Improving External Validity of Epidemiologic Analyses: A Kernel Weighting Approach

Lingxiao Wang

Joint Program in Survey Methodology, University of Maryland, College Park, MD

wanglx89@umd.edu

November, 14, 2018
Overview

1. Introduction

2. Methods

3. Simulation Studies

4. Data Analysis: The NIH-AARP Cohort Study

5. Conclusion, Discussion, Limitations, and Future Work
Introduction

Non-Probability
Epidemiologic Cohorts

Population

Catchment Areas

Coverage

Invitation

Invited Sample

Volunteer

Cohort

Population-Based
Household Surveys

Form Strata in Population

Stage 1 – Clusters

Stage 2 – Households

Stage 3 – Individuals

Wang, L. (JPSM, UMD)

KW for External Validity

November, 14, 2018
## Introduction

### Nonprobability Samples v.s. Probability-Based Samples

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Nonprobability samples</th>
<th>Probability-based samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Less expensive</td>
<td>Representativeness</td>
</tr>
<tr>
<td></td>
<td>Quicker</td>
<td>Population Inference</td>
</tr>
<tr>
<td></td>
<td>Easier to implement</td>
<td>(External Validity)</td>
</tr>
<tr>
<td></td>
<td>Large sample sizes</td>
<td></td>
</tr>
</tbody>
</table>

| Disadvantages       | ?                       | Representativeness        |
|                     |                         | High costs                |
Low Representativeness of Non-probability Cohorts

- “Healthy volunteer effect” (Pinsky et al., 2017)
  Disease prevalences or mortality rates < national survey estimates or population values

- Example: UK Biobank (Fry et al., 2017)
  All-cause mortality rate in UK Biobank = Half of UK population

Using a Non-probability cohort to estimate population means/ prevalences requires addressing the representativeness of the cohort
Propensity-Score-Based Adjustments

- In observational studies
  Match and balance the distributions of confounders to estimate treatment effect (Rosenbaum & Rubin, 1983)

- In survey research
  Estimate propensity of responding to adjust nonresponse bias (Czajka et al., 1992)

⭐ Improve the representativeness of non-probability samples?
Introduction

Propensity-Score Weighting Methods in Survey Research

- **Inverse of propensity score weighting (IPSW)**
    - Correct bias under the true propensity score model ✓
    - How sensitive to model specification ?
    - Can produce highly variable estimates due to extreme weights ?

- **Propensity Score Adjustment by subclassification (PSAS)**
  - Lee & Valliant (2009), Valliant & Dever (2011)
    - Avoid extreme weights ✓
    - How much Bias reduced ?
    - Limited guidance on number of classes ?
    - Assumption of approximately equal propensity scores ?

- **Variance estimation ?**
  - Intra-cluster correlation
  - Predicting propensity scores
Introduction

Propensity-Score Weighting Methods in Survey Research

- **Inverse of propensity score weighting (IPSW)**
    - Correct bias under the true propensity score model ✓
    - How sensitive to model specification ?
    - Can produce highly variable estimates due to extreme weights ?

- **Propensity Score Adjustment by subclassification (PSAS)**
  - Lee & Valliant (2009), Valliant & Dever (2011)
    - Avoid extreme weights ✓
    - How much Bias reduced ?
    - Limited guidance on number of classes ?
    - Assumption of approximately equal propensity scores ?

- **Variance estimation ?**
  - Intra-cluster correlation
  - Predicting propensity scores
**Introduction**

### Propensity-Score Weighting Methods in Survey Research

- **Inverse of propensity score weighting (IPSW)**
  - Correct bias under the true propensity score model ✓
  - How sensitive to model specification ?
  - Can produce highly variable estimates due to extreme weights ?

- **Propensity Score Adjustment by subclassification (PSAS)**
  - Lee & Valliant (2009), Valliant & Dever (2011)
  - Avoid extreme weights ✓
  - How much Bias reduced ?
  - Limited guidance on number of classes ?
  - Assumption of approximately equal propensity scores ?

