Robust and Efficient Methods of Inference for Non-Probability Samples: Application to Naturalistic Driving Data

Ali Rafei[1], Michael R. Elliott[1], Carol A.C. Flannagan[2]

[1]Michigan Program in Survey Methodology
[2]University of Michigan Transportation Research institute

JPSM/MPSM Seminar 2020

September 30

Problem statement

- Probability sampling is the gold standard for finite population inference.

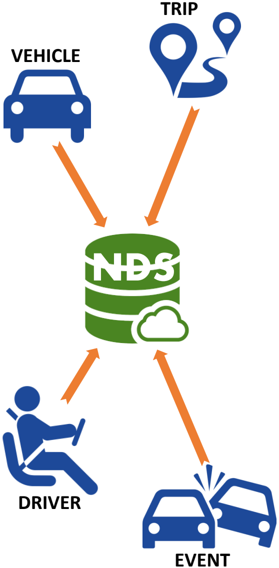
- The 21^{st} century witnesses re-emerging non-probability sampling.

  1. The response rate is steadily declining.
  2. Massive unstructured data are increasingly available.
  3. Convenience samples are easier, cheaper and faster to collect.
  4. Rare events, such as crashes, require long-term followup.

Naturalistic Driving Studies (NDS)

- One real-world application of sensor-based Big Data.
- Driving behaviors are monitored via instrumented vehicles.

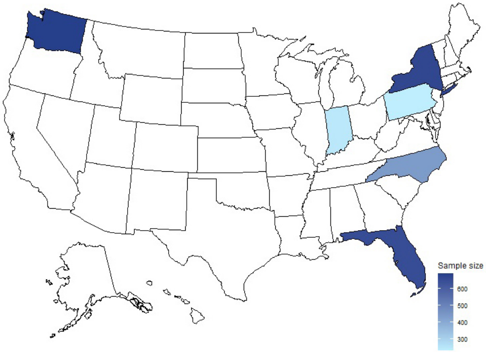
- A rich resource for exploring crash causality, traffic safety, and travel dynamics.



Strategic Highway Research Program 2

SHRP2 SOLUTIONS
STRATEGIC HIGHWAY
RESEARCH PROGRAM

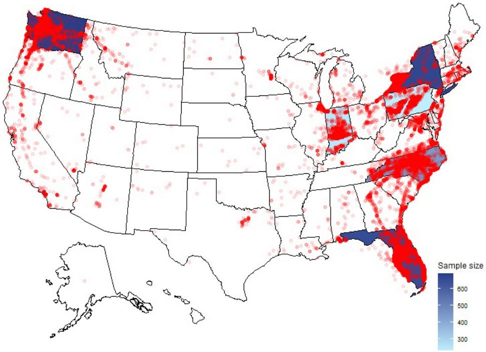
- Launched in 2010, SHRP2 is the largest NDS conducted to date.
- Participants were ∼3,150 volunteers from six sites across the U.S.
- ∼5M trips & ∼50M driven miles were recorded (Trip? time interval during which vehicle is on)
- **Major challenges:**
  1. SHRP2 is a non-probability sample.
  2. Youngest/eldest groups were oversampled.
  3. Only six sites have been studied.



Strategic Highway Research Program 2

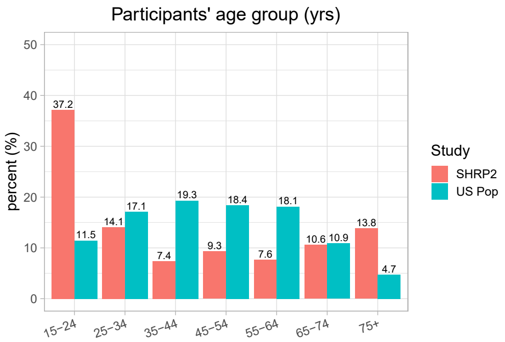
- Launched in 2010, SHRP2 is the largest NDS conducted to date.
- Participants were ∼3,150 volunteers from six sites across the U.S.
- ∼5M trips & ∼50M driven miles were recorded (Trip? time interval during which vehicle is on)
- **Major challenges:**
  1. SHRP2 is a non-probability sample.
  2. Youngest/eldest groups were oversampled.
  3. Only six sites have been studied.

SHRP2 SOLUTIONS
STRATEGIC HIGHWAY RESEARCH PROGRAM



Strategic Highway Research Program 2

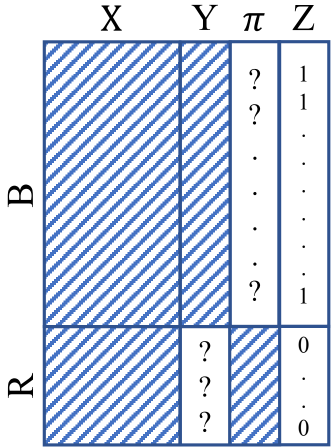
- Launched in 2010, SHRP2 is the largest NDS conducted to date.
- Participants were ∼3,150 volunteers from six sites across the U.S.
- ∼5M trips & ∼50M driven miles were recorded. (Trip? time interval during which vehicle is on)



- **Major challenges:**
  1. SHRP2 is a non-probability sample.
  2. Youngest/eldest groups were oversampled.
  3. Only six sites have been studied.

- Let's define the following notations:
  1. B: Big non-probability sample
  2. R: Reference survey
  3. X: Set of common auxiliary vars
  4. Y: Outcome var of interest
  5. Z: Indicator of being in B

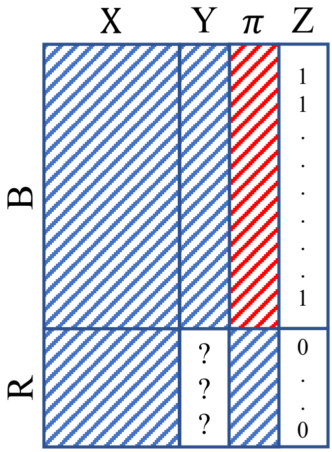
- Considering MAR+positivity assumptions given X:
  1. **Quasi-randomization (QR):** Estimating pseudo-inclusion probabilities (π^B) in B
  2. **Prediction modeling (PM):** Predicting the outcome var (Y) for units in R
  3. **Doubly robust Adjustment (DR):** Combining the two to further protect against model misspecification



Combined sample

Basic framework

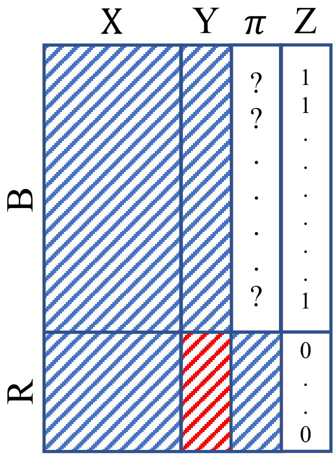
- Let's define the following notations:
  1. B: Big non-probability sample
  2. R: Reference survey
  3. X: Set of common auxiliary vars
  4. Y: Outcome var of interest
  5. Z: Indicator of being in B
- Considering MAR+positivity assumptions given X:
  1. **Quasi-randomization (QR):**
     Estimating pseudo-inclusion probabilities (π^B) in B
  2. **Prediction modeling (PM):**
     Predicting the outcome var (Y) for units in R
  3. **Doubly robust Adjustment (DR):**
     Combining the two to further protect against model misspecification



Combined sample

Basic framework

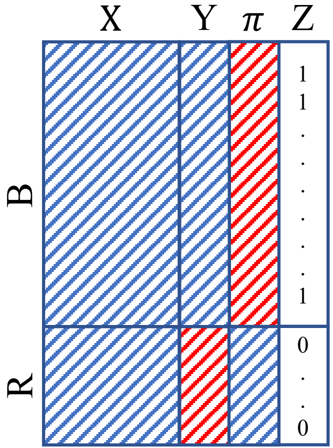
- Let's define the following notations:
  1. B: Big non-probability sample
  2. R: Reference survey
  3. X: Set of common auxiliary vars
  4. Y: Outcome var of interest
  5. Z: Indicator of being in B
- Considering MAR+positivity assumptions given X:
  1. **Quasi-randomization (QR):**
     Estimating pseudo-inclusion probabilities (π^B) in B
  2. **Prediction modeling (PM):**
     Predicting the outcome var (Y) for units in R
  3. **Doubly robust Adjustment (DR):**
     Combining the two to further protect against model misspecification



Combined sample

Basic framework

- Let's define the following notations:
  1. B: Big non-probability sample
  2. R: Reference survey
  3. X: Set of common auxiliary vars
  4. Y: Outcome var of interest
  5. Z: Indicator of being in B
- Considering MAR+positivity assumptions given X:
  1. **Quasi-randomization (QR):**
     Estimating pseudo-inclusion probabilities (π^B) in B
  2. **Prediction modeling (PM):**
     Predicting the outcome var (Y) for units in R
  3. **Doubly robust Adjustment (DR):**
     Combining the two to further protect against model misspecification



Combined sample

Quasi-randomization

- Traditionally, propensity scores are used to estimate pseudo-weights (Lee 2006).

PS weighting when R is *epsem*:

$$\bar{y}_{PW} = \frac{1}{N}\sum_{i=1}^{n_B} \frac{y_i}{\pi^B(x_i)}$$

where under a logistic regression model, we have

$$\pi^B(x_i) \propto p_i(\beta) = P(Z_i = 1|x_i;\beta) = \frac{exp\{x_i^T\beta\}}{1+exp\{x_i^T\beta\}}, \quad \forall i \in B$$

- When R is NOT *epsem*, β can be estimated through a PMLE approach by solving:
  1. $\sum_{i\in B} x_i[1-p_i(\beta)] - \sum_{i\in R} x_i p_i(\beta)/\pi_i^R = 0$ (odds of PS) (Wang et al. 2020)
  2. $\sum_{i\in B} x_i - \sum_{i\in R} x_i p_i(\beta)/\pi_i^R = 0$ (Chen et al. 2019)
  3. $\sum_{i\in B} x_i/p_i(\beta) - \sum_{i\in R} x_i/\pi_i^R = 0$ (Kim 2020)

Quasi-randomization

- Traditionally, propensity scores are used to estimate pseudo-weights (Lee 2006).

