Reducing the bias of non-probability sample estimators through inverse probability weighting with an application to Statistics Canada's crowdsourcing data

Jean-François Beaumont

Collaborators: K. Bosa, A. Brennan, J. Charlebois and K. Chu

Morris Hansen Memorial Lecture, March, 1, 2022

Delivering insight through data, for a better Canada

Motivation: Crowdsourcing experiments

- In 2020, Statistics Canada advertised a series of online questionnaires on its website
  - This approach is called crowdsourcing
  - 1st "crowdsourcing" sample: 200,000 participants
- **Why use crowdsourcing?**
  - Desire to have **timely** and **inexpensive** information (e.g. pandemic)
- **Why being careful with crowdsourcing?**
  - **Participation bias** and measurement errors

Overview

- Data integration scenario
- Inverse probability weighting (Chen, Li and Wu, 2020)
- Developed two extensions that account for the data structure:
  - Variable selection procedure using a modified AIC (Akaike Information Criterion)
  - nppCART: a modified CART algorithm
- Illustration using crowdsourcing data

- **Disclaimer:** The content of this presentation represents the authors' opinions and not necessarily those of Statistics Canada.

Data integration scenario

- Estimation of the population total: $\theta = \sum_{k \in U} y_k$
- Non-probability sample: $s_{NP} \subset U$
  - **Observed**: variable of interest y_k and auxiliary variables $\mathbf{x}_k$
  - Participation indicator: δ_k
- Probability sample: $s_P \subset U$
  - **Observed**: $\mathbf{x}_k$ and a survey weight w_k
  - **Missing**: y_k and δ_k
- Assumption: No measurement errors

Inverse probability weighting

- Model the participation probability: $p_k = \Pr(\delta_k = 1 \mid \mathbf{x}_k) > 0$
- Assumption: Non-informative participation
  - $\Pr(\delta_k = 1 \mid \mathbf{x}_k, y_k) = \Pr(\delta_k = 1 \mid \mathbf{x}_k)$
  - Requires powerful auxiliary variables
- Pseudo weights: $\hat{w}_k^{NP} = \hat{p}_k^{-1}$
- Estimator of θ : $\hat{\theta}_{NP} = \sum_{k \in s_{NP}} \hat{w}_k^{NP} y_k$
- Pseudo weights can be calibrated to increase efficiency and achieve double robustness
- Alternative: Model y_k (e.g., statistical matching)

Estimation of p_k

- Logistic model: $p_k(\boldsymbol{\alpha}) = \left[1 + \exp\left(-\mathbf{x}_k'\boldsymbol{\alpha}\right)\right]^{-1}$
- $\hat{p}_k = p_k(\hat{\boldsymbol{\alpha}})$. How to estimate $\boldsymbol{\alpha}$?
- Maximum Likelihood: $\boxed{\sum_{k \in s_{NP}} \mathbf{x}_k - \sum_{k \in U} p_k(\boldsymbol{\alpha})\mathbf{x}_k = \mathbf{0}}$
    - Requires $\mathbf{x}_k$ to be available for $k \in U$
    - Similar to weighting for survey nonresponse
- Chen, Li and Wu (2020): $\boxed{\sum_{k \in s_{NP}} \mathbf{x}_k - \sum_{k \in s_P} w_k p_k(\boldsymbol{\alpha})\mathbf{x}_k = \mathbf{0}}$
    - Pseudo ML: Requires knowing $\mathbf{x}_k$ for $k \in s_{NP}$ and $k \in s_P$

Estimation of p_k

- Homogeneous group model: $p_k \equiv p_g$, $k \in U_g$
  - Special case of the logistic model
  - Using Chen, Li and Wu (2020), the estimated participation probability for unit k in group g :

$$\hat{p}_g = n_g^{NP} \Big/ \hat{N}_g$$

- In practice, homogeneous groups are often created

Estimation of p_k

- Two main reasons:
  - Robust with respect to a misspecification of the logistic model (Haziza and Lesage, 2016)
  - Avoids very small estimated probabilities
- How to form homogeneous groups?
  - First, compute $\hat{p}_k^{\text{logistic}}$ and then create groups homogeneous with respect to $\hat{p}_k^{\text{logistic}}$
  - Use classification trees

