata)

Subject and Methods

Notation Cont'd

- y: variable of interest.
- x : $q \times 1$ vector of covariates available in both s_s and s_c.
- z: indicator for cohort membership (=1 for $r \in s_c$)
- $p(x) = Pr\{z = 1 \mid x\}$: propensity score.

- w_i: sample weight for $i \in s_s$.
- $\widehat{N} = \sum_{i \in s_s} w_i$: survey estimate of population total.

Subject and Methods

Notation Cont'd

- y: variable of interest.
- x : $q \times 1$ vector of covariates available in both s_s and s_c.
- z: indicator for cohort membership (=1 for $r \in s_c$)
- $p(x) = Pr\{z = 1 \mid x\}$: propensity score.

- w_i: sample weight for $i \in s_s$.
- $\widehat{N} = \sum_{i \in s_s} w_i$: survey estimate of population total.

4-Step Kernel Weighting Method to Create Pseudo-weights for a Cohort

1. Fit logistic regression model for predicting $p(x)$

$$log\frac{p(x_r)}{1-p(x_r)} = \alpha + \gamma^T x_r \text{ for } r \in s_s \cup s_c \tag{1}$$

   Get estimated propensity score $\hat{p}\left(x_i^{(s)}\right), \hat{p}\left(x_j^{(c)}\right)$ for $i \in s_s$, and $j \in s_c$ respectively.

2. For each individual $i \in s_s$, compute

$$d\left(x_i^{(s)}, x_j^{(c)}\right) = \hat{p}\left(x_i^{(s)}\right) - \hat{p}\left(x_j^{(c)}\right) \text{ for each } j \in s_c$$

4-Step Kernel Weighting Method to Create Pseudo-weights for a Cohort

1. Fit logistic regression model for predicting $p(x)$

$$log\frac{p(x_r)}{1-p(x_r)} = \alpha + \gamma^T x_r \text{ for } r \in s_s \cup s_c \tag{1}$$

   Get estimated propensity score $\hat{p}\left(x_i^{(s)}\right), \hat{p}\left(x_j^{(c)}\right)$ for $i \in s_s$, and $j \in s_c$ respectively.

2. For each individual $i \in s_s$, compute

$$d\left(x_i^{(s)}, x_j^{(c)}\right) = \hat{p}\left(x_i^{(s)}\right) - \hat{p}\left(x_j^{(c)}\right) \text{ for each } j \in s_c$$

Kernel Weighting Method to Create Pseudo-weights for a Cohort Cont'd

3. Obtain kernel weight (KW) for each $j \in s_c$ from the unit i

$$k_{ij} = \frac{K\left(d\left(x_i^{(s)}, x_j^{(c)}\right)/h\right)}{\sum_{j \in s_c} K\left(d\left(x_i^{(s)}, x_j^{(c)}\right)/h\right)} \text{ for } j \in s_c$$

h: bandwidth; $K(\cdot)$: kernel function.
Note: $\sum_{j \in s_c} k_{ij} = 1;\ k_{ij} \in [0, 1)$.

- The closer the distance;
- The higher similarity in **x** distribution;
- Larger portion of w_i assigned to cohort unit j

Therefore, relax the assumption of PSAS

Subject and Methods

3. Compute the KW pseudo-weight for $j \in s_c$

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

The sum of pseudo-weights across cohort units:

$$\sum_{j \in s_c} w_j^{kw} = \sum_{i \in s_s} w_i$$

The cohort KW estimate of prevalence is

$$\widehat{\overline{Y}}^{kw} = \left(\sum_{j \in s_c} w_j^{kw} \right)^{-1} \sum_{j \in s_c} w_j^{kw} \cdot y_j$$

Subject and Methods

3. Compute the KW pseudo-weight for $j \in s_c$

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

The sum of pseudo-weights across cohort units:

$$\sum_{j \in s_c} w_j^{kw} = \sum_{i \in s_s} w_i$$

The cohort KW estimate of prevalence is

$$\widehat{\overline{Y}}^{kw} = \left(\sum_{j \in s_c} w_j^{kw} \right)^{-1} \sum_{j \in s_c} w_j^{kw} \cdot y_j$$