PS weighting when R is *epsem*:

$$\bar{y}_{PW} = \frac{1}{N}\sum_{i=1}^{n_B}\frac{y_i}{\pi^B(x_i)}$$

where under a logistic regression model, we have

$$\pi^B(x_i) \propto p_i(\beta) = P(Z_i = 1|x_i;\beta) = \frac{exp\{x_i^T\beta\}}{1+exp\{x_i^T\beta\}}, \quad \forall i \in B$$

- When R is NOT *epsem*, β can be estimated through a PMLE approach by solving:
  1. $\sum_{i\in B} x_i[1-p_i(\beta)] - \sum_{i\in R} x_i p_i(\beta)/\pi_i^R = 0$ (odds of PS) (Wang et al. 2020)
  2. $\sum_{i\in B} x_i - \sum_{i\in R} x_i p_i(\beta)/\pi_i^R = 0$ (Chen et al. 2019)
  3. $\sum_{i\in B} x_i/p_i(\beta) - \sum_{i\in R} x_i/\pi_i^R = 0$ (Kim 2020)

Quasi-randomization

- However, the PMLE approach is limited to the parametric models.
- One may be interested in applying more flexible non-parametric methods.
- Denote $\delta_i = \delta_i^B + \delta_i^R$. With an additional assumption $B \cap R = \emptyset$, one can show

$$\pi_i^B = P(\delta_i^B = 1|x_i, \pi_i^R) = P(\delta_i = 1|x_i, \pi_i^R)P(Z_i = 1|x_i, \pi_i^R)$$
$$\pi_i^R = P(\delta_i^R = 1|x_i, \pi_i^R) = P(\delta_i = 1|x_i, \pi_i^R)P(Z_i = 0|x_i, \pi_i^R)$$

Propensity Adjusted Probability weighting (PAPW):

$$\pi_i^B(x_i^*; \beta^*) = \pi_i^R \frac{p_i(\beta^*)}{1 - p_i(\beta^*)}, \quad \forall i \in B$$

where $x_i^* = [x_i, \pi_i^R]$, and β^* can be estimated through the regular MLE.

- This is especially advantageous when applying a broader range of predictive methods.

Quasi-randomization

- However, the PMLE approach is limited to the parametric models.
- One may be interested in applying more flexible non-parametric methods.
- Denote $\delta_i = \delta_i^B + \delta_i^R$. With an additional assumption $B \cap R = \emptyset$, one can show

$$\pi_i^B = P(\delta_i^B = 1|x_i, \pi_i^R) = P(\delta_i = 1|x_i, \pi_i^R)P(Z_i = 1|x_i, \pi_i^R)$$
$$\pi_i^R = P(\delta_i^R = 1|x_i, \pi_i^R) = P(\delta_i = 1|x_i, \pi_i^R)P(Z_i = 0|x_i, \pi_i^R)$$

Propensity Adjusted Probability weighting (PAPW):

$$\pi_i^B(x_i^*; \beta^*) = \pi_i^R \frac{p_i(\beta^*)}{1 - p_i(\beta^*)}, \qquad \forall i \in B$$

where $x_i^* = [x_i, \pi_i^R]$, and β^* can be estimated through the regular MLE.

- This is especially advantageous when applying a broader range of predictive methods.

Quasi-randomization

- However, the PMLE approach is limited to the parametric models.
- One may be interested in applying more flexible non-parametric methods.
- Denote $\delta_i = \delta_i^B + \delta_i^R$. With an additional assumption $B \cap R = \emptyset$, one can show

$$\pi_i^B = P(\delta_i^B = 1|x_i, \pi_i^R) = P(\delta_i = 1|x_i, \pi_i^R)P(Z_i = 1|x_i, \pi_i^R)$$
$$\pi_i^R = P(\delta_i^R = 1|x_i, \pi_i^R) = P(\delta_i = 1|x_i, \pi_i^R)P(Z_i = 0|x_i, \pi_i^R)$$

Propensity Adjusted Probability weighting (PAPW):

$$\pi_i^B(x_i^*; \beta^*) = \pi_i^R \frac{p_i(\beta^*)}{1 - p_i(\beta^*)}, \qquad \forall i \in B$$

where $x_i^* = [x_i, \pi_i^R]$, and β^* can be estimated through the regular MLE.

- This is especially advantageous when applying a broader range of predictive methods.

Quasi-randomization

- Under certain regularity conditions, one can prove that $\hat{\bar{y}}_{PW} = \bar{y}_U + O_p(n_B^{-1/2})$.
- When π_i^R is unknown for $i \in B$, Elliott & Valliant (2017) show that

Propensity Adjusted Probability Prediction (PAPP):

$$\pi_i^B(x_i;\beta,\gamma) = P(\delta_i^R = 1|x_i;\gamma)\frac{p_i(\beta)}{1-p_i(\beta)}, \qquad \forall i \in B$$

where γ is the vector of parameters in modeling δ_i^R on x_i.

- To predict $P(\delta_i^R = 1|x_i;\gamma)$ for $i \in B$, one can model π_i^R on x_i instead of δ_i^R because

$$P(\delta_i^R = 1|x_i) = \int_0^1 P(\delta_i^R = 1|\pi_i^R, x_i)P(\pi_i^R|x_i)d\pi_i^R$$
$$= \int_0^1 \pi_i^R P(\pi_i^R|x_i)d\pi_i^R = E(\pi_i^R|x_i) \qquad i \in R$$

Quasi-randomization

- Under certain regularity conditions, one can prove that $\hat{\bar{y}}_{PW} = \bar{y}_U + O_p(n_B^{-1/2})$.
- When π_i^R is unknown for $i \in B$, Elliott & Valliant (2017) show that

Propensity Adjusted Probability Prediction (PAPP):

$$\pi_i^B(x_i;\beta,\gamma) = P(\delta_i^R = 1|x_i;\gamma)\frac{p_i(\beta)}{1-p_i(\beta)}, \qquad \forall i \in B$$

where γ is the vector of parameters in modeling δ_i^R on x_i.

- To predict $P(\delta_i^R = 1|x_i;\gamma)$ for $i \in B$, one can model π_i^R on x_i instead of δ_i^R because

$$\begin{aligned} P(\delta_i^R = 1|x_i) &= \int_0^1 P(\delta_i^R = 1|\pi_i^R, x_i)P(\pi_i^R|x_i)d\pi_i^R \\ &= \int_0^1 \pi_i^R P(\pi_i^R|x_i)d\pi_i^R = E(\pi_i^R|x_i) \qquad i \in R \end{aligned}$$

Doubly robust adjustment

- Augmented Inverse Propensity weighting (AIPW) was proposed by Robins et al (1994).

Chen et al (2019) extend AIPW to a non-probability sample setting

$$\hat{\bar{y}}_{DR} = \frac{1}{N}\sum_{i=1}^{n_B}\frac{\{y_i - m(x_i;\theta)\}}{\pi_i^B(x_i;\beta)} + \frac{1}{N}\sum_{j=1}^{n_R}\frac{m(x_j;\theta)}{\pi_j^R}$$

where $m(.)$ is a continuous differentiable function w.r.t. θ.

- Parameteres $\eta = (\beta, \theta)$ are estimated by simultaneously solving (Kim & Haziza 2014):

$$\frac{\partial}{\partial\beta}\left[\bar{y}_{DR} - \bar{y}_U\right] = \frac{1}{N}\sum_{i=1}^{N}\delta_i^B\left[\frac{1}{\pi_i^B(x_i;\beta)} - 1\right]\{y_i - m(x_i;\theta)\}x_i = 0$$

$$\frac{\partial}{\partial\theta}\left[\bar{y}_{DR} - \bar{y}_U\right] = \frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i^B}{\pi^B(x_i;\beta)}\dot{m}(x_i;\theta) - \sum_{i=1}^{n_R}\frac{\dot{m}(x_i;\theta)}{\pi_i^R} = 0$$

Doubly robust adjustment

- Augmented Inverse Propensity weighting (AIPW) was proposed by Robins et al (1994).

Chen et al (2019) extend AIPW to a non-probability sample setting

$$\hat{\bar{y}}_{DR} = \frac{1}{N}\sum_{i=1}^{n_B}\frac{\{y_i - m(x_i;\theta)\}}{\pi_i^B(x_i;\beta)} + \frac{1}{N}\sum_{j=1}^{n_R}\frac{m(x_j;\theta)}{\pi_j^R}$$

where $m(.)$ is a continuous differentiable function w.r.t. θ.

- Parameteres $\eta = (\beta, \theta)$ are estimated by simultaneously solving (Kim & Haziza 2014):

$$\frac{\partial}{\partial\beta}\left[\bar{y}_{DR} - \bar{y}_U\right] = \frac{1}{N}\sum_{i=1}^{N}\delta_i^B\left[\frac{1}{\pi_i^B(x_i;\beta)} - 1\right]\{y_i - m(x_i;\theta)\}x_i = 0$$

$$\frac{\partial}{\partial\theta}\left[\bar{y}_{DR} - \bar{y}_U\right] = \frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i^B}{\pi^B(x_i;\beta)}\dot{m}(x_i;\theta) - \sum_{i=1}^{n_R}\frac{\dot{m}(x_i;\theta)}{\pi_i^R} = 0$$

Adjusted DR estimator

- However, if both QR and PM are incorrectly specified, the estimates are still biased.
- To avoid using PMLE, we recommend using PAPW/PAPP approach for predicting π_i^B.

Proposed AIPW estimator when π_i^R is calculable for $i \in B$:

$$\bar{y}_{DR} = \frac{1}{N}\sum_{i=1}^{n_B}\frac{1}{\pi_i^R}\left[\frac{1-p_i(\beta^*)}{p_i(\beta^*)}\right]\{y_i - m(x_i^*;\theta^*)\} + \frac{1}{N}\sum_{j=1}^{n_R}\frac{m(x_j^*;\theta^*)}{\pi_j^R}$$

where θ^* is the vector of parameters associated with $x_i^* = [x_i, \pi_i^R]$.

- Assuming that y_i is observed for $i \in R$, denote $\bar{y}_R = N^{-1}\sum_{i=1}^{n_R} y_i/\pi_i^R$. We have

$$\bar{y}_{DR} - \bar{y}_R = \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i^R}\left[\frac{Z_i}{p_i(\beta^*)} - 1\right]\{y_i - m(x_i^*;\theta^*)\}$$

which is identical to what Kim & Haziza (2014) derived for incomplete data inference.

Adjusted DR estimator

- However, if both QR and PM are incorrectly specified, the estimates are still biased.
- To avoid using PMLE, we recommend using PAPW/PAPP approach for predicting π_i^B.

Proposed AIPW estimator when π_i^R is calculable for $i \in B$:

$$\bar{y}_{DR} = \frac{1}{N}\sum_{i=1}^{n_B}\frac{1}{\pi_i^R}\left[\frac{1-p_i(\beta^*)}{p_i(\beta^*)}\right]\{y_i - m(x_i^*;\theta^*)\} + \frac{1}{N}\sum_{j=1}^{n_R}\frac{m(x_j^*;\theta^*)}{\pi_j^R}$$

where θ^* is the vector of parameters associated with $x_i^* = [x_i, \pi_i^R]$.