Choice of auxiliary variables / groups

- Choice of relevant auxiliary variables and interactions (or homogeneous groups) is key to reduce bias
  - Auxiliary variables are often categorical and crossing them all is usually not an option
- Standard procedures cannot be used:
  - The pooled sample is not an i.i.d. sample
  - The probability sampling design must be taken into account
- Developed two extensions of Chen, Li and Wu (2020):
  - Stepwise selection procedure using a modified AIC
  - nppCART: a modified CART (also based on modified AIC)

Modified AIC

- AIC is a likelihood-based criterion used to select a model
- $\mathrm{AIC} = -2l(\hat{\boldsymbol{\alpha}}) + 2q$ Assumption: ML estimation ($s_P = U$)
  - $l(\hat{\boldsymbol{\alpha}})$: Log likelihood
- Borrow from Lumley and Scott (2015): modified the classical AIC when pseudo maximum likelihood is used to estimate model parameters from survey data

$$\boxed{\text{modified AIC} = -2\hat{l}(\hat{\boldsymbol{\alpha}}) + 2q + (\text{penalty for using } s_P \text{ instead of } U)}$$

  - $\hat{l}(\hat{\boldsymbol{\alpha}})$: Pseudo log likelihood

nppCART: a modified CART

- CART creates homogeneous groups (Breiman et al., 1984)
  - Implicitly and automatically select relevant auxiliary variables and interactions
  - Does not account for the data structure and probability sampling design
- Growing step:
  - CART: Recursively split the sample by minimizing an objective function
    - Entropy distance $\propto$ - (log likelihood for homog. group model)
  - nppCART: Replace log likelihood by pseudo log likelihood (as in Chen, Li and Wu, 2020)

nppCART: an modified CART

- Pruning step:
  - Determine a sequence of subtrees of decreasing size
  - Choose the best subtree: nppCART minimizes the modified AIC for the homogeneous group model
- nppCART accounts for the probability sampling design in both steps
- May be used to create homogeneous groups based on
  - All the auxiliary variables
  - Only one variable: $\hat{p}_k^{\text{logistic}}$

Bootstrap variance estimation

- Need to account for two sources of variability: probability sampling design and participation model
- Two sets of bootstrap weights:
  - A set of bootstrap weights that accounts for the probability sampling design (e.g., Rao, Wu and Yue, 1992)
  - A set of bootstrap weights obtained by modelling the participation mechanism as Poisson sampling (Beaumont and Patak, 2012)
- Simplification: Treat homogeneous groups as fixed