Subject and Methods

3. Compute the KW pseudo-weight for $j \in s_c$

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

The sum of pseudo-weights across cohort units:

$$\sum_{j \in s_c} w_j^{kw} = \sum_{i \in s_s} w_i$$

The cohort KW estimate of prevalence is

$$\widehat{\overline{Y}}^{kw} = \left(\sum_{j \in s_c} w_j^{kw} \right)^{-1} \sum_{j \in s_c} w_j^{kw} \cdot y_j$$

Property of Kernel Pseudo-Weights

Theorem Under the following conditions:

- $\int K(u)du = 1$
- $\sup_u |K(u)| < \infty,\ \int |K(u)|du < \infty,\ \lim_{|u|\to\infty} |u| \cdot |K(u)| = 0,$
- $n_c \to \infty,\ h_{n_c} \to 0,\ n_c \cdot h_{n_c} \to \infty$
- $\boldsymbol{E}(Y|p(x), cohort) = \boldsymbol{E}(Y|p(x), survey)$
- $\boldsymbol{E}(Y) = \mu,\ \boldsymbol{E}(Y^2) < \infty$

KW estimator of population means is consistent with the target population mean

$$\left(\widehat{\overline{Y}}^{kw} - \overline{Y}\right) \xrightarrow{\boldsymbol{P}} 0$$

Subject and Methods

Jackknife Variance Estimation

- Total number of strata $= H + 1$.
- Number of replicates $K = \sum_h^{H+1} a_h$.

For replicate $(h\alpha)$,

1. Leave out α-th cluster in stratum h.
2. Calculate the weight adjustment factor $f_{r(h\alpha)}$

$$f_{r(h\alpha)} = \begin{cases} 0, & \text{stratum } h \text{ cluster } \alpha; \\ \frac{a_h}{a_h - 1}, & \text{stratum } h \text{ cluster } \alpha' \neq \alpha; \\ 1, & \text{otherwise.} \end{cases}$$

3. Refit model 1 with $f_{r(h\alpha)}$ in 2, and re-estimate propensity scores.

Subject and Methods

Jackknife Variance Estimation

- Total number of strata $= H + 1$.
- Number of replicates $K = \sum_h^{H+1} a_h$.

For replicate $(h\alpha)$,

1. Leave out α-th cluster in stratum h.
2. Calculate the weight adjustment factor $f_{r(h\alpha)}$

$$f_{r(h\alpha)} = \begin{cases} 0, & \text{stratum } h \text{ cluster } \alpha; \\ \frac{a_h}{a_h - 1}, & \text{stratum } h \text{ cluster } \alpha' \neq \alpha; \\ 1, & \text{otherwise.} \end{cases}$$

3. Refit model 1 with $f_{r(h\alpha)}$ in 2, and re-estimate propensity scores.

Subject and Methods

Jackknife Variance Estimation

- Total number of strata $= H + 1$.
- Number of replicates $K = \sum_h^{H+1} a_h$.

For replicate $(h\alpha)$,

1. Leave out α-th cluster in stratum h.
2. Calculate the weight adjustment factor $f_{r(h\alpha)}$

$$f_{r(h\alpha)} = \begin{cases} 0, & \text{stratum } h \text{ cluster } \alpha; \\ \frac{a_h}{a_h - 1}, & \text{stratum } h \text{ cluster } \alpha' \neq \alpha; \\ 1, & \text{otherwise.} \end{cases}$$

3. Refit model 1 with $f_{r(h\alpha)}$ in 2, and re-estimate propensity scores.

Jackknife Variance Estimation Cont'd

4. Calculate kernel weight for $j \in s_c^{(h\alpha)}$ associated with $i \in s_s^{(h\alpha)}$

$$k_{ij(h\alpha)} = \frac{K\left(d\left(x_i^{(s)}, x_j^{(c)}\right)/h\right)}{\sum_{j \in s_c^{(h\alpha)}} K\left(d\left(x_i^{(s)}, x_j^{(c)}\right)/h\right)}, \text{ for } j \in s_c^{(h\alpha)}.$$

5. The KW pseudo-weight for $j \in s_c^{(h\alpha)}$ is.

$$w_{j(h\alpha)}^{kw} = \sum_{i \in s_s^{(h\alpha)}} k_{ij(h\alpha)} \cdot w_i \cdot f_{i(h\alpha)}, \text{ for } j \in s_c^{(h\alpha)}.$$