- Assuming that y_i is observed for $i \in R$, denote $\bar{y}_R = N^{-1}\sum_{i=1}^{n_R} y_i/\pi_i^R$. We have

$$\bar{y}_{DR} - \bar{y}_R = \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i^R}\left[\frac{Z_i}{p_i(\beta^*)} - 1\right]\{y_i - m(x_i^*;\theta^*)\}$$

which is identical to what Kim & Haziza (2014) derived for incomplete data inference.

Adjusted DR estimator

- Therefore, under GLM, we recommend estimating $\eta^* = (\beta^*, \theta^*)$ by solving:

$$\frac{\partial}{\partial \beta^*}\left[\bar{y}_{DR} - \bar{y}_R\right] = \frac{1}{N}\sum_{i=1}^{n}\frac{Z_i}{\pi_i^R}\left[\frac{1}{p_i(\beta^*)} - 1\right]\{y_i - m(x_i^*; \theta^*)\}x_i^* = 0$$

$$\frac{\partial}{\partial \theta^*}\left[\bar{y}_{DR} - \bar{y}_R\right] = \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i^R}\left[\frac{Z_i}{p_i(\beta^*)} - 1\right]\dot{m}(x_i^*; \theta^*) = 0$$

- Under some regularity conditions, one can prove that $\hat{\bar{y}}_{DR} = \bar{y}_{DR} + O_p(n^{-1/2})$.

- Note that using π_i^R as a predictor in $m(.)$ further weakens the modeling assumption.

Proposed AIPW estimator when π_i^R is unknown for $i \in B$:

$$\bar{y}_{DR} = \frac{1}{N}\sum_{i=1}^{n_B}\frac{1}{\pi_i^R(x_i; \gamma)}\left[\frac{1 - p_i(\beta)}{p_i(\beta)}\right]\{y_i - m(x_i; \theta)\} + \frac{1}{N}\sum_{j=1}^{n_R}\frac{m(x_j; \theta)}{\pi_j^R}$$

Adjusted DR estimator

- Therefore, under GLM, we recommend estimating $\eta^* = (\beta^*, \theta^*)$ by solving:

$$\frac{\partial}{\partial\beta^*}\left[\bar{y}_{DR} - \bar{y}_R\right] = \frac{1}{N}\sum_{i=1}^{n}\frac{Z_i}{\pi_i^R}\left[\frac{1}{p_i(\beta^*)} - 1\right]\{y_i - m(x_i^*;\theta^*)\}x_i^* = 0$$

$$\frac{\partial}{\partial\theta^*}\left[\bar{y}_{DR} - \bar{y}_R\right] = \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i^R}\left[\frac{Z_i}{p_i(\beta^*)} - 1\right]\dot{m}(x_i^*;\theta^*) = 0$$

- Under some regularity conditions, one can prove that $\hat{\bar{y}}_{DR} = \bar{y}_{DR} + O_p(n^{-1/2})$.
- Note that using π_i^R as a predictor in $m(.)$ further weakens the modeling assumption.

Proposed AIPW estimator when π_i^R is unknown for $i \in B$:

$$\bar{y}_{DR} = \frac{1}{N}\sum_{i=1}^{n_B}\frac{1}{\pi_i^R(x_i;\gamma)}\left[\frac{1 - p_i(\beta)}{p_i(\beta)}\right]\{y_i - m(x_i;\theta)\} + \frac{1}{N}\sum_{j=1}^{n_R}\frac{m(x_j;\theta)}{\pi_j^R}$$

Bayesian Additive Regression Trees (BART)

- BART is a flexible sum-of-trees regression method (Chipman et al 2010).

BART structure:

$$y_i = \sum_{j=1}^{m} f(x_i, T_j, M_j) + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and T_j is the jth tree with M_j being terminal node parameters.

Bayesian Additive Regression Trees (BART)

- BART is a flexible sum-of-trees regression method (Chipman et al 2010).

BART structure:

$$y_i = \sum_{j=1}^{m} f(x_i, T_j, M_j) + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and T_j is the jth tree with M_j being terminal node parameters.

- BART is Bayesian assigning prior distributions to T (length & decision rules), M, and σ.
- Considering independent structure between trees:

$$p[(T_1, M_1), ..., (T_m, M_m), \sigma^{-2}] = [\prod_{j=1}^{m}\{\prod_{i=1}^{b_j} P(\mu_{ij}|T_j)\}P(T_j)]P(\sigma^{-2})$$

- Given the data, posterior distribution is simulated using a backfitting MCMC method.

Bayesian Additive Regression Trees (BART)

- Advantages of BART: automatic variable selection, quantifying uncertainty using PPD.
- For a binary outcome, BART uses a data augmentation approach to transform Y into $\mathbb{R}$.

Extending the modified DR method using BART:

$$log\left(\frac{\pi_i^R}{1-\pi_i^R}\right) = k(x_i) + \epsilon_i, \qquad \Phi^{-1}[P(Z_i = 1|x_i)] = h(x_i), \qquad y_i = f(x_i) + \epsilon_i$$

For a given MCMC draw, m ($m = 1, 2, ..., M$), we have

$$\hat{\bar{y}}_{DR}^{(m)} = \frac{1}{\hat{N}_B}\sum_{i=1}^{n_B}\left\{\frac{1+exp[\hat{k}^{(m)}(x_i)]}{exp[\hat{k}^{(m)}(x_i)]}\right\}\left\{\frac{1-\Phi[\hat{h}^{(m)}(x_i)]}{\Phi[\hat{h}^{(m)}(x_i)]}\right\}\{y_i - \hat{f}^{(m)}(x_i)\} + \frac{1}{\hat{N}_R}\sum_{j=1}^{n_R}\frac{\hat{f}^{(m)}(x_j)}{\pi_j^R}$$

Final AIPW estimator under BART: $\hat{\bar{y}}_{DR} = \frac{1}{M}\sum_{m=1}^{M}\hat{\bar{y}}_{DR}^{(m)}$

Variance estimation

- To estimate variance, one has to incorporate uncertainty due to sampling, imputing pseudo-weights, and predicting the outcome. Two methods are proposed:
- Asymptotic variance estimator when π_i^R is known for $i \in B$
  - For pseudo-weighting approach based on PAPW:

$$\widehat{Var}\left(\hat{\bar{y}}_{PW}\right) = \frac{1}{N^2}\sum_{i=1}^{n_B}\{1-\hat{\pi}_i^B\}\left(\frac{y_i-\hat{\bar{y}}_{PW}}{\hat{\pi}_i^B}\right)^2 - 2\frac{\hat{b}^T}{N^2}\sum_{i=1}^{n_B}\{1-p_i(\hat{\beta}_1)\}\left(\frac{y_i-\hat{\bar{y}}_{PW}}{\hat{\pi}_i^B}\right)x_i + \hat{b}^T\left[\frac{1}{N^2}\sum_{i=1}^{n}p_i(\hat{\beta}_1)x_i x_i^T\right]\hat{b}$$

where $\hat{b}^T = \left\{\frac{1}{N}\sum_{i=1}^{n_B}\left(\frac{y_i-\hat{\bar{y}}_{PW}}{\hat{\pi}_i^B}\right)x_i^T\right\}\left\{\frac{1}{N}\sum_{i=1}^{n}p_i(\hat{\beta}_1)x_i x_i^T\right\}^{-1}$

Variance estimation

- To estimate variance, one has to incorporate uncertainty due to sampling, imputing pseudo-weights, and predicting the outcome. Two methods are proposed:
- Asymptotic variance estimator when π_i^R is known for $i \in B$
  - For pseudo-weighting approach based on PAPW:

$$\widehat{Var}\left(\hat{\bar{y}}_{PW}\right) = \frac{1}{N^2}\sum_{i=1}^{n_B}\{1-\hat{\pi}_i^B\}\left(\frac{y_i-\hat{\bar{y}}_{PW}}{\hat{\pi}_i^B}\right)^2 - 2\frac{\hat{b}^T}{N^2}\sum_{i=1}^{n_B}\{1-p_i(\hat{\beta}_1)\}\left(\frac{y_i-\hat{\bar{y}}_{PW}}{\hat{\pi}_i^B}\right)x_i + \hat{b}^T\left[\frac{1}{N^2}\sum_{i=1}^{n}p_i(\hat{\beta}_1)x_ix_i^T\right]\hat{b}$$

where $\hat{b}^T = \left\{\frac{1}{N}\sum_{i=1}^{n_B}\left(\frac{y_i-\hat{\bar{y}}_{PW}}{\hat{\pi}_i^B}\right)x_i^T\right\}\left\{\frac{1}{N}\sum_{i=1}^{n}p_i(\hat{\beta}_1)x_ix_i^T\right\}^{-1}$

  - For the modified AIPW estimator (Chen et al 2019):

$$\widehat{Var}(\hat{\bar{y}}_{DR}) = \hat{V}_1 + \hat{V}_2 - \hat{B}(\hat{V}_2)$$

where

$$\hat{V}_1 = \widehat{Var}(\hat{\bar{y}}_{PM}), \qquad \hat{V}_2 = \frac{1}{N^2}\sum_{i=1}^{n_B}\left[\frac{1-\hat{\pi}_i^B}{(\hat{\pi}_i^B)^2}\right]\{y_i - m(x_i^*;\hat{\theta}_1)\}^2, \qquad \hat{B}(\hat{V}_2) = \frac{1}{N^2}\sum_{i=1}^{n}\left[\frac{Z_i}{\hat{\pi}_i^B} - \frac{1-Z_i}{\pi_i^R}\right]\hat{\sigma}_i^2$$

Variance estimation

- Variance estimation when π_i^R is incomputable for $i \in B$:
- Under GLM:
  - A modified **bootstrap resampling** method (Rao & Wu, 1991)
    1. Draw M bootstrap samples of sizes $n_B - 1$ and $n_R - 1$ from B and R to estimate $\hat{\bar{y}}_{DR}^{(m)}$'s.
    2. Update the sampling weights in R to $w_i^{(m)} = w_i \frac{n_R}{n_R - 1} t_i$.

$$\widehat{Var}(\hat{\bar{y}}_{DR}^{(m)}) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{\bar{y}}_{DR}^{(m)} - \bar{\bar{y}}_{DR} \right]^2$$

Variance estimation

- Variance estimation when π_i^R is incomputable for $i \in B$:
- Under GLM:
  - A modified **bootstrap resampling** method (Rao & Wu, 1991)
    1. Draw M bootstrap samples of sizes $n_B - 1$ and $n_R - 1$ from B and R to estimate $\hat{\bar{y}}_{DR}^{(m)}$'s.
    2. Update the sampling weights in R to $w_i^{(m)} = w_i \frac{n_R}{n_R - 1} t_i$.

$$\widehat{Var}(\hat{\bar{y}}_{DR}^{(m)}) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{\bar{y}}_{DR}^{(m)} - \bar{\bar{y}}_{DR} \right]^2$$