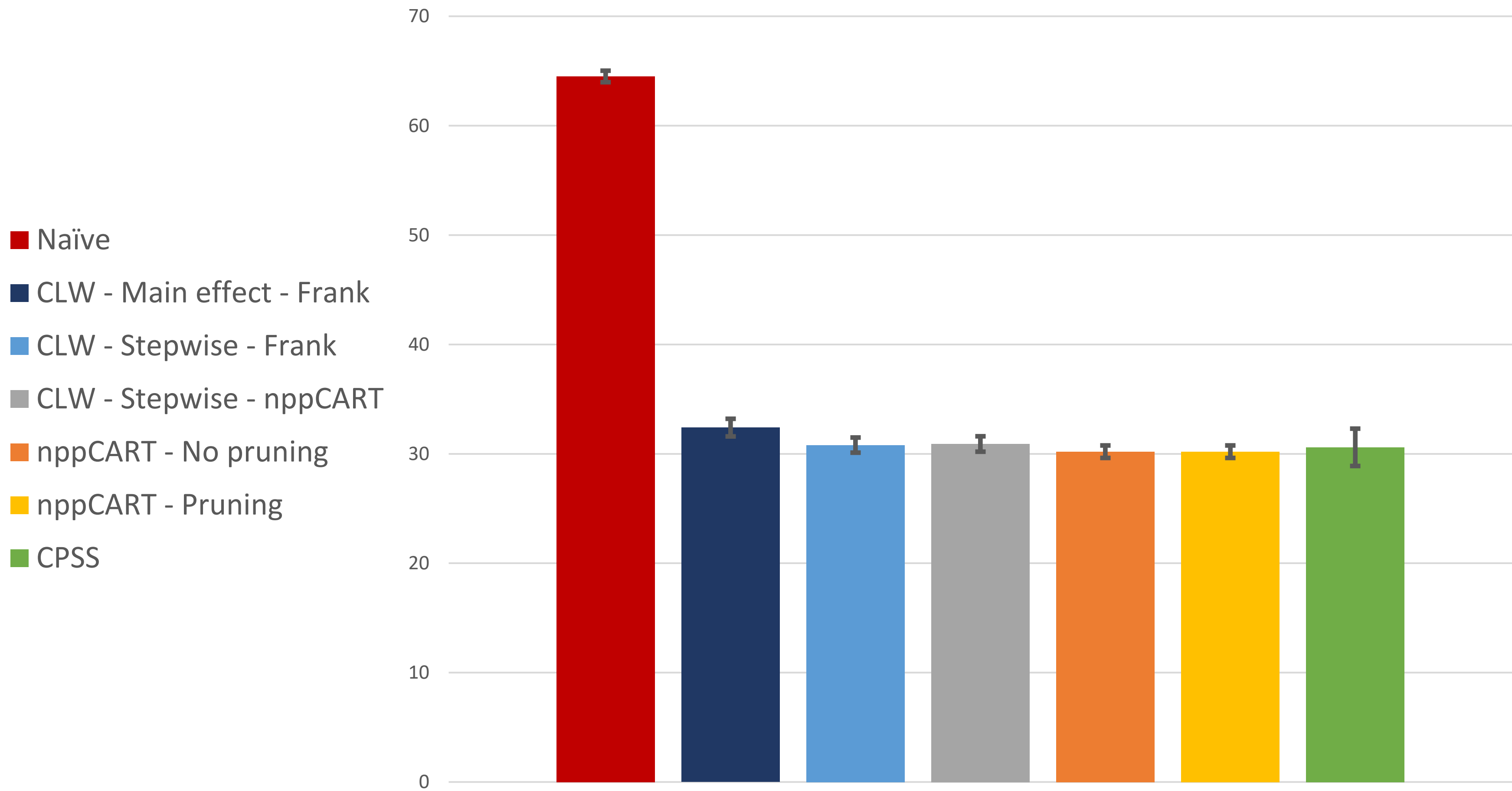
Illustration

- Non-probability sample: Crowdsourcing (31,415 participants)
- Probability sample: LFS (87,779 respondents + response rate around 80%)
- Auxiliary variables: education (8), region (56), age (13), sex (2), immigration (3), employment (3), marital (6), household size (6)
- Reference for comparison: CPSS (probability sample with 4,209 respondents and response rate around 15%)