- **Variance estimation ?**
  - Intra-cluster correlation
  - Predicting propensity scores
Research Goal

Propose a propensity-score-based kernel weighting approach for mean / prevalence

- Bias reduction
- Efficiency improvement
- No ad-hoc subclassification
- Appropriate variance estimation for weighted estimates
**Methods**

**Basic Idea**

- **Cohort Sample** ($s_c$)
  - Population
  - Coverage
  - Catchment Areas
  - Invitation
  - Invited Sample
  - Volunteer
  - Cohort
  - $c$ study centers $\rightarrow$ individuals

- **Survey Sample** ($s_s$)
  - $H$ strata $\rightarrow a_h$ PSUs in stratum $h$
  - Sample weights $\{w_i\}_{i \in s_s}$

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

We distribute survey samples weights $\{w_i\}_{i \in s_s}$ to the cohort units based on their similarity measured by propensity scores.

Wang, L. (JPSM, UMD)

KW for External Validity

November, 14, 2018
Methods

Kernel Weighting Method

1. Fit logistic regression as PSAS method

\[
\log \left\{ \frac{p(x_r)}{1 - p(x_r)} \right\} = \gamma_0 + \gamma^T f(x_r), \quad r \in s
\]

Note: complex survey designs are not considered in propensity model (1)

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

2. For individual \( i \in s_s \), compute

\[
d \left( x_{i}^{(s)}, x_{j}^{(c)} \right) = \hat{p}(x_{i}^{(s)}) - \hat{p}(x_{j}^{(c)}) \quad \text{for each} \quad j \in s_c
\]
Methods

Kernel Weighting Method

1. Fit logistic regression as PSAS method

\[
\log \left\{ \frac{p(x_r)}{1 - p(x_r)} \right\} = \gamma_0 + \gamma^T f(x_r), \quad r \in s
\]  

(1)

**Note:** complex survey designs are not considered in propensity model (1)

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

2. For individual \( i \in s_s \), compute

\[
d \left( x_i^{(s)}, x_j^{(c)} \right) = \hat{p}(x_i^{(s)}) - \hat{p}(x_j^{(c)}) \quad \text{for each } j \in s_c\]
Kernel Weighting Method – Cont’d

Obtain kernel weight for $j \in s_c$ with $i \in s_s$

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

for $j \in s_c$

$K(\cdot)$: kernel function; $h$: bandwidth.

$k_{ij} \cdot w_i$ is distributed from $i$ to $j$.

NOTE

- $\sum_{j \in s_c} k_{ij} = 1; k_{ij} \in [0, 1)$.
- The larger $k_{ij}$ of unit $j$, the more similar to unit $i$, the larger portion of $w_i$ assigned to unit $j$.
- Relaxing the PSAS assumption of equal p-score within subclasses.
Kernel Weighting Method – Cont’d

4. Compute the KW pseudo-weight for \( j \in s_c \)

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

KW Estimate of Population Prevalence

The population prevalence is estimated by

\[
\hat{Y}^{KW} = \left( \sum_{j \in s_c} w_j^{KW} \right)^{-1} \sum_{j \in s_c} w_j^{KW} \cdot y_j
\]

\[
\sum_{j \in s_c} w_j^{KW} = \sum_{i \in s_s} w_i
\]
Kernel Weighting Method – Cont’d

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

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

KW Estimate of Population Prevalence

The population prevalence is estimated by

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

$$\sum_{j \in s_c} w_j^{KW} = \sum_{i \in s_s} w_i$$
Property of KW weighted Mean/Prevalence

Theorem Under the following conditions:

1. \( \int K(u)du = 1 \)
2. \( \sup_u |K(u)| < \infty, \int |K(u)|du < \infty, \lim_{|u| \to \infty} |u| \cdot |K(u)| = 0 \)
3. \( n_c \to \infty, \ h_{nc} \to 0, \ n_c \cdot h_{nc} \to \infty \)
4. \( E(Y|p(x), \text{cohort}) = E(Y|p(x), \text{survey}) \)
5. \( E(Y^2) < \infty, \ E(w_i) < \infty \)

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

\[
\left( \hat{Y}_{KW} - \bar{Y} \right) \xrightarrow{P} 0; \quad E \left( \hat{Y}_{KW} - \bar{Y} \right) \to 0
\]
### Three Propensity-Score-Based Weighting Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Samples</th>
<th>Propensity score</th>
<th>Kernel function</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPSW</td>
<td>$s_c \cup U$</td>
<td>Estimating participation rates</td>
<td>—</td>
</tr>
<tr>
<td>PSAS</td>
<td>$s_c \cup s_s$</td>
<td>Measure of similarity</td>
<td>Stratified uniform</td>
</tr>
<tr>
<td>KW</td>
<td></td>
<td></td>
<td>Gaussian, triangular etc.</td>
</tr>
</tbody>
</table>
Methods