- Under BART:
  - A multiple imputation method using the **posterior predictive** draws
    1. Randomly select a sample of size M from posterior predictive draws, and estimate $\hat{\bar{y}}_{DR}^{(m)}$.
    2. Use Rubin's combining rules to construct point/variance estimates.

$$\widehat{Var}(\hat{\bar{y}}_{DR}) = \bar{V}_W + (1 + 1/M) V_B$$

where $\bar{V}_w = \sum_{m=1}^{M} var\{\hat{\bar{y}}_{DR}^{(m)}\}/M$ and $V_B = \sum_{m=1}^{M} [\hat{\bar{y}}_{DR}^{(m)} - \bar{\bar{y}}_{DR}]^2/(M-1)$

Simulation study I (Chen et al 2019)

- A pop. of size $N = 1,000,000$ was generated with the following variables:

$$z_{1i} \sim Ber(p = 0.5) \qquad z_{2i} \sim U(0,2) \qquad z_{3i} \sim Exp(\mu = 1) \qquad z_{4i} \sim \chi^2_{(4)}$$

$$x_{1i} = z_{1i} \qquad x_{2i} = z_{2i} + 0.3z_{1i} \qquad x_{3i} = z_{3i} + 0.2(x_{1i} + x_{2i}) \qquad x_{4i} = z_{4i} + 0.1(x_{1i} + x_{2i} + x_{3i})$$

- Y is a continuous outcome with normal distribution as below:

$$Y_i = 2 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + 0.5\epsilon_i \qquad where \quad \epsilon_i \sim N(0,1)$$

- Two sets of unequal selection probabilities, are generated as below:

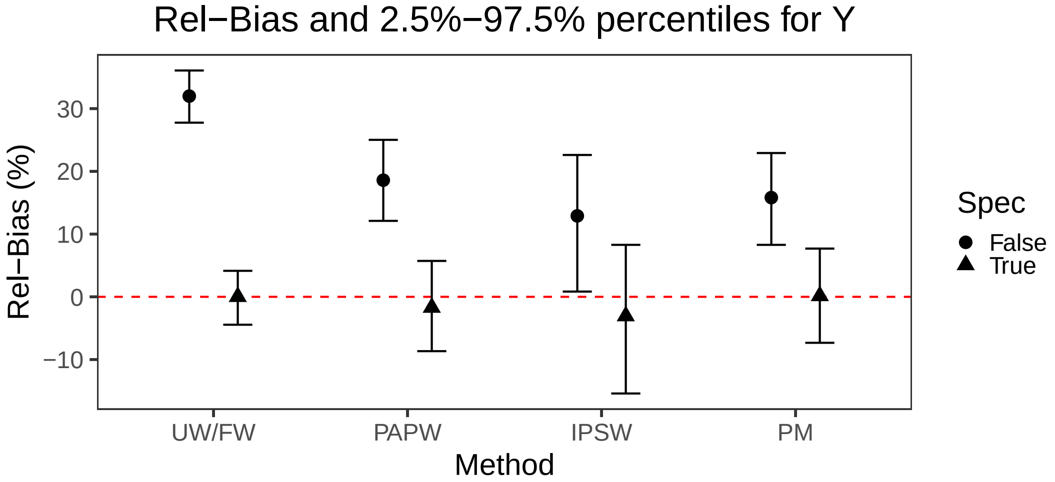
$$\pi_i^R \propto \gamma_1 + z_{3i}, \qquad log\left(\frac{\pi_i^B}{1-\pi_i^B}\right) = \gamma_0 + 0.1x_{1i} + 0.2x_{2i} + 0.1x_{3i} + 0.2x_{4i}$$

- The simulation was iterated $K = 1000$ times, and rel-Bias, rMSE, 95%CI coverage rates and SE ratio were computed.

- Different scenarios of model misspecification were examined.

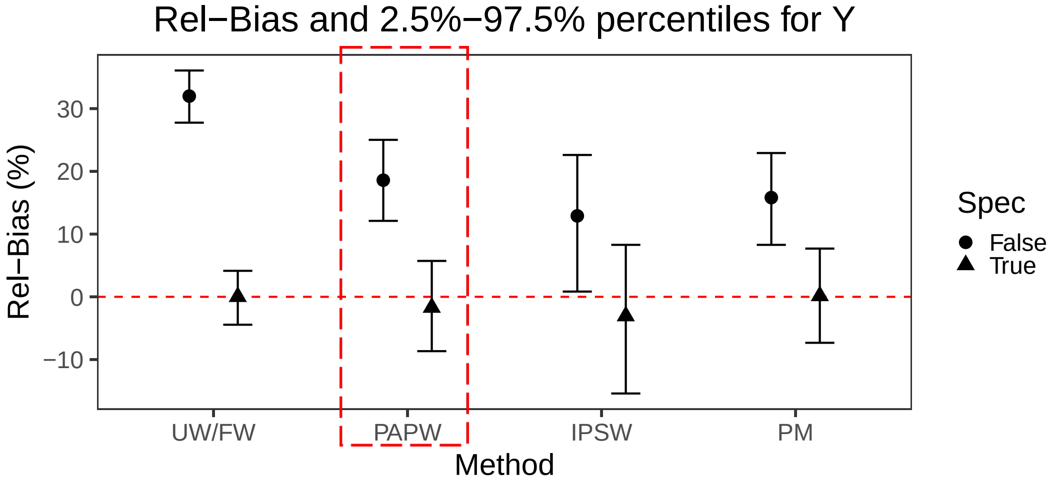
Simulation results I

- The simulation results for $n_R = 100$ and $n_B = 1,000$



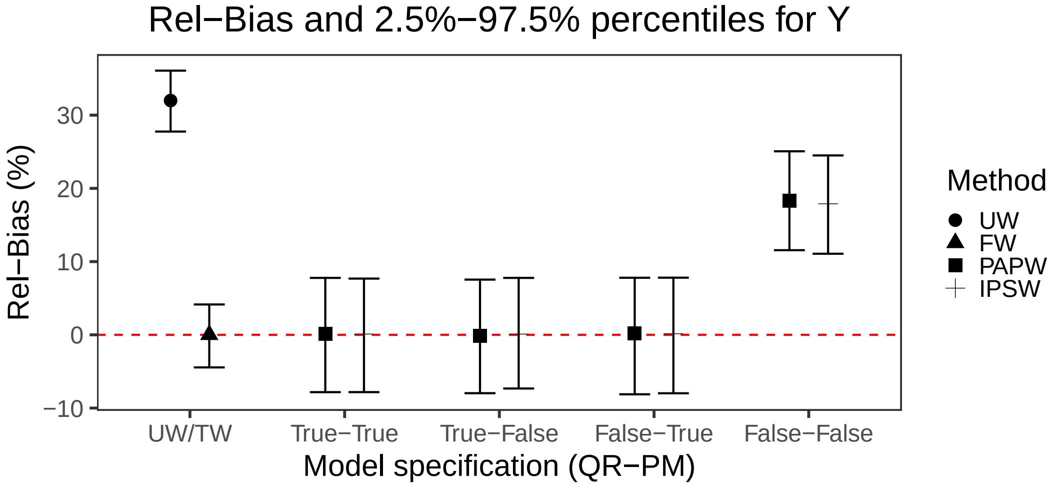
Simulation results I

- The simulation results for $n_R = 100$ and $n_B = 1,000$

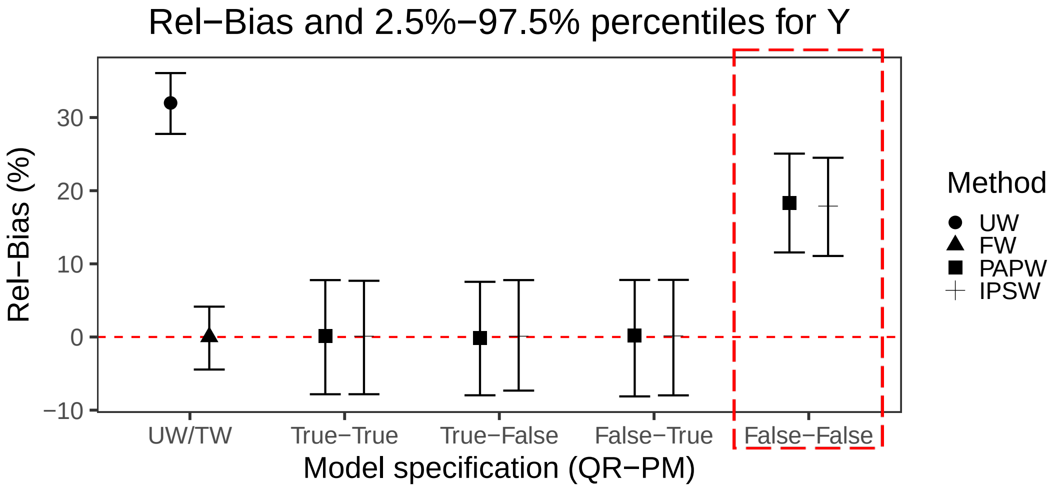


Simulation results I

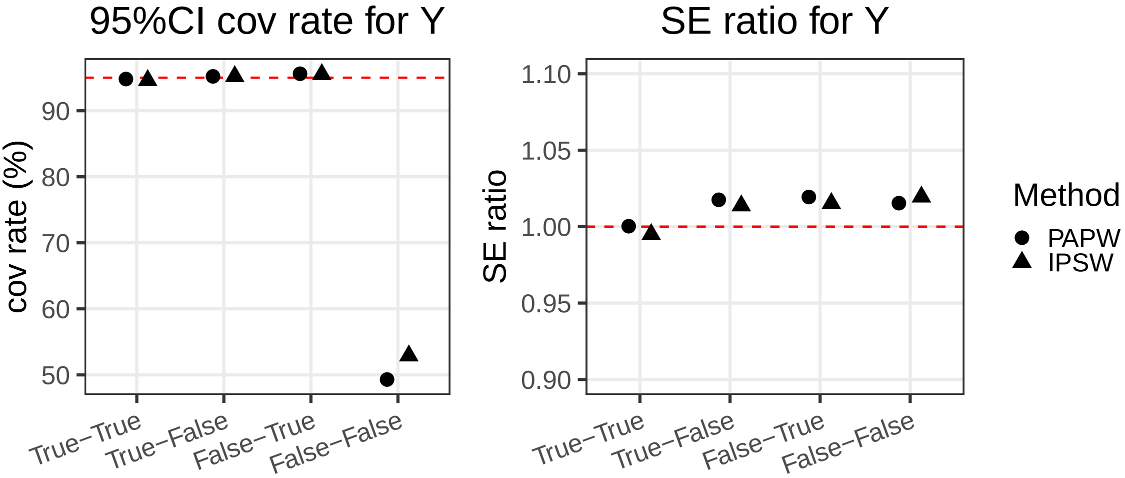
- The simulation results for $n_R = 100$ and $n_B = 1,000$