Jackknife Variance Estimation Cont'd

4. Calculate kernel weight for$j \in s_c^{(h\alpha)}$ associated with $i \in s_s^{(h\alpha)}$

$$k_{ij(h\alpha)} = \frac{K\left(d\left(x_i^{(s)}, x_j^{(c)}\right)/h\right)}{\sum_{j \in s_c^{(h\alpha)}} K\left(d\left(x_i^{(s)}, x_j^{(c)}\right)/h\right)}, \text{ for } j \in s_c^{(h\alpha)}.$$

5. The KW pseudo-weight for $j \in s_c^{(h\alpha)}$ is.

$$w_{j(h\alpha)}^{kw} = \sum_{i \in s_s^{(h\alpha)}} k_{ij(h\alpha)} \cdot w_i \cdot f_{i(h\alpha)}, \text{ for } j \in s_c^{(h\alpha)}.$$

JK Variance for Mean/ Prevalence

For each replicate, re-estimate the population mean/ prevalence with replicate KW pseudo-weights

$$\widehat{\overline{Y}}^{kw}_{(h\alpha)} = \left(\sum_{j \in s_c^{(h\alpha)}} w^{kw}_{j(h\alpha)}\right)^{-1} \cdot \sum_{j \in s_c^{(h\alpha)}} w^{kw}_{j(h\alpha)} \cdot y_j$$

The JK variance estimate for $\widehat{\overline{Y}}^{kw}$ is

$$var_{JK}\left(\widehat{\overline{Y}}^{kw}\right) = \sum_{h=1}^{H+1} \frac{a_h - 1}{a_h} \sum_{\alpha=1}^{a_h} \left(\widehat{\overline{Y}}^{kw}_{(h\alpha)} - \widehat{\overline{Y}}^{kw}\right)^2.$$

Finite Population Generation

1. $M = 3,000$ clusters with size=3,000 (population size $N = 9 \times 10^6$)
2. Population Generation
   - age, sex, Hisp, income, and urban/rural (2015 ACS)
   - Continuous exposure Env
   - **Disease status** y (=1 for having disease; 0 otherwise)

$$logit\{Pr(y = 1)\} = \beta_0 + \beta_1 age + \beta_2 sex + \beta_3 Hisp + \beta_4 Env$$

Sample to Assemble the Survey Sample and Cohort

Two-stage Probability Proportional to Size (PPS) Design

Sample	Design	Measure of Size	Inclusion Probability
Cohort	Clusters	$\sum_{i \in C_a} s_i^b$	$\frac{n_c \cdot s_i^b}{\sum_{i=1}^N s_i^b}$
$(n_c = 75 \times 150)$	Individuals	s_i^b	
Survey	Clusters	$\sum_{i \in C_a} s_i^{b'}$	$\frac{n_s \cdot s_i^{b'}}{\sum_{i=1}^N s_i^{b'}}$
$(n_s = 150 \times 10)$	Individuals	$s_i^{b'}$	

C_a: a^{th} cluster $(a = 1, \cdots, M)$; b and b': real numbers

s_i : generated by $s = \exp\{\boldsymbol{\gamma x}\}$, where $\boldsymbol{x}$=(1, age, income, Env, v)
where $v = Pr(y = 1) + u,\ u \sim N(0, 0.01)$

As the result:

$$logit\{Pr(z = 1|s_c \cup s_s)\} = const. + (b - b') \cdot \boldsymbol{\gamma x}$$

$$logit\{Pr(z = 1|s_c \cup U)\} = const.^* + b \cdot \boldsymbol{\gamma x}$$

Results under 1+3 Propensity Score Models

- **Weighting methods**

  IPSW, PSAS, KW

- **Propensity score models**

Model	Covariates
True model	age, income, Env, z
Underfit model	age, income, Env
Mixed model	age, income, Env, race/ethnicity, sex
Overfit model	age, income, Env, z, urban/rural(age, income, Env, z)