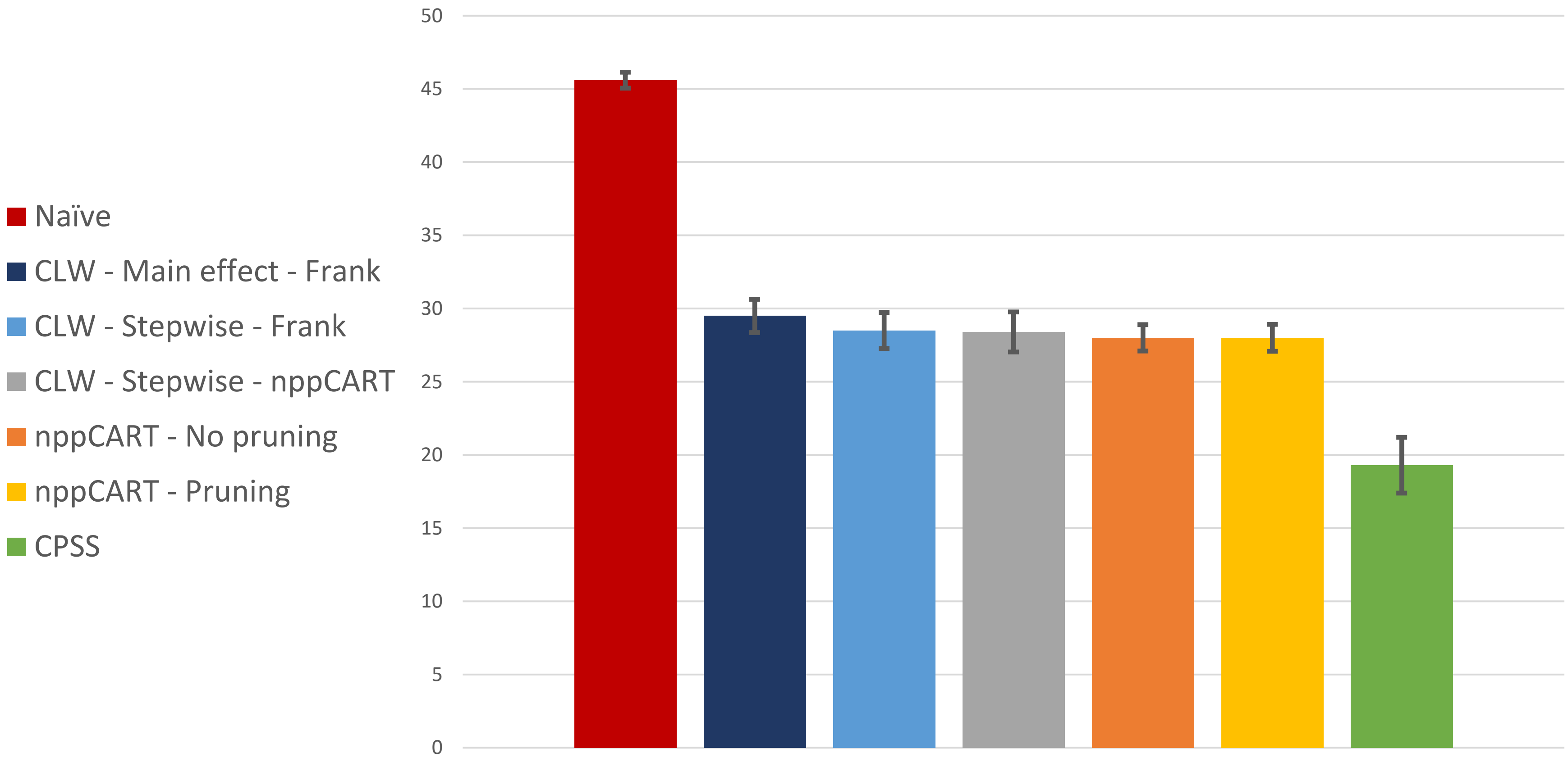
Methods

- Naïve (1 group)
- CLW – Main effect – Frank (100 groups)
- CLW – Stepwise – Frank (100 groups)
- CLW – Stepwise – nppCART (1,276 groups)
- nppCART – No pruning (3,165 groups)
- nppCART – Pruning (1,772 groups)

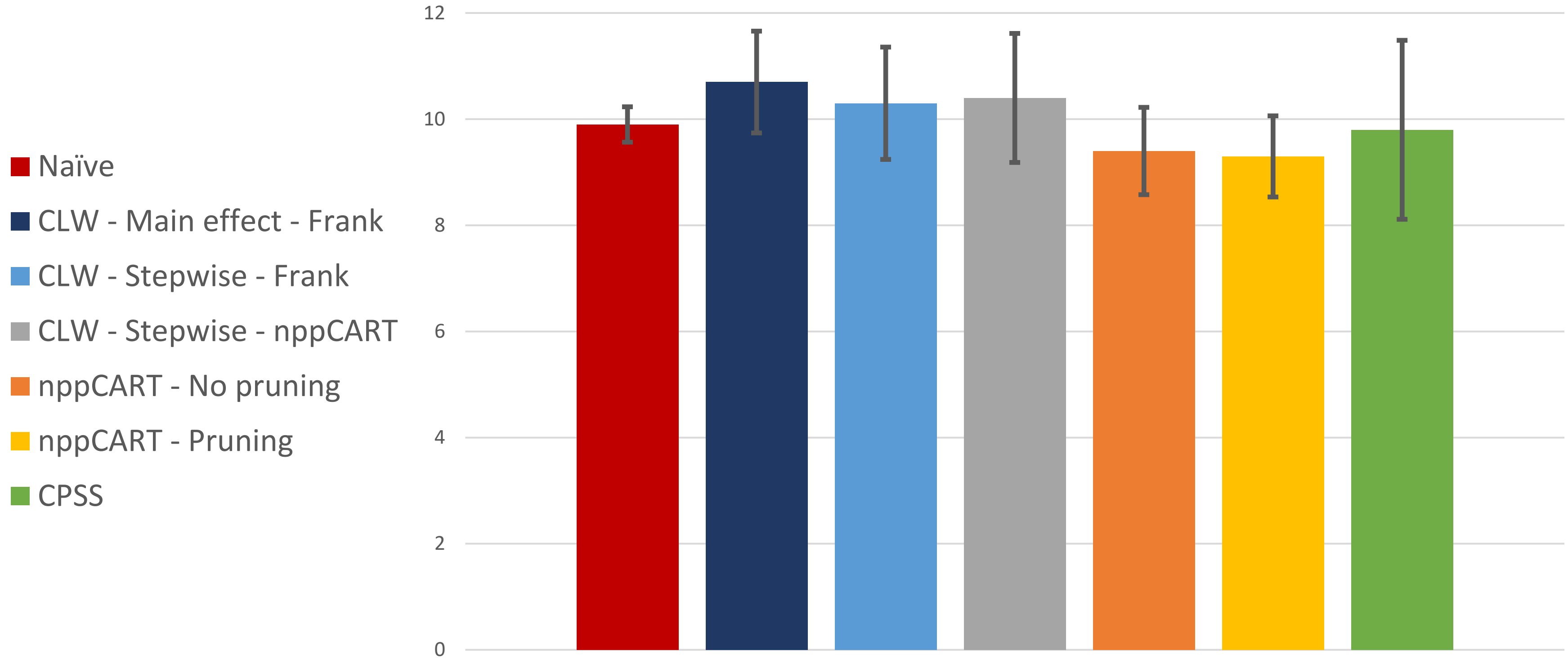
Proportion of people having a university degree