- The simulation results for $n_R = 100$ and $n_B = 1,000$



- The simulation results for $n_R = 100$ and $n_B = 1,000$



Simulation study II

- A clustered pop. of size $A = 1,000$ and $n_\alpha = 1,000$ was generated as below:

$$\begin{pmatrix} X_{1\alpha} \\ D_\alpha \end{pmatrix} \sim MVN(\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}) \quad , \quad X_{2\alpha} \sim Ber(p = 0.5)$$

- Y is a continuous outcome with normal distribution as below:

$$Y_{\alpha i}|X_\alpha, d_\alpha \sim N(\mu = 2 + 0.4x_{1\alpha}^2 + 0.3x_{1\alpha}^3 - 0.2x_{2\alpha} - 0.1x_{1\alpha}x_{2\alpha} - d_\alpha + u_\alpha, \sigma^2 = 1)$$

- Two sets of unequal selection probabilities, are generated as below:

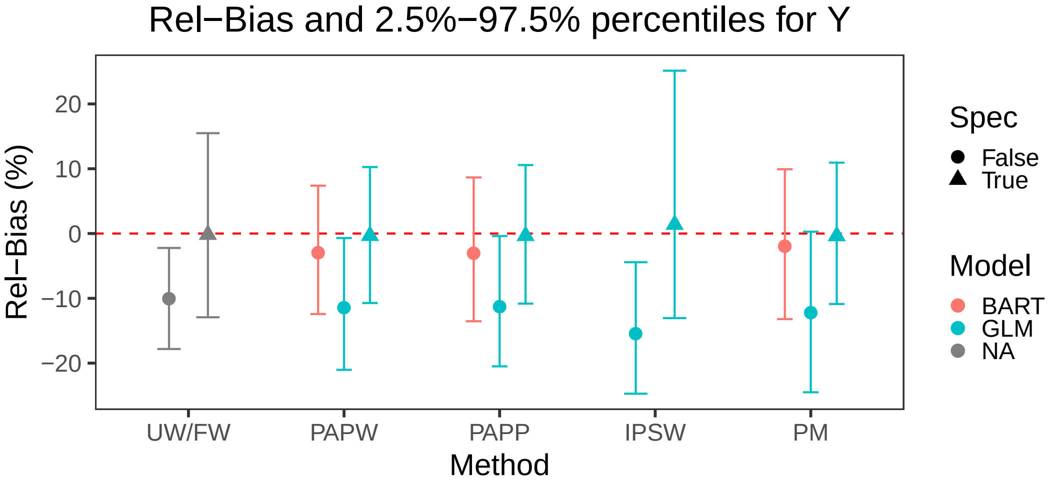
$$P(\delta_\alpha^R = 1|d) = \frac{e^{\gamma_0 + 0.5d_\alpha}}{1 + e^{\gamma_0 + 0.5d_\alpha}}, \qquad P(\delta_\alpha^B = 1|x) = \frac{e^{\gamma_1 + 0.4x_{1\alpha} - 0.2x_{1\alpha}^2 + 0.6x_{2\alpha} + 0.1x_{1\alpha}x_{2\alpha}}}{1 + e^{\gamma_1 + 0.4x_{1\alpha} - 0.2x_{1\alpha}^2 + 0.6x_{2\alpha} + 0.1x_{1\alpha}x_{2\alpha}}}$$

- The simulation was iterated $K = 1000$ times, and rel-Bias, rMSE, 95%CI coverage rates and SE ratio were computed.

- Different scenarios of model misspecification were examined.

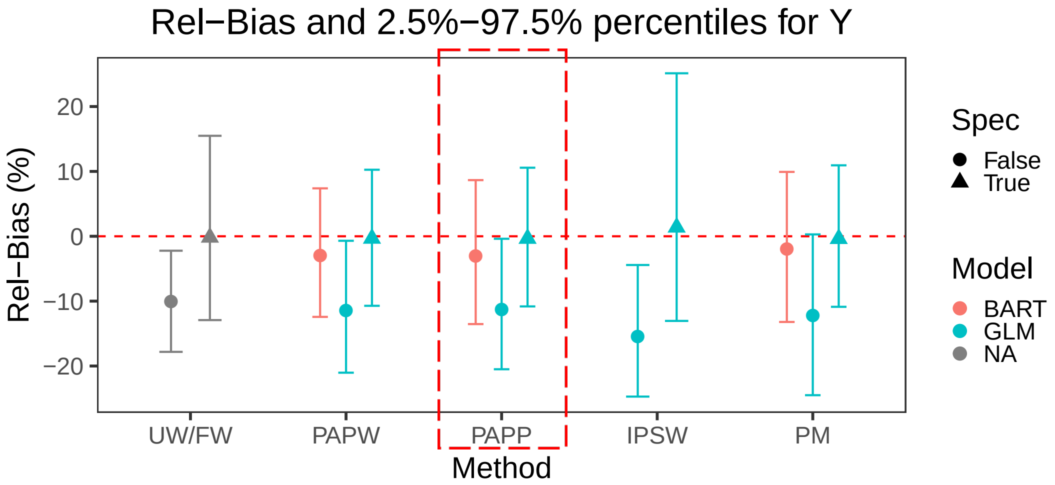
Simulation results II

- The simulation results for $n_{R\alpha} = 100$ and $n_{B\alpha} = 50$ and $a = 200$:



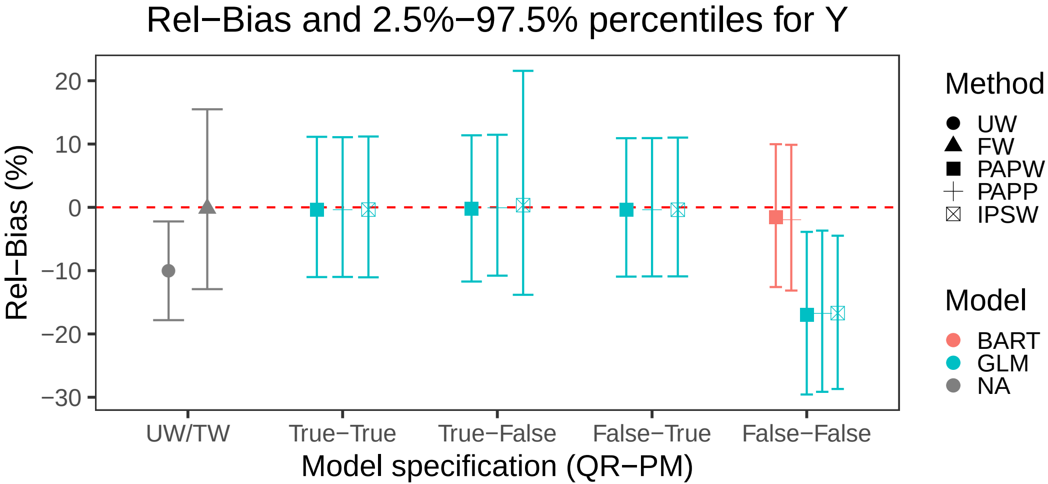
Simulation results II

- The simulation results for $n_{R\alpha} = 100$ and $n_{B\alpha} = 50$ and $a = 200$:



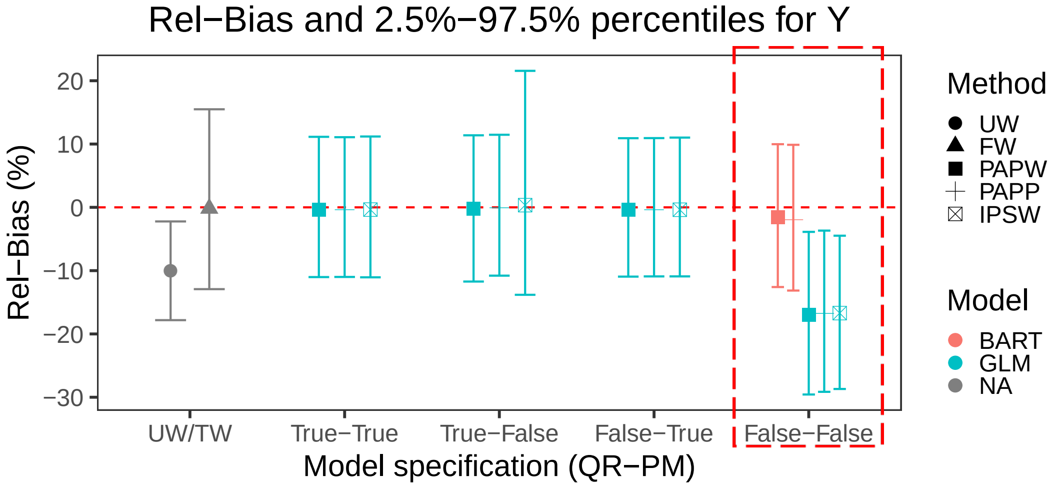
Simulation results II

- The simulation results for $n_{R\alpha} = 100$ and $n_{B\alpha} = 50$ and $a = 200$:

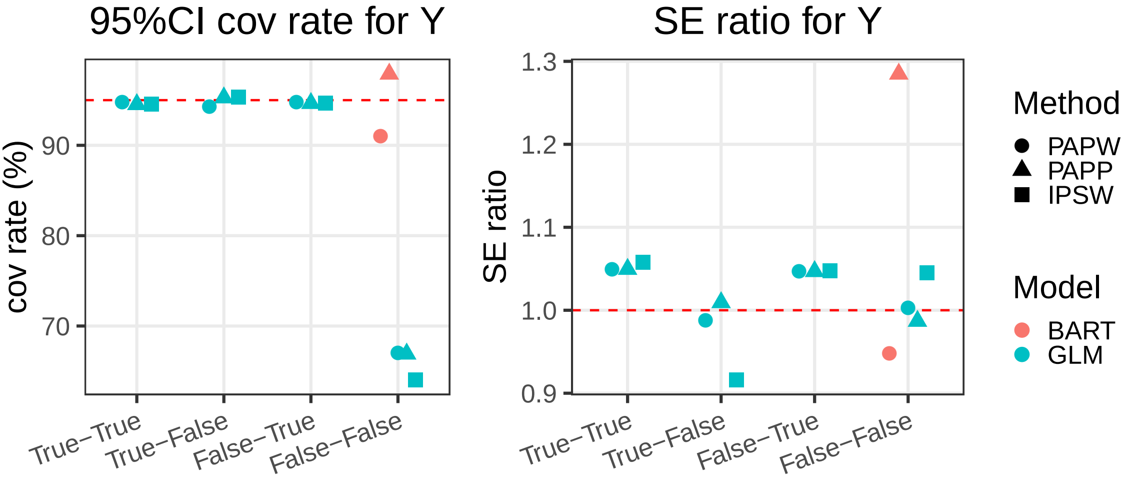


Simulation results II

- The simulation results for $n_{R\alpha} = 100$ and $n_{B\alpha} = 50$ and $a = 200$:

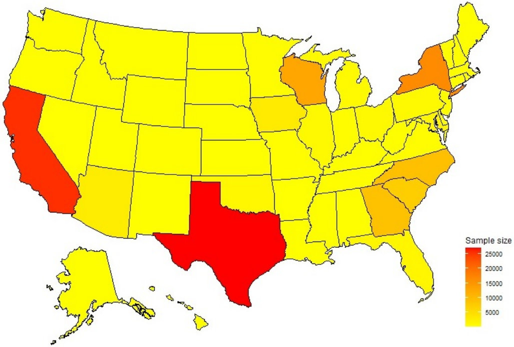


- The simulation results for $n_{R\alpha} = 100$ and $n_{B\alpha} = 50$ and $a = 200$:



Results on SHRP2: reference survey

- The 2017 **National Household Travel Survey** (NHTS) as the reference survey
- A nationally representative survey of U.S. citizens aged ≥ 5 years ($n_R = 129,112$)
- An address-based sample with a stratified design
- Initial recruitment through mailing (RR: 30.4%)
- Responded HH assigned randomly to weekdays
- Travel log using web/telephone (RR: 51.4%)
- NHTS data were combined with SHRP2 data at the day level ($n_B = 874,211$)



Results on SHRP2: data integration

- **Common variables in SHRP2 and NHTS 2017 data sets**

Individual level	Vehicle level	Trip level
gender, age, race, ethnicity, urban size, birth country, education, HH income home ownership, job status	vehicle make, vehicle type vehicle age, mileage	duration, distance, average speed, start time, weekday, month

- **Differences between SHRP2 and NHTS in sample composition**

Feature	NHTS	SHRP2
Age range	≥ 5	16-80
Transportation mode	walk, bicycle, motorbike, car, ...	car, SUV, van, light truck
Driving status	driver, passenger	driver
Vehicle ownership	owned, rental, public transportation	owned
Trip measurement	self-reported	sensor-recorded
Followup duration	one day	months or years

Results on SHRP2: data integration

- **Common variables in SHRP2 and NHTS 2017 data sets**

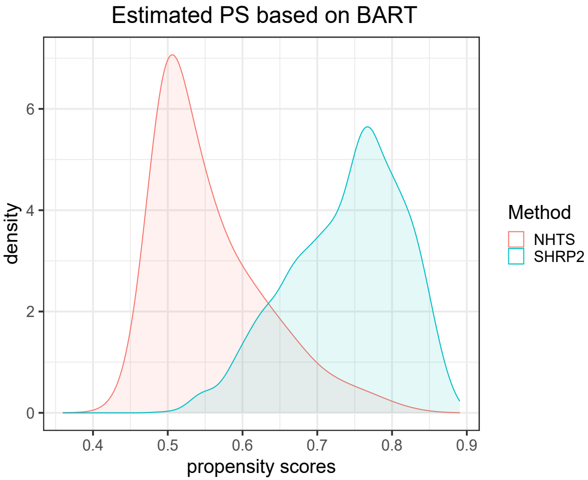
Individual level	Vehicle level	Trip level
gender, age, race, ethnicity, urban size, birth country, education, HH income home ownership, job status	vehicle make, vehicle type vehicle age, mileage	duration, distance, average speed, start time, weekday, month

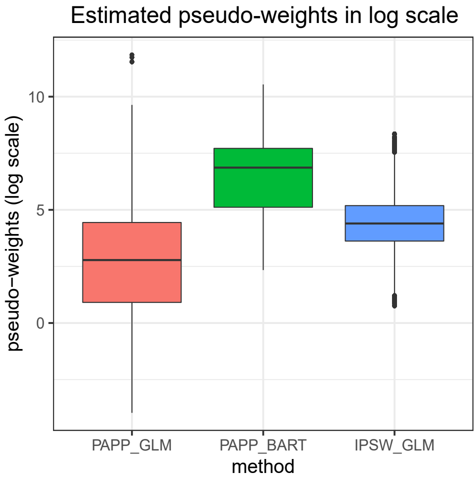
- **Differences between SHRP2 and NHTS in sample composition**

Feature	NHTS	SHRP2
Age range	≥ 5	16-80
Transportation mode	walk, bicycle, motorbike, car, ...	car, SUV, van, light truck
Driving status	driver, passenger	driver
Vehicle ownership	owned, rental, public transportation	owned
Trip measurement	self-reported	sensor-recorded
Followup duration	one day	months or years

Results on SHRP2: pseudo-weighting

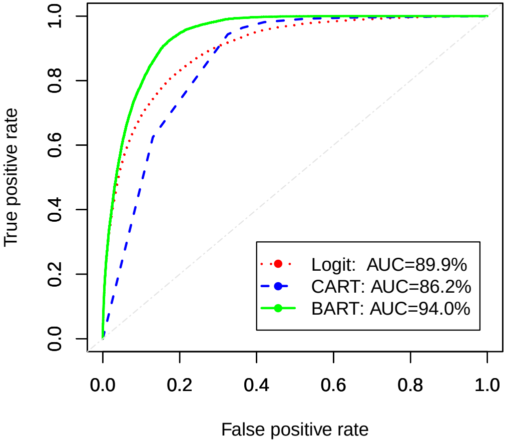
Assessing the common support of the distribution of estimated PS in SHRP2 vs NHTS

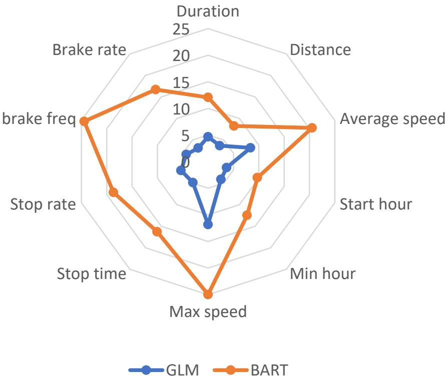




Results on SHRP2: model specification

Comparing the performance of BART with GLM in estimating PS and trip-related outcomes

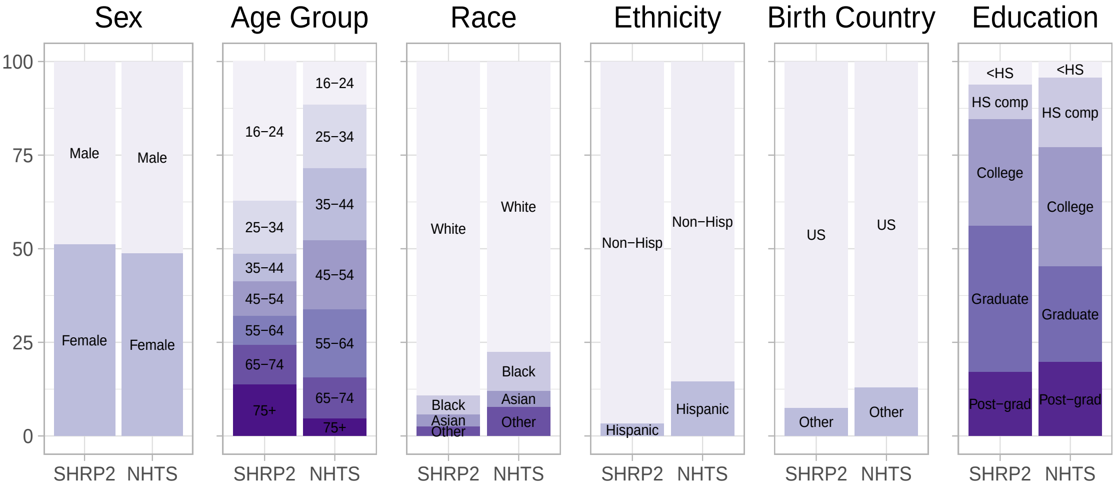




(pseudo)-R^2 in modeling trip-related outcomes

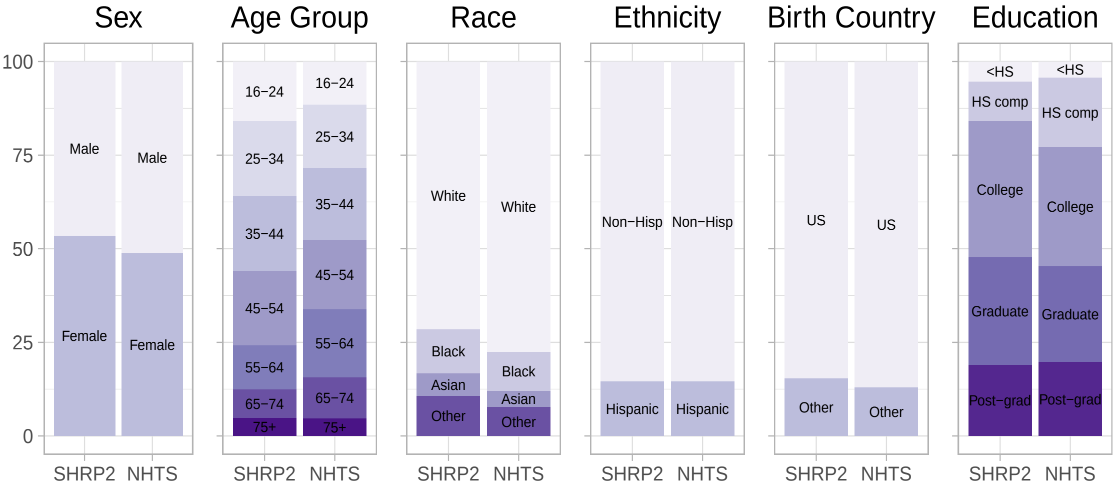
Results on SHRP2: sample composition

Comparing dist. of common covariates: unweighted SHRP2 vs weighted NHTS 2017



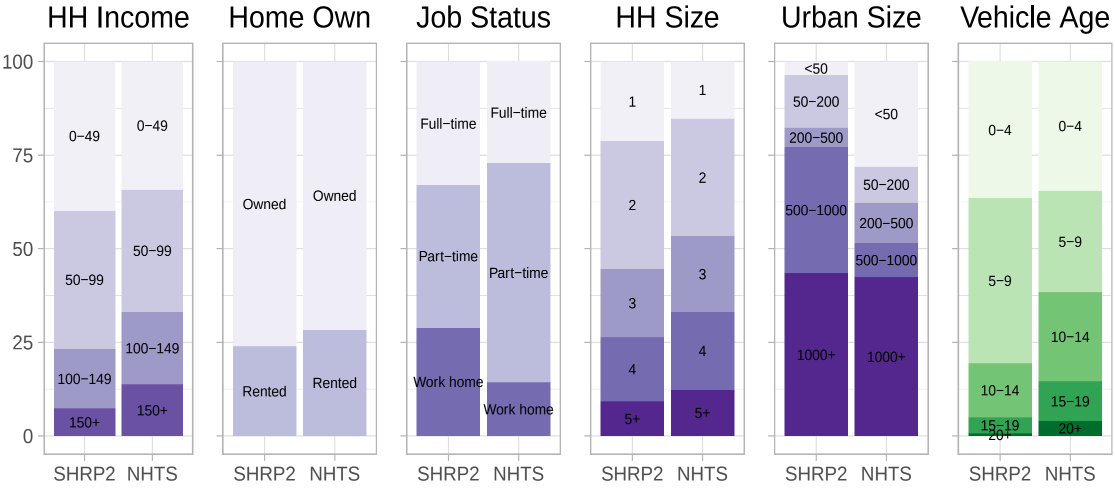
Results on SHRP2: sample composition

Comparing dist. of common covariates: pseudo-weighted SHRP2 vs weighted NHTS 2017