- **Analytical Statistic**

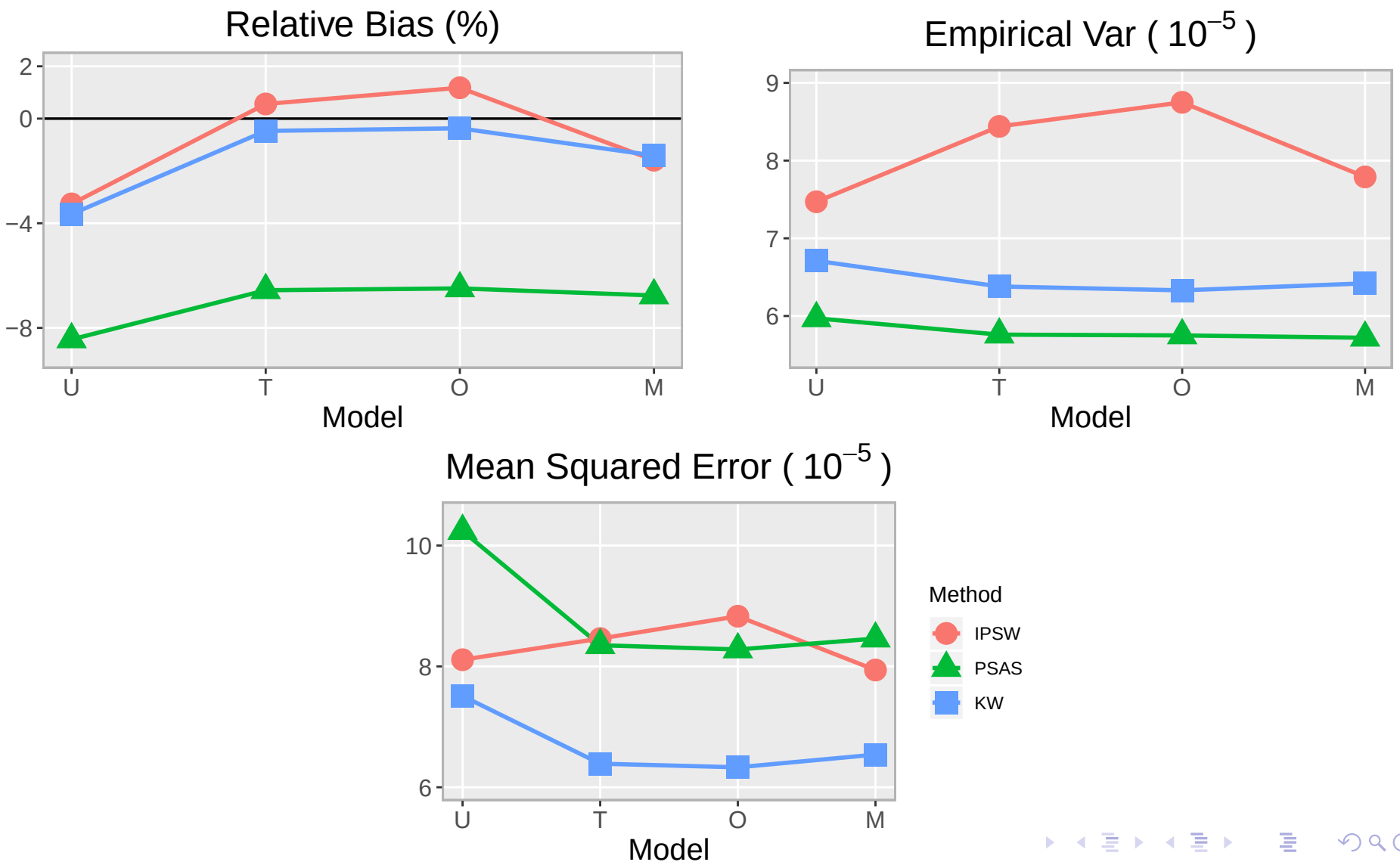
  Estimate of disease prevalence $\bar{y}$

- **Criteria**

  Relative bias, empirical variance, variance ratio $= \frac{\text{analytical variance (TSL, JK)}}{\text{empirical variance}}$, coverage probability