Jackknife Variance Estimation

For replicate \((h \alpha)\),

1. Leave out \(\alpha\)-th cluster in stratum \(h\).
2. calculate the weight adjustment factor \(f(h \alpha)\) to weight up the remaining clusters in stratum \(h\)
3. Refit propensity model (1) with \(f(h \alpha)\) in 2, and re-estimate propensity scores.
4. Calculate KW pseudo weights

Considers sources of variability due to

- Cluster effect & unequal weights
- Estimating propensity scores

Also applies to IPSW & PSAS
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 \( Env \)
   - Disease status \( y (=1 \text{ for having disease}) \)
     \[ y \sim \text{Bernoulli}(\mu) \quad (\text{Lunn & Davis, 1998}) \]
     \[ \mu = \{1 + \exp(-\beta_0 - \beta_1 \text{age} - \beta_2 \text{sex} - \beta_3 \text{Hisp} - \beta_4 \text{Env})\}^{-1} \]
   - Continuous variable \( z = \mu + u, \ u \sim N(0, 0.01) \)

Note: All variables have intra-cluster correlation.
Simulation Studies – Prevalence Estimation

Sample to Assemble the Survey Sample and Cohort

Two-stage Probability Proportional to Size (PPS) design

<table>
<thead>
<tr>
<th>Sample</th>
<th>Design</th>
<th>Measure of Size</th>
<th>Inclusion Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort</td>
<td>Clusters</td>
<td>$\sum_{i \in C_{\alpha}} q_i^b$</td>
<td>$\frac{n_c \cdot q_i^b}{\sum_{i=1}^N q_i^b}$</td>
</tr>
<tr>
<td></td>
<td>Individuals</td>
<td>$q_i^b$</td>
<td>$\sum_{i=1}^N q_i^b$</td>
</tr>
<tr>
<td>($n_c = 6000$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Survey</td>
<td>Clusters</td>
<td>$\sum_{i \in C_{\alpha}} s_i^b$</td>
<td>$\frac{n_s \cdot s_i^b}{\sum_{i=1}^N s_i^b}$</td>
</tr>
<tr>
<td></td>
<td>Individuals</td>
<td>$s_i^b$</td>
<td>$\sum_{i=1}^N s_i^b$</td>
</tr>
<tr>
<td>($n_s = 1500$)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$C_{\alpha}$: $\alpha^{th}$ cluster ($\alpha = 1, \cdots, M$)

$q_i = \{1 + \exp(-\gamma_0 - \gamma_1 age_i - \gamma_2 hh\_inc_i - \gamma_3 Env_i - \gamma_4 z_i)\}^{-1}, s_i = 1 - q_i$

Propensity ($p$) model is:

$$\logit\{p\} = const. + b\gamma_1 age + b\gamma_2 hh\_inc + b\gamma_3 Env + b\gamma_4 z$$
Results under 1+3 Propensity Score Models

- **Analytical Statistic**
  Estimate of disease prevalence $\bar{y}$

- **Weighting methods**
  IPSW, PSAS, KW

- **Propensity score models**

<table>
<thead>
<tr>
<th>Model</th>
<th>Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>True model</td>
<td>age, income, $Env$, $z$</td>
</tr>
<tr>
<td>Underfit model</td>
<td>age, income, $Env$</td>
</tr>
<tr>
<td>Mixed model</td>
<td>age, income, $Env$, race/ethnicity, sex</td>
</tr>
<tr>
<td>Overfit model</td>
<td>age, income, $Env$, $z$, urban/rural</td>
</tr>
</tbody>
</table>

- **Criteria**
  Relative bias, empirical variance, variance ratio (TSL, JK), 95% coverage probability
Simulation Studies – Prevalence Estimation

Relative Bias and Empirical Variance

Relative Bias (%)

Empirical Var (10^{-5})

Method IPSW PSAS KW

Wang, L. (JPSM, UMD)
Simulation Studies – Prevalence Estimation

Estimated Variance and Coverage Probabilities

Variance Ratio (TSL)

Variance Ratio (JK)

Coverage Probability (TSL)

Coverage Probability (JK)

Method
- IPSW
- PSAS
- KW

Wang, L. (JPSM, UMD)
KW for External Validity
November, 14, 2018
Data Analysis: The NIH-AARP Cohort Study

Data Materials

- **Aim**
  Estimate 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. \( n_s = 9,306 \)
     \( \hat{N} = 49,761,895 \). 339 strata. 2 PSU’s per stratum.

