Combining Data from Probability and Non-probability Surveys

Michael Elliott[1,2]

[1]Department of Biostatistics, University of Michigan
[2]Survey Methodology Program, Institute for Social Research

Motivation for Utilizing Non-Probability Samples

- Non-probability samples are an increasing part of life for the survey analyst.
  - Non-response.
  - Sampling frame coverage.
  - Increasing cost.
  - Detailed outcomes of interest not present in probability samples.
  - Larger sample size than equivalent probability sample, especially in small domains.
- Offers possibility of improved inference if increase in precision is not overwhelmed by bias from the non-probability sample.

Framework for Nonprobability Sample Inference

Consider the joint density of a population vector of analysis variable $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$ and of 0-1 indicator variables $\delta_{\mathbf{s}} = (\delta_1, \delta_2, \ldots, \delta_N)$ for a sample s:

$$f(\mathbf{Y}, \delta_{\mathbf{s}} | \mathbf{X}; \Theta, \Phi) = f(\mathbf{Y} | \mathbf{X}; \Theta) f(\delta_{\mathbf{s}} | \mathbf{Y}, \mathbf{X}; \Phi)$$

where $\mathbf{X}$ is an $N \times p$ matrix of covariates that govern $\mathbf{Y}$ through unknown parameter Θ, and unknown parameter Φ governs $f(\delta_{\mathbf{s}}$ through both $\mathbf{Y}$ and $\mathbf{X}$ (Smith 1983; Rubin 1976; Little 1982).

- Probability sampling: $f(\delta_{\mathbf{s}} | \mathbf{Y}, \mathbf{X}; \Phi) = f(\delta_{\mathbf{s}} | \mathbf{X})$.
- Non-probability sampling: $\delta_{\mathbf{s}}$ can depend on $\mathbf{Y}$ and/or Φ in addition to $\mathbf{X}$.

Framework for Nonprobability Sample Inference

1. Quasi-randomization: model $f(\delta_{\mathbf{s}}|\mathbf{Y}, \mathbf{X}; \Phi)$.
   - Ideally, the probability of being in the sample is not NMAR and a model can be found for $f(\delta_{\mathbf{s}}|\mathbf{X}; \Phi)$.
2. Superpopulation: model $f(\mathbf{Y}|\mathbf{X}; \Theta)$.
   - Calibration a broad special case where model-based estimates are adjusted to known or estimated quantities outside of the non-probability sample.

Quasi-randomization

- Estimation of the most general NMAR model $f(\delta_{\mathbf{s}}|\mathbf{Y},\mathbf{X},\Phi)$ typically requires information on non-sampled units that is available only in specialized applications.
    - Typically assume MAR $f(\delta_{\mathbf{s}}|\mathbf{X},\Phi)$.
- Even here, estimation typically requires some heroic assumptions – unless there is a “reference” probability survey available.

Quasi-randomization: Generating Pseudo-Weights

- Elliott and Davis (2005) developed method to account for non-response bias and frame coverage.
  - Extend to estimate over- and under-representation of sample elements in the non-probability sample based on covariates available in both samples.
- By repeated application of Bayes' Rule and discriminant analysis we can approximate when sampling fractions are small the probability that a nonprobability case would have been sampled by

$$P(S_i^* = 1 \mid \mathbf{x}_i = \mathbf{x}_o) \propto P(S_i = 1 \mid \mathbf{x}_i = \mathbf{x}_o) \frac{P(Z_i = 1 \mid \mathbf{x}_i = \mathbf{x}_o)}{P(Z_i = 0 \mid \mathbf{x}_i = \mathbf{x}_o)}.$$

  - S^* = sampling indicator for being in the nonprobability sample.
  - S = indicator for being in the probability sample.
  - $\mathbf{x}_i$ = covariates that determine probability of selection.