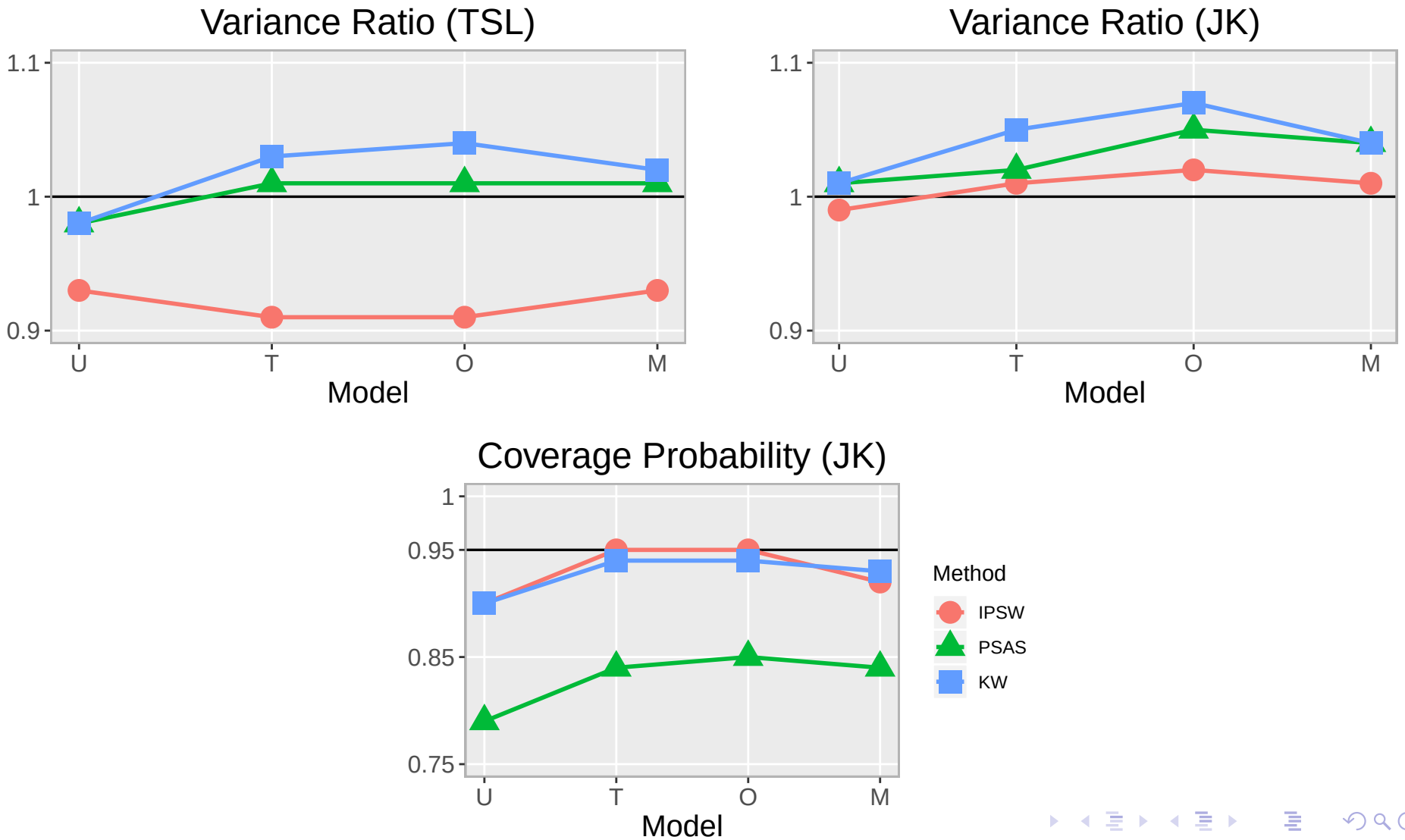
Simulation Studies

Relative Bias, Empirical Variance, and MSE of Prevalence Estimates



Simulation Studies

Variance Ratios and Coverage Probabilities of Prevalence Estimates



Data Analysis: The NIH-AARP Cohort Study

Data Materials

- **Aim**

  Estimate prevalence of multiple diseases, and prospective nine-year all-cause mortality for people aged 50 to 71 in the US from 1996.