**Note**: Both datasets were linked to National Death Index (NDI) for mortality information.
Nine-Year All-Cause Mortality in 1997 NHIS v.s. NIH-AARP

<table>
<thead>
<tr>
<th>Age Group</th>
<th>NHIS(weighted)</th>
<th>NIH-AARP</th>
</tr>
</thead>
<tbody>
<tr>
<td>50−54</td>
<td>13.67</td>
<td>14.22</td>
</tr>
<tr>
<td>55−59</td>
<td>6.27</td>
<td>5.62</td>
</tr>
<tr>
<td>60−64</td>
<td>9.71</td>
<td>8.26</td>
</tr>
<tr>
<td>64+</td>
<td>24.09</td>
<td>20.68</td>
</tr>
</tbody>
</table>

Overall:

- NHIS(weighted): 14.22%
- NIH-AARP: 13.01%
Selected Demographic Characteristics in 1997 NHIS v.s. NIH-AARP

Age Group

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Percentage (%) NHIS(weighted)</th>
<th>Percentage (%) NIH−AARP</th>
</tr>
</thead>
<tbody>
<tr>
<td>50−54</td>
<td>30.27</td>
<td>13.07</td>
</tr>
<tr>
<td>55−59</td>
<td>23.07</td>
<td>13.07</td>
</tr>
<tr>
<td>60−64</td>
<td>22.17</td>
<td>22.17</td>
</tr>
<tr>
<td>65−69</td>
<td>28.08</td>
<td>32.96</td>
</tr>
<tr>
<td>69+</td>
<td>32.96</td>
<td>3.74</td>
</tr>
</tbody>
</table>

Sex

<table>
<thead>
<tr>
<th>Sex</th>
<th>Percentage (%) NHIS(weighted)</th>
<th>Percentage (%) NIH−AARP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>59.33</td>
<td>47.28</td>
</tr>
<tr>
<td>Female</td>
<td>40.67</td>
<td>52.72</td>
</tr>
</tbody>
</table>

Race/Ethnicity

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>Percentage (%) NHIS(weighted)</th>
<th>Percentage (%) NIH−AARP</th>
</tr>
</thead>
<tbody>
<tr>
<td>NH−White</td>
<td>91.65</td>
<td>91.65</td>
</tr>
<tr>
<td>NH−Black</td>
<td>79.51</td>
<td>79.51</td>
</tr>
<tr>
<td>Hispanic</td>
<td>6.97</td>
<td>6.97</td>
</tr>
<tr>
<td>NH−Other</td>
<td>3.7</td>
<td>3.7</td>
</tr>
</tbody>
</table>

Education

<table>
<thead>
<tr>
<th>Education</th>
<th>Percentage (%) NHIS(weighted)</th>
<th>Percentage (%) NIH−AARP</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;High School</td>
<td>55.39</td>
<td>37.85</td>
</tr>
<tr>
<td>Post−High/Some College</td>
<td>22.99</td>
<td>23.28</td>
</tr>
<tr>
<td>College/Postgrad</td>
<td>21.62</td>
<td>38.87</td>
</tr>
</tbody>
</table>
Distribution of Self-Reported Health Status in 1997 NHIS v.s. NIH-AARP

<table>
<thead>
<tr>
<th>Health Status</th>
<th>NHIS (weighted)</th>
<th>NIH-AARP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>22.04%</td>
<td>16.51%</td>
</tr>
<tr>
<td>Very Good</td>
<td>16.51%</td>
<td>30.06%</td>
</tr>
<tr>
<td>Good</td>
<td>36.08%</td>
<td>29.65%</td>
</tr>
<tr>
<td>Fair</td>
<td>13.27%</td>
<td>11.09%</td>
</tr>
<tr>
<td>Poor</td>
<td>2.71%</td>
<td>1.85%</td>
</tr>
</tbody>
</table>