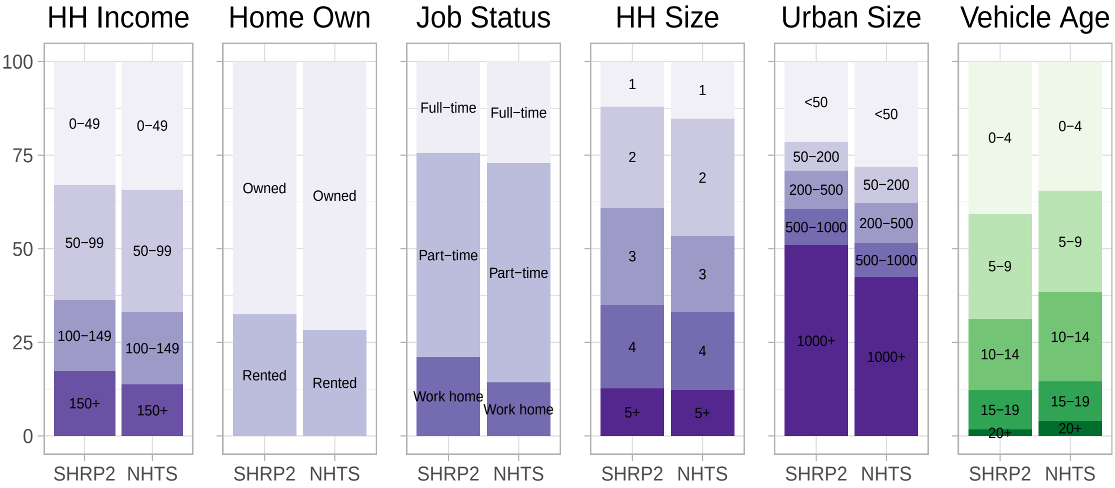
Results on SHRP2: sample composition

Comparing dist. of common covariates: unweighted SHRP2 vs weighted NHTS 2017



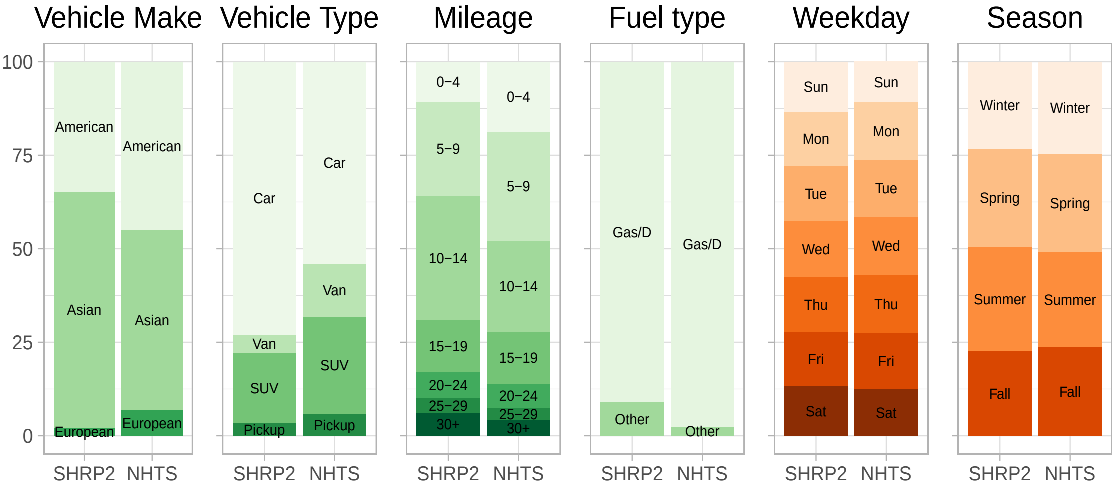
Results on SHRP2: sample composition

Comparing dist. of common covariates: pseudo-weighted SHRP2 vs weighted NHTS 2017



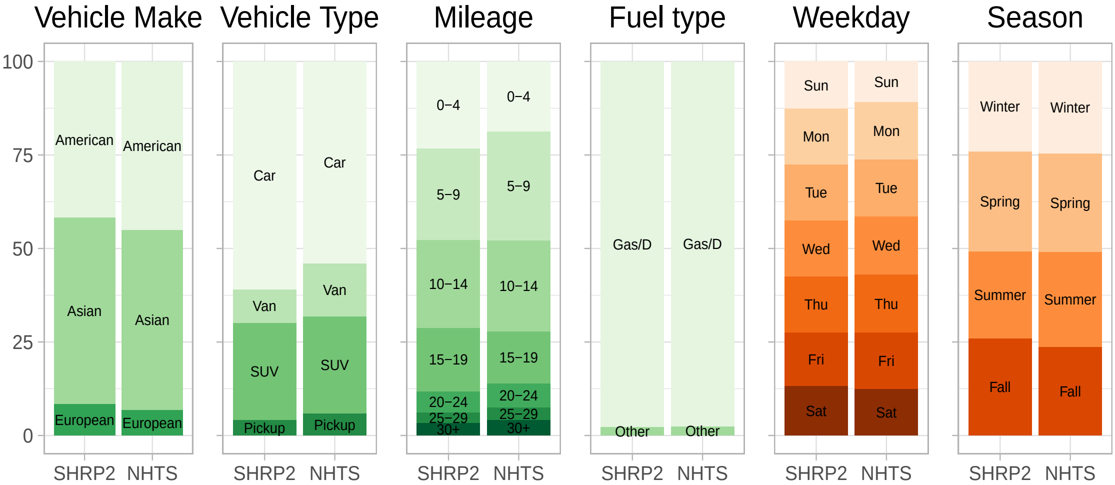
Results on SHRP2: sample composition

Comparing dist. of common covariates: unweighted SHRP2 vs weighted NHTS 2017



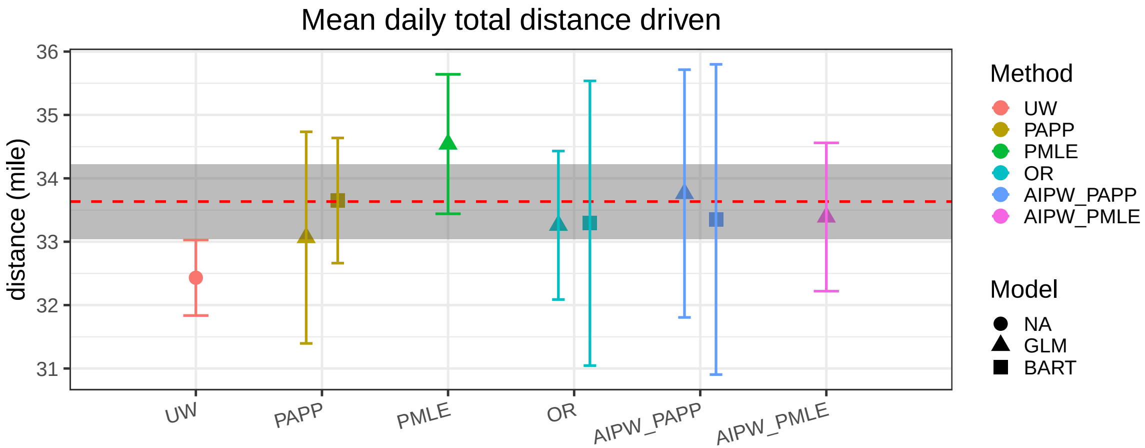
Results on SHRP2: sample composition

Comparing dist. of common covariates: pseudo-weighted SHRP2 vs weighted NHTS 2017