- **Data**

  1. National Institutes of Health and the American Association of Retired Persons (NIH-AARP) Diet and Health Study

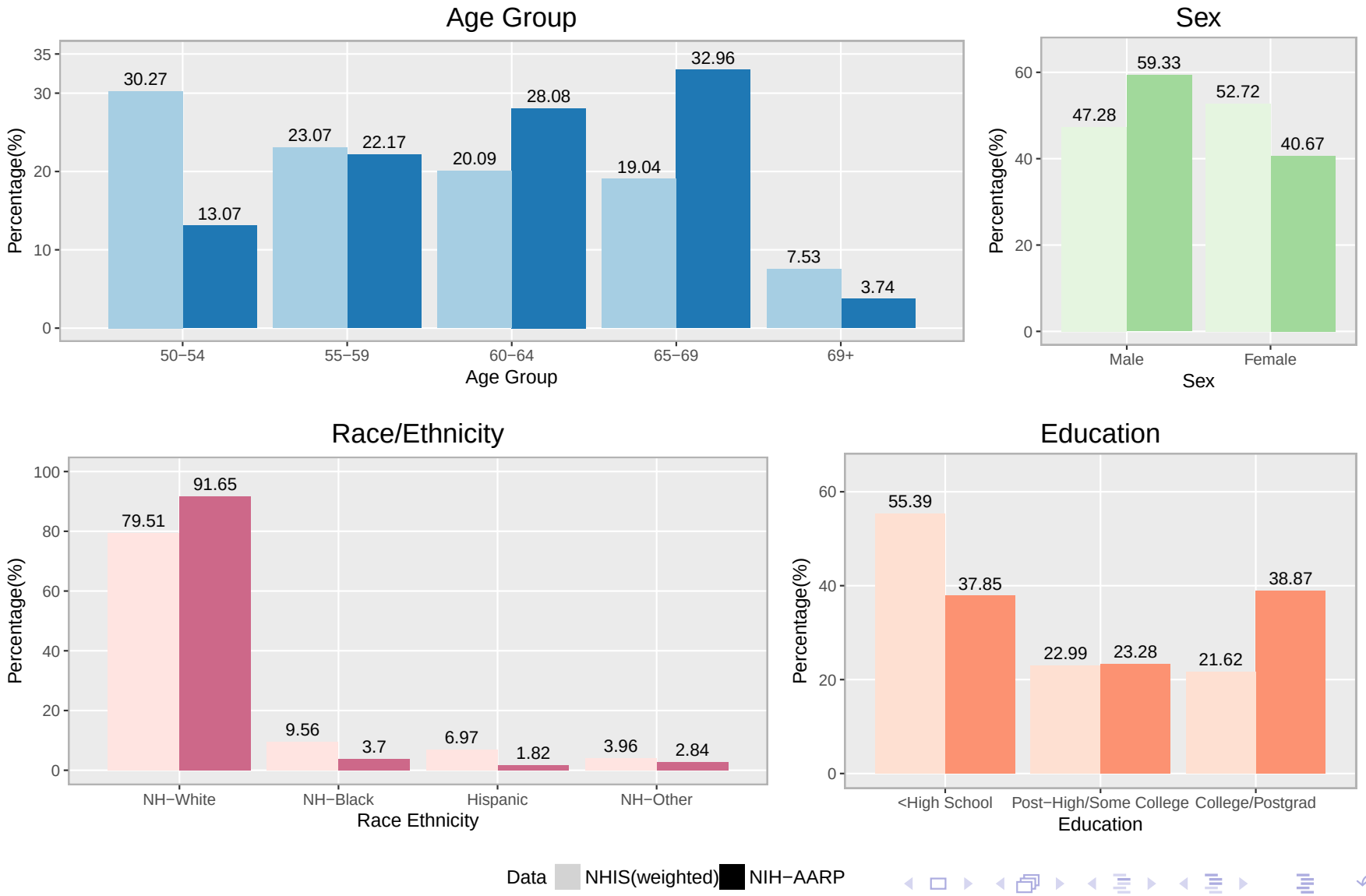
     AARP members from 1995-1996, aged 50 to 71 years, in six states or in two metropolitan areas. ($n_c = 529,708$)

  2. 1997 US National Health Interview Survey (NHIS)

     A cross-sectional household interview survey of the civilian noninstitutionalized US population. ($n_s = 9,306$)
     $\widehat{N} = 49,761,895$. 339 strata. 2 PSU's per stratum.