Proportion of people who worked most of their hours at home during the reference week

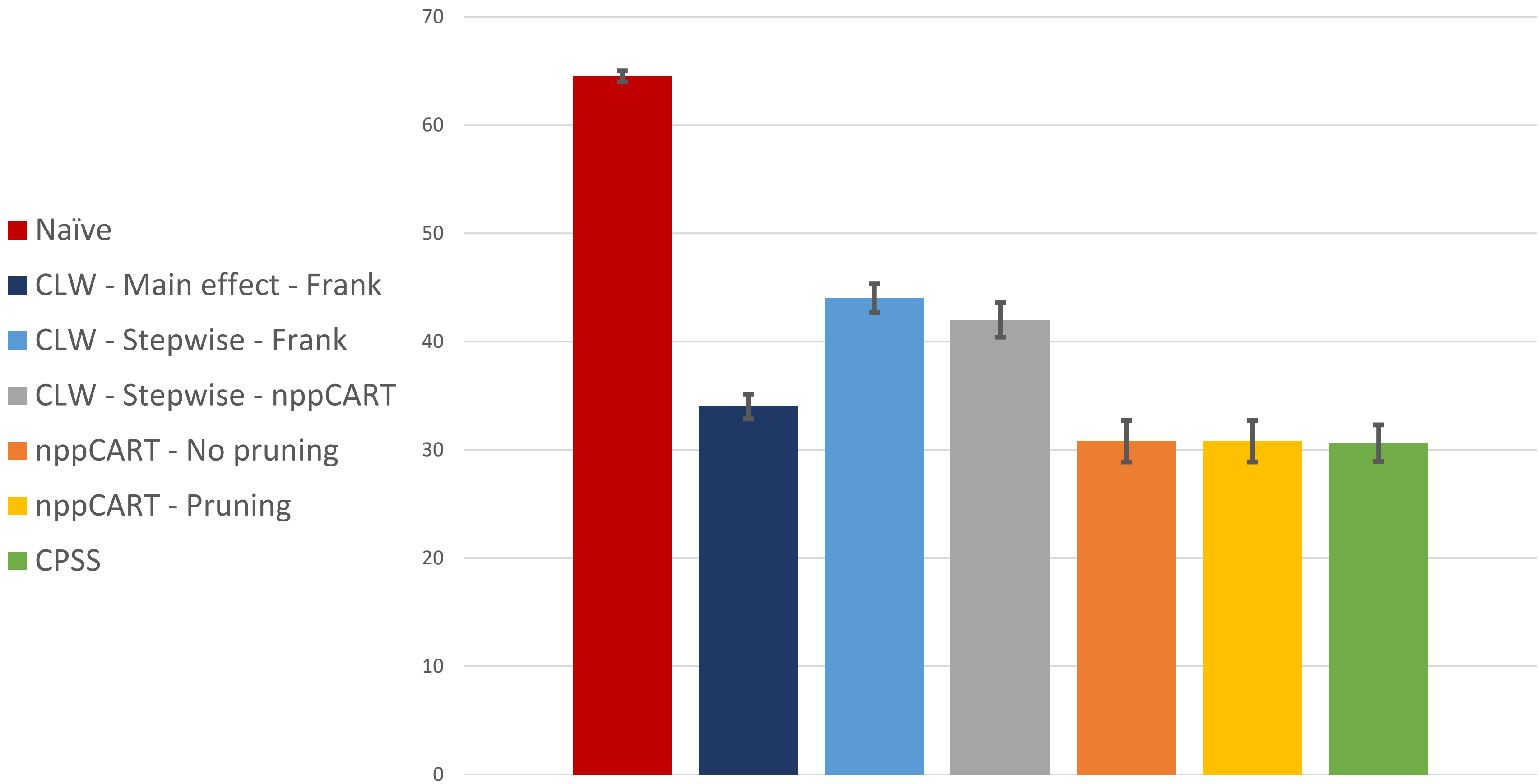


Proportion of people who "fear being a target for putting others at risk" because they do not always wear a mask in public