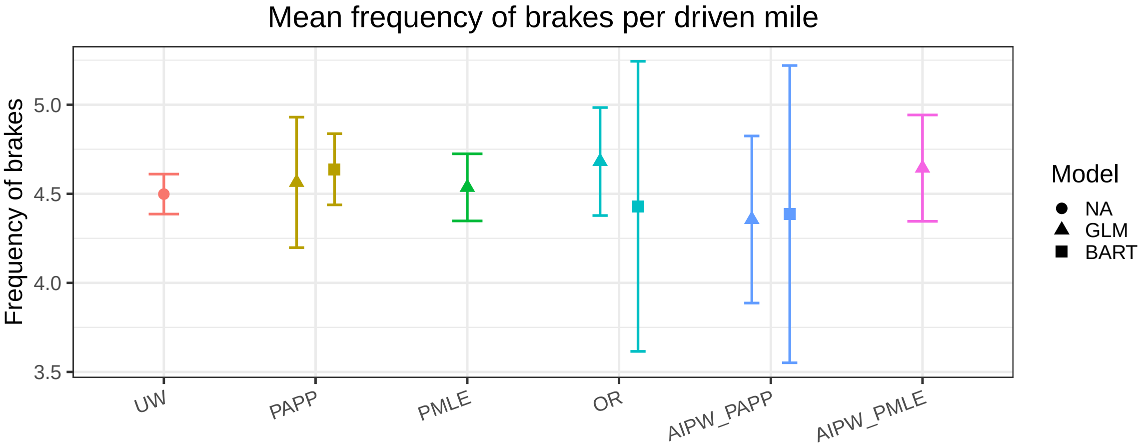
Results on SHRP2: bias adjustment

Comparing adjusted estimates of some trip-related outcome vars in SHRP2 vs NHTS



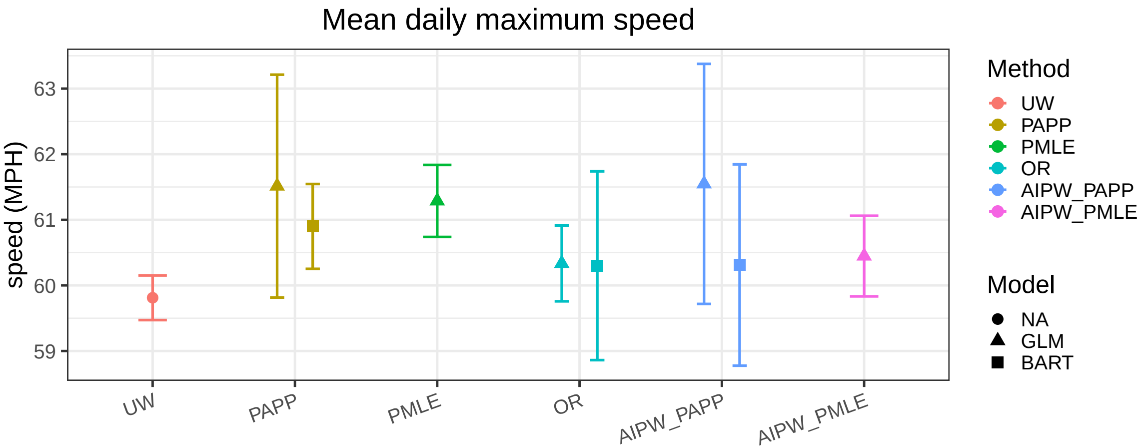
Result on SHRP2: bias adjustment

Comparing adjusted estimates of some SHRP2-specific outcome vars in SHRP2



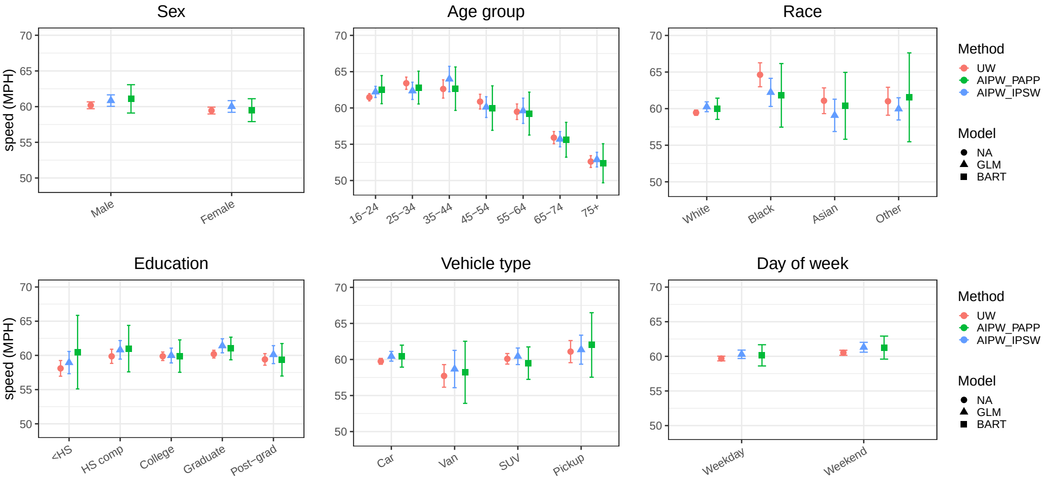
Results on SHRP2: bias adjustment

Comparing adjusted estimates of some SHRP2-specific outcome vars in SHRP2



Result on SHRP2: bias adjustment

Comparing adjusted estimates of maximum speed stratified by different factors



Discussion

- We proposed a robust method for inference in non-prob. samples.
- The AIPW method under BART produced approximately unbiased estimates, especially when both QR and PM are unknown.
- Compared to PMLE, our proposed estimator was more efficient.
- Under GLM both point and variance estimators were DR.
- The proposed asymptotic/bootstrap variance estimator performed well in simulations.
- However, the results of SHRP2 data were poor for some outcome vars.

Discussion

- **Weaknesses:**
  1. Auxiliary variables in SHRP2 were poor predictors of trip-related outcomes.
  2. Variance estimate under BART was not as accurate as alternative methods
  3. Computationally demanding, especially in high-dimensional data or when n is too large.
- **Future directions:**
  1. To develop a model-assisted method using penalized spline of propensity prediction
  2. To expand a sandwich-type variance estimator under GLM when π_i^R is unknown for $i \in B$
  3. To apply divide-and-recombine techniques to reduce the computational burden

Penalized spline propensity prediction

- Estimate π_i^R for $i \in S_B$ given x_i by modeling $E(\pi_i^R|x_i)$ if it is unknown for units of B.
- Estimate π_i^B based on $B \cup R$ using one of the methods discussed, PAPW/PAPP/IPSW.
- Predict y_i for $i \in S_R$ given $[\hat{\pi}_i^R, \hat{\pi}_i^B, x_i]$ using a penalized spline model as below:

Penalized spline model for a continuous outcome

$$y_i|x_i, \hat{\pi}_i^R, \hat{\pi}_i^B; \theta \sim N(\theta_0 + x_i^T\theta_1 + u_{i1}^T(\hat{\pi}_i^R - K_R)_+^p + u_{i2}^T(\hat{\pi}_i^B - K_B)_+^p, \tau^2)$$

where $u_{ij} \sim N(0, \sigma_j^2 I)$, a vector of q random effects and K a vector of q fixed knots.

- Use design-based methods in R to estimate the population unknown quantity:

$$\hat{\bar{y}}_{PM} = \frac{1}{N}\sum_{i=1}^{n_R}\frac{\hat{y}_i}{\pi_i^R}$$

Penalized spline propensity prediction

- Estimate π_i^R for $i \in S_B$ given x_i by modeling $E(\pi_i^R|x_i)$ if it is unknown for units of B.
- Estimate π_i^B based on $B \cup R$ using one of the methods discussed, PAPW/PAPP/IPSW.
- Predict y_i for $i \in S_R$ given $[\hat{\pi}_i^R, \hat{\pi}_i^B, x_i]$ using a penalized spline model as below:

Penalized spline model for a continuous outcome

$$y_i|x_i, \hat{\pi}_i^R, \hat{\pi}_i^B; \theta \sim N(\theta_0 + x_i^T\theta_1 + u_{i1}^T(\hat{\pi}_i^R - K_R)_+^p + u_{i2}^T(\hat{\pi}_i^B - K_B)_+^p, \tau^2)$$

where $u_{ij} \sim N(0, \sigma_j^2 I)$, a vector of q random effects and K a vector of q fixed knots.

- Use design-based methods in R to estimate the population unknown quantity:

$$\hat{\bar{y}}_{PM} = \frac{1}{N}\sum_{i=1}^{n_R}\frac{\hat{y}_i}{\pi_i^R}$$

Penalized spline propensity prediction

- Estimate π_i^R for $i \in S_B$ given x_i by modeling $E(\pi_i^R|x_i)$ if it is unknown for units of B.
- Estimate π_i^B based on $B \cup R$ using one of the methods discussed, PAPW/PAPP/IPSW.
- Predict y_i for $i \in S_R$ given $[\hat{\pi}_i^R, \hat{\pi}_i^B, x_i]$ using a penalized spline model as below:

Penalized spline model for a continuous outcome

$$y_i|x_i, \hat{\pi}_i^R, \hat{\pi}_i^B; \theta \sim N(\theta_0 + x_i^T\theta_1 + u_{i1}^T(\hat{\pi}_i^R - K_R)_+^p + u_{i2}^T(\hat{\pi}_i^B - K_B)_+^p, \tau^2)$$

where $u_{ij} \sim N(0, \sigma_j^2 I)$, a vector of q random effects and K a vector of q fixed knots.

- Use design-based methods in R to estimate the population unknown quantity:

$$\hat{\bar{y}}_{PM} = \frac{1}{N}\sum_{i=1}^{n_R}\frac{\hat{y}_i}{\pi_i^R}$$

Penalized spline propensity prediction

- Estimate π_i^R for $i \in S_B$ given x_i by modeling $E(\pi_i^R|x_i)$ if it is unknown for units of B.
- Estimate π_i^B based on $B \cup R$ using one of the methods discussed, PAPW/PAPP/IPSW.
- Predict y_i for $i \in S_R$ given $[\hat{\pi}_i^R, \hat{\pi}_i^B, x_i]$ using a penalized spline model as below:

Penalized spline model for a continuous outcome

$$y_i|x_i, \hat{\pi}_i^R, \hat{\pi}_i^B; \theta \sim N(\theta_0 + x_i^T\theta_1 + u_{i1}^T(\hat{\pi}_i^R - K_R)_+^p + u_{i2}^T(\hat{\pi}_i^B - K_B)_+^p, \tau^2)$$

where $u_{ij} \sim N(0, \sigma_j^2 I)$, a vector of q random effects and K a vector of q fixed knots.

- Use design-based methods in R to estimate the population unknown quantity:

$$\hat{\bar{y}}_{PM} = \frac{1}{N}\sum_{i=1}^{n_R}\frac{\hat{y}_i}{\pi_i^R}$$

Thanks for your attention

Email address: arafei@umich.edu

Acknowledgements:

- Professor Michael R. Elliott
- Research Professor Carol A.C. Flannagan
- Research Associate Professor Brady T. West

References

Elliott, M., Valliant, R. (2017)
Inference for nonprobability samples
Statistical Science 32(2), 249–264.

Wu, Changbao & Sitter, Randy R. (2001)
A model-calibration approach to using complete auxiliary information from survey data
Journal of the American Statistical Association 96(453), 185–193.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994)
Estimation of regression coefficients when some regressors are not always observed
Journal of the American statistical Association 89(427), 846-866.

Elliott, M., Resler, A., Flannagan, C., Rupp, J. (2010)
Appropriate analysis of CIREN data: Using NASS-CDS to reduce bias in estimation of injury risk factors in passenger vehicle crashes
Accident analysis and prevention 42(2), 530–539.

Deville, J. C., & Särndal, C. E. (1992)
Calibration estimators in survey sampling
Journal of the American statistical Association 87(418), 376-382.

References

Rafei, A., Flannagan, A. C. F., Elliott, M. R. (2020)
Big Data for Finite Population Inference: Applying Quasi-Random Approaches to Naturalistic Driving Data Using Bayesian Additive Regression Trees
Journal of Survey Statistics and Methodology 8 (1), 148-180.

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

Chen, Y., Li, P., Wu, C. (2018).
Doubly robust inference with non-probability survey samples.
Journal of American Statistical Association. 1-11.

Wang, L., Valliant, R., Li, Y (2020).
Adjusted Logistic Propensity Weighting Methods for Population Inference using Nonprobability Volunteer-Based Epidemiologic Cohorts.
arXiv preprint arXiv:2007.02476.

Lee, S. (2006).
Propensity score adjustment as a weighting scheme for volunteer panel web surveys.
Journal of official statistics, 22(2), 329