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

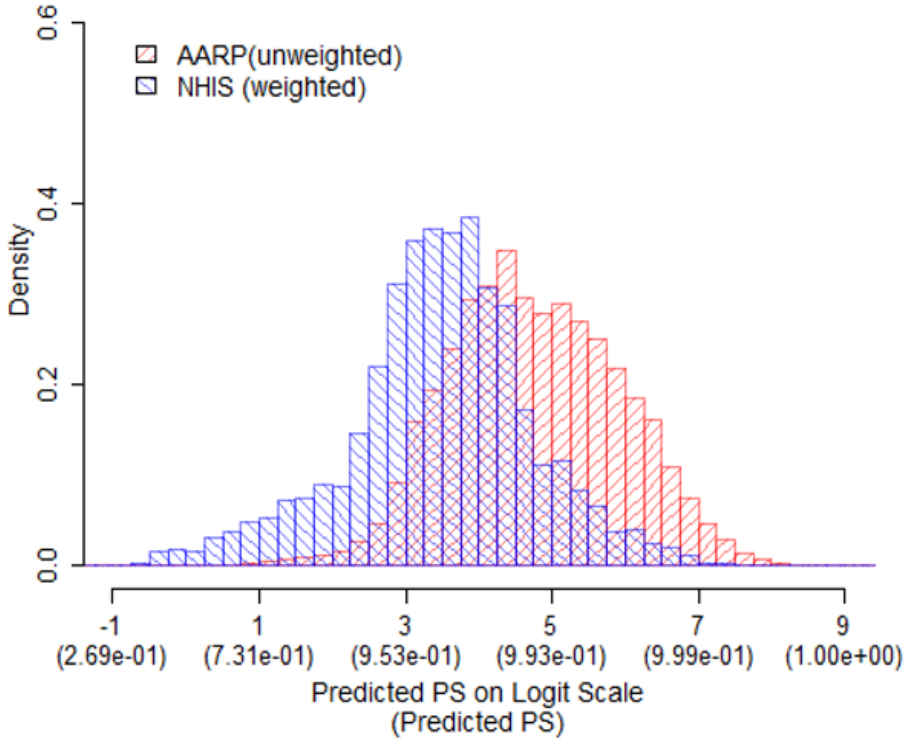
Selected Demographic Characteristics in 1997 NHIS v.s. NIH-AARP



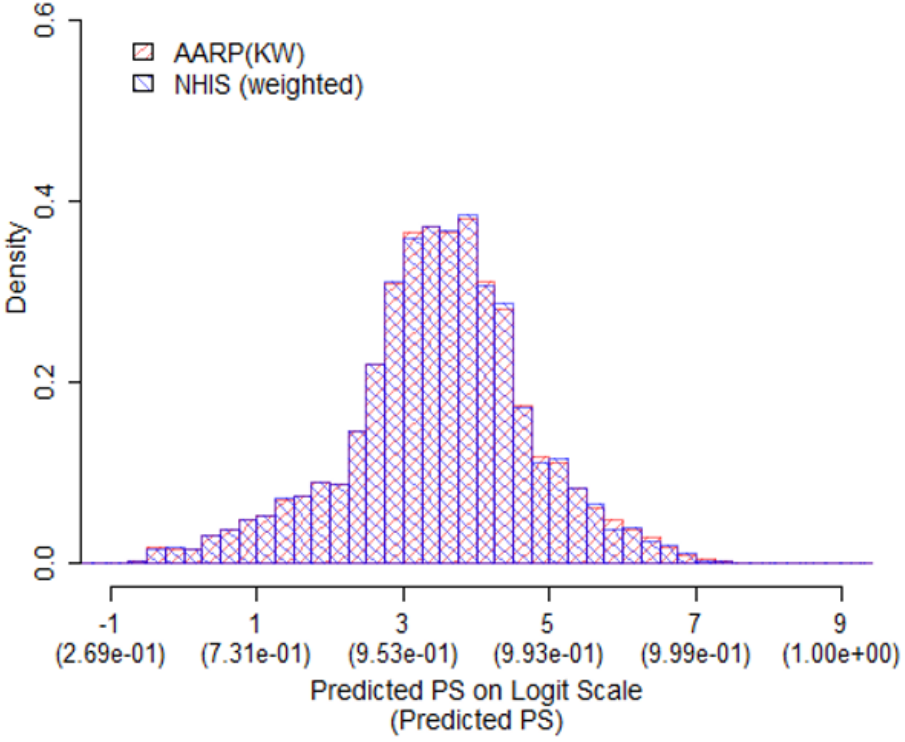
Data Analysis: The NIH-AARP Cohort Study

Histograms of Predicted Logit Propensity Scores

AARP v.s. NHIS



KW weighted AARP v.s. NHIS



Note: The propensity score model did not include NHIS sample weights.