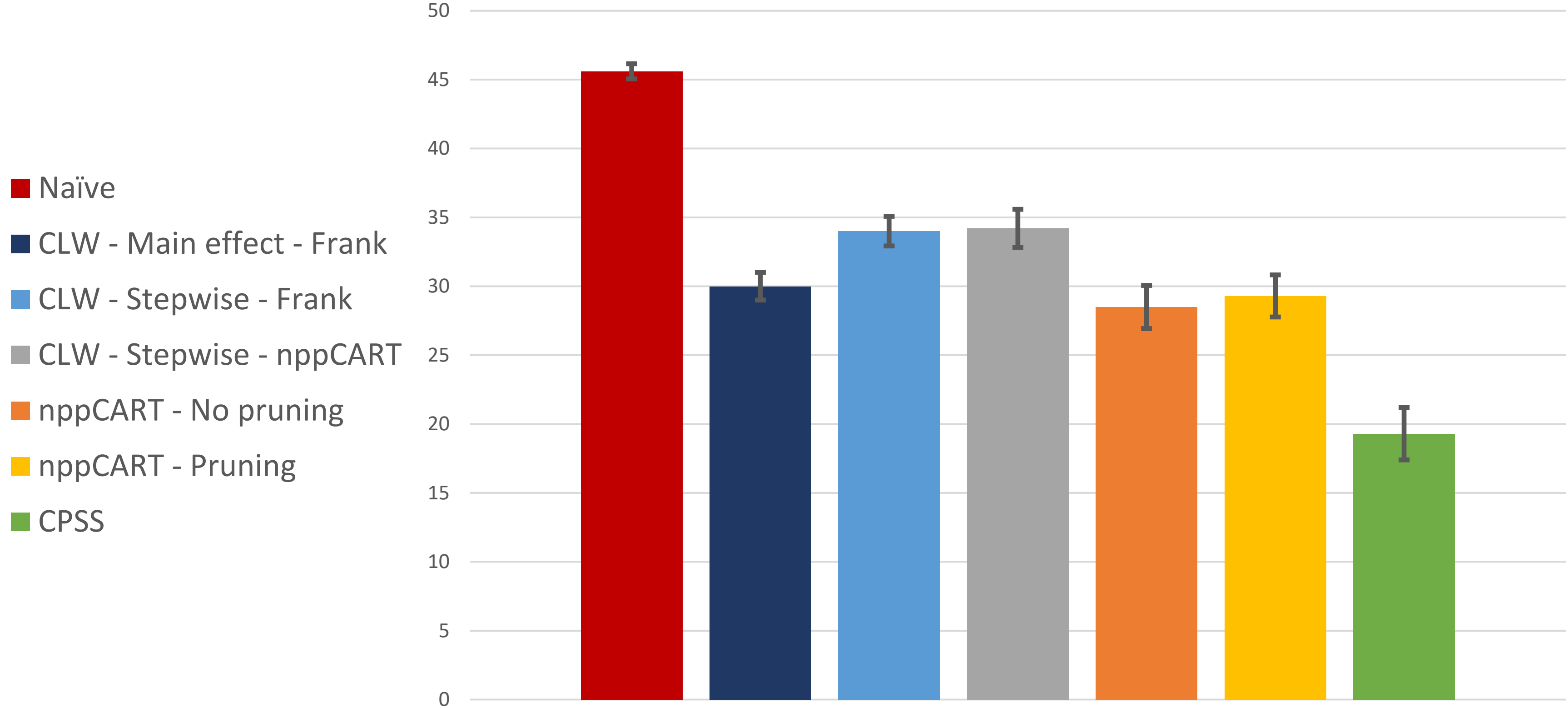
Probability sample: CPSS

Proportion of people having a university degree



Probability sample: CPSS

Proportion of people who worked most of their hours at home during the reference week



Some conclusion of our experimentations

- The variable Education is by far the most important to explain participation
- Interactions are not strong
- All IPW methods performed similarly, especially when the probability sample was large
- Larger differences may be expected
  - for smaller domains
  - for other data sets (with stronger interactions)
- Future work: Random forests?