Wang, L. (JPSM, UMD)

KW for External Validity
November, 14, 2018
Data Analysis: The NIH-AARP Cohort Study

Histograms of Predicted Propensity Scores in 1997 NHIS v.s. NIH-AARP

AARP v.s. NHIS

KW weighted AARP v.s. NHIS

Note: The propensity score model did not include NHIS sample weights.
Histograms of Predicted Propensity Scores

AARP v.s. NHIS

PSAS weighted AARP v.s. NHIS

Note: The propensity score model did not include NHIS sample weights.
Histograms of Predicted Propensity Scores

AARP v.s. NHIS

IPSW weighted AARP v.s. NHIS

Note: The propensity score model included NHIS sample weights.
Evaluation Criteria

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

- \( p_{NHIS} \): Estimate of mortality rate from NHIS
- \( p_{AARP} \): Naive estimate of mortality rate from AARP
- \( p^* \): Estimate of mortality rate from (IPSW, PSAS or KW) -weighted AARP
Bias Reduction(%) of Weighted NIH-ARRP Estimates

Wang, L. (JPSM, UMD)  KW for External Validity  November, 14, 2018
Kernel-weighting approach for cohort in 3 steps:

1. Predict propensity to be in the cohort (versus in the survey)
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 weights.

Properties

1. Unbiased estimate of population total
2. Consistent estimates of population means/ prevalences

Variance Estimation

JK variance considers all sources of variability. TSL variance is fine.
Conclusion

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

1. Less bias and the best coverage probability.
3. IPSW: more extreme weights, inflated variance, inappropriate $\text{var}_{TSL}$.
4. PSAS: special case of KW, oversmoothed.
5. Reduce more, but cannot eliminate bias in practice.

KW estimate of regression coefficient (v.s. IPSW, PSAS)

1. Unbiased estimates and smallest MSE
Discussion and Limitations

Discussion

Kernel function
Bias reduction: \( N(0, \sigma^2) \); variance control: \( Tri(-b, b, 0) \).

Bandwidth selection
Silverman’s (Silverman, 1986) or Scott’s (Scott, 1992) method.

Limitations

1. Requires overlapping distributions
2. Depends on the predictivity of propensity score model
Future Work – Simulation Studies for Regression Analyses

Finite Population Generation

- **Disease:** \( y \sim Bernoulli(\mu), \quad \mu = \{1 + \exp(-\beta_0 - \beta_1 age - \beta_2 Env)\}^{-1} \)
- Continuous variable \( z = y - u, u \sim N(0, 0.01) \)

Assemble the Survey Sample and Cohort Informative Design

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** \( \text{logit} (p) = \text{const} + b\gamma_1 Env + b\gamma_2 z_i \cdot Env \)

**Analysis model** \( \text{logit} (P\{y = 1\}) = \beta_0 + \beta_1 age + \beta_2 Env \)

Ignoring the weights will introduce bias to estimate of \( \beta_2 \)!
Inference for $\beta_2$ under the True Propensity Score Models

<table>
<thead>
<tr>
<th>Method</th>
<th>Rel Bias(%)</th>
<th>Emp Var</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>-65.19</td>
<td>7.9E-03</td>
<td>9.6E-01</td>
</tr>
<tr>
<td>IPSW</td>
<td>-0.27</td>
<td>1.2E-02</td>
<td>1.2E-02</td>
</tr>
<tr>
<td>PSAS</td>
<td>-38.53</td>
<td>8.7E-03</td>
<td>3.4E-01</td>
</tr>
<tr>
<td>KW</td>
<td>-0.47</td>
<td>4.6E-03</td>
<td>4.6E-03</td>
</tr>
</tbody>
</table>
Finite Population Generation

- **Disease:**
  
  \[ y \sim \text{Bernoulli}(\mu), \quad \mu = \left\{ 1 + \exp(-\beta_0 - \beta_1 \text{age} - \beta_2 \text{Env} - \beta_3 \text{age} \cdot \text{Env}) \right\}^{-1} \]

- Continuous variable \( z^* = py - u, u \sim N(0, 0.01) \)