Analysis: The NIH-AARP Cohort Study

Evaluation Criteria

- p_{NHIS}: Estimate of Disease Prevalence from NHIS
- p_{AARP}: Estimate of Disease Prevalence from naive AARP
- p^*: Estimate of Disease Prevalence from (IPSW, PSAS or KW)-weighted AARP

$$BiasReduction(\%) = \frac{p_{AARP} - p^*}{p_{AARP} - p_{NHIS}} \times 100$$

Bias Reduction(%) for NIH-ARRP Estimates

Discussion, Conclusion, and Limitations

Summary

Kernel-weighting approach for cohort:

1. Predict propensity scores
2. Compute kernel weights by kernel-smoothing the distances of predicted propensity scores between survey and cohort units.
3. Create pseudo-weights by the sum of the survey weights, weighted by the kernel weight.

Properties

1. Unbiased estimate of population size
2. Consistent estimate of population mean/prevalence under conditions

Variance Estimation

JK variance considers all sources of variability.

Conclusion

Performance of KW prevalence (v.s. IPSW, PSAS)

1. IPSW: Extreme weights and sensitive to model mis-specification
2. PSAS: Special case of KW, but oversmoothed.
3. Less bias, reduced Variance and best MSE.

   Note: reduce, but cannot eliminate bias in practice.

Discussion, Conclusion, and Limitations

Discussion

Kernel function

Bias reduction: $N(0, \sigma)$; Variance control: $Tri(-b, b, 0)$.

Bandwidth selection

Silverman's (Silverman, 1986) or Scott's (Scott, 1992) method.

Discussion, Conclusion, and Limitations

Limitations and Future Research

1. Requires overlapping distributions
2. Depends on the predictivity of propensity score model
3. Model selection and diagnostics
4. Doubly robust estimators
   (Kim & Wang 2018; Chen et al., 2018)
   - Design unbiased if propensity model is correct
   - Model unbiased if outcome model is correct

References I

Chen, Y., Li, P., and Wu, C. (2018)
Doubly Robust Inference with Non-probability Survey Samples

Czajka, J. L., Hirabayashi, S. M., Little, R. J., and Rubin, D. B. (1992).
Projecting from advance data using propensity modeling: An application to income and tax statistics.
Journal of Business & Economic Statistics, 10(2), 117-131.

Elliott, M.R., Valliant, R., Chen, J. K-T (2017)
Inference for Non-probability Samples.
Statistical Science, 32(2), 249-264.

Elliott, M. R. (2009)
Combining data from probability and non-probability samples using pseudo-weights.
Survey Practice, 2(6), 1-7.

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., et al. (2017).
Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population.
American journal of epidemiology, 186(9), 1026-1034.

Kim, J., K., & Wang, Z. (2018)
Sampling techniques for big data analysis in finite population inference.

Lee, S. and Valliant, R. (2009).
Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment.
Sociological Methods & Research, 37:319-343.

Pinsky, P. F., Miller, A., Kramer, B. S., Church, T., Reding, D., et al. (2007).
Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial.
American journal of epidemiology 165(8), 874-881.

Rosenbaum, P. R., & Rubin, D. B. (1983).
The central role of the propensity score in observational studies for causal effects.
Biometrika, 70(1), 41-55.

Valliant, R., & Dever, J. A. (2011).
Estimating propensity adjustments for volunteer web surveys.
Sociological Methods & Research 40(1), 105-137.

Acknowledgement

Collaborators:

Barry I. Graubard	**Lingxiao Wang**	**Hormuzd I. Katki**
Senior Investigator	PhD candidate and Predoctoral fellow	Senior Investigator
Biostatistics Branch DCEG	JPSM, University of Maryland	Biostatistics Branch DCEG,
National Cancer Institute	BB, DCEG, NCI	National Cancer Institute

This work motivated by the pioneer work

Michael Elloitt and Richard Valliant 2017 in Statistical Science

Thank You!

Contact Information:

Yan Li

Associate Professor

Joint Program in Survey Methodology

University of Maryland, College Park, MD, U.S.A 20872

Email: yli6@umd.edu