Assemble the Survey Sample and Cohort Non-informative Design

Two-stage PPS design with measures of size calculated from

- \( q_i = \exp\{\gamma_0 + \gamma_1 \text{Env}_i + \gamma_2 z^*_i \cdot \text{Env}_i\} \) for cohort selection

- \( s_i = \exp\{\frac{1}{2}(-\gamma_0 - \gamma_1 \text{Env}_i - \gamma_2 z^*_i \cdot \text{Env}_i)\} \) for survey sample selection

True propensity score model

\[
\logit(p) = \text{const} + b\gamma_1 \text{Env} + b\gamma_2 z^*_i \cdot \text{Env}
\]

Analysis model

\[
\logit(P\{y = 1\}) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{Env} + \beta_3 \text{age} \cdot \text{Env}
\]

Ignoring the weights will NOT introduce bias to estimates of \( \beta_1, \beta_2, \) or \( \beta_3 \)!
## Inference for $\beta$ under the True Propensity Score Models

<table>
<thead>
<tr>
<th>Method</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>Mean Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>1.56</td>
<td>0.96</td>
<td>0.21</td>
<td>$4.6E-03$</td>
</tr>
<tr>
<td>IPSW</td>
<td>1.38</td>
<td>-0.14</td>
<td>0.52</td>
<td>$5.3E-03$</td>
</tr>
<tr>
<td>PSAS</td>
<td>1.78</td>
<td>1.25</td>
<td>0.21</td>
<td>$4.7E-03$</td>
</tr>
<tr>
<td>KW</td>
<td>1.27</td>
<td>-0.44</td>
<td>0.60</td>
<td>$5.2E-03$</td>
</tr>
</tbody>
</table>
Future Work – Simulation Studies for Regression Analyses

True analysis model

\[
\text{logit } (P\{y = 1\}) = \beta_0 + \beta_1 \text{age} + \beta_2 E\text{nv} + \beta_3 \text{age} \cdot E\text{nv}
\]

Fitted analysis model

\[
\text{logit } (P\{y = 1\}) = \beta_0 + \beta_2 E\text{nv}
\]

Ignoring the weights will introduce bias to estimate of \(\beta_2\)!

<table>
<thead>
<tr>
<th>Method</th>
<th>Rel Bias(%)</th>
<th>Emp Var</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>32.67</td>
<td>2.8E-03</td>
<td>1.8E-01</td>
</tr>
<tr>
<td>IPSW</td>
<td>0.90</td>
<td>4.3E-03</td>
<td>4.5E-03</td>
</tr>
<tr>
<td>PSAS</td>
<td>14.42</td>
<td>2.5E-03</td>
<td>3.8E-02</td>
</tr>
<tr>
<td><strong>KW</strong></td>
<td><strong>1.46</strong></td>
<td><strong>2.1E-03</strong></td>
<td><strong>2.5E-03</strong></td>
</tr>
</tbody>
</table>
Future Work

KW method

1. Improving efficiency for regression coefficient estimates
   (Pfeffermann & Sverchkov, 1999; Beaumont, 2008; Kim & Skinner, 2013)
2. More statistical models
   e.g. Cox model

Propensity score-based methods

1. Model selection and diagnostics
2. Non-parametric method for propensity score estimation
Acknowledgment

- National Institute of Health/ National Cancer Institute
  DCEG, Biostatistics Branch, and NIH Graduate Partnerships Program
  Mentors, Dr. Barry Graubard and Dr. Hormuzd Katki
  Restricted data from NIH-AARP Diet and Health Study

- The Joint Program in Survey Methodology, UMD
  Dissertation advisor, Dr. Yan Li
A new approach to weighting and inference in sample surveys.  
Biometrika 95, 539-53

Projecting from advance data using propensity modeling: An application to income and tax statistics.  

Inference for Non-probability Samples.  

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.

Weighting in survey analysis under informative sampling.  
Biometrika, 100(2), 385-398.

Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment.  

Parametric and semi-parametric estimation of regression models fitted to survey data.  
Sankhya B 61, 16686.

Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial.  
American journal of epidemiology 165(8), 874-881.

The central role of the propensity score in observational studies for causal effects.  
Biometrika, 70(1), 41-55.

Estimating propensity adjustments for volunteer web surveys.  
Sociological Methods & Research 40(1), 105-137.
Thank You!