Generating Pseudo-Weights

- Resulting pseudo-weight is given by

$$w_i = 1/\hat{P}(S_i^* = 1 \mid \mathbf{x}_i = \mathbf{x}_o) \propto$$

$$1/\hat{P}(S_i = 1 \mid \mathbf{x}_i = \mathbf{x}_o)\frac{\hat{P}(Z_i = 0 \mid \mathbf{x}_i = \mathbf{x}_o)}{\hat{P}(Z_i = 1 \mid \mathbf{x}_i = \mathbf{x}_o)}.$$

- If the probability sample weight as a function of $\mathbf{x}_o$ is known, $1/\hat{P}(S_i = 1 \mid \mathbf{x}_i = \mathbf{x}_o)$ can be replaced with $1/P(S_i = 1 \mid \mathbf{x}_i = \mathbf{x}_o)$ and computed directly.
  - Otherwise $\hat{P}(S_i = 1 \mid \mathbf{x}_i = \mathbf{x}_o)$ can be estimated using, e.g., beta regression (Ferrari and Cribari 2004).
- Obtain $\hat{P}(Z_i = z \mid \mathbf{x}_i = \mathbf{x}_o)$ via logistic regression.
  - LASSO (Tibshirani 1996).
  - Bayesian additive regression trees (Chipman et al. 2010).
  - Super learner algorithms (Van der Laan et al. 2007).

Inference Using Pseudo-Weights

- For point estimation, use normalized pseudo-weights and probability sample weights as case weights in combined dataset to obtain the estimator of interest $\hat{\theta}$.
- For variance estimation, use a jackknife estimators that treats the non-probability sample as a single stratum with IID observations and the probability sample following the appropriate sample design.

Quasi-randomization Example: CIREN and NASS-CDS (Elliott et al. 2010)

- Crash Injury Research Engineering Network (CIREN) database contains detailed medical and crash information on motor vehicle crash patients admitted to Level 1 trauma centers around the US.
- Non-probability sample: Centers compete to get grants (only Level 1 eligible).
- Inclusion criteria: model year, injury severity, crash type, and occupant restraint condition.
- Extensive medical and biomechanical information about each occupant and crash.
  - Careful case-by-case assessment of injury-causation scenarios.
- Use CIREN data from 2000-2006, and restricted to 1,393 occupants 16 and older that actually met specific criteria for CIREN inclusion.

Quasi-randomization Example: CIREN and NASS-CDS

- National Automotive Sampling Survey – Crashworthiness Data System (NASS-CDS) (NHTSA, 2008) is a representative three-stage probability sample selected annually from all police-reported crashes that resulted in at least one vehicle having to be towed from the scene for damage.
- Oversamples crashes:
  - fatal/serious injuries.
  - transported to ER/hospital.
- A subset of 4,099 NASS-CDS 2000-2006 subjects eligible for inclusion in CIREN based on their injury outcomes was used to create the CIREN pseudo-weights.
- Limitations of NASS-CDS
  - Detail of injury type (e.g., know had pelvic fracture, but type is unknown).
  - Sample size of severe (AIS 3+) injuries somewhat limited.

Constructing Pseudo-Weights

- Predict NASS weights using injury severity, medical treatment, model year of vehicle, deformation location, light condition, year of interview, and vehicle make.
- Balance based on age, gender, restraint use, type of crash, damage distribution and extent, days hospitalized, driver (vs. passenger), injury severity, model year.

	CIREN unweighted	CIREN pseudo-weighted	NASS (CIREN-elig)
Age (yr)	41.4	42.5	41.8
Days Hosp.	10.3	6.3	6.0
Mean AIS	3.45	3.35	3.31
% 35+ kph	61.6	48.1	50.3
% Driver	50.8	81.3	78.5
$>$ 4 years	30.2	47.4	45.5
% American	57.6	60.8	66.6
% Daylight	55.1	58.4	50.6

NASS-CIREN Analysis: Predictors of Lower Extremity Injury

- Outcome available in both NASS and CIREN: AIS 3+ lower-extremity injury.
  - Increase the size of an injured sample to better estimate the effects of predictors.
- Three analyses: use data from NASS-CDS alone, use data from CIREN alone, and and use data from NASS and CIREN combined using CIREN-pseudo weights.
- A total of 884 lower extremity injuries were available in the NASS-CDS dataset; an additional 387 lower extremity injuries were available in the CIREN dataset.
- Restrict to frontal crashes.

NASS-CIREN Analysis: Predictors of Lower Extremity Injury

Odds ratio AIS 3+ of lower-extremity injury.

	NASS only	CIREN (unweighted)	NASS-CIREN
Age (vs. 65)			
16-19	**.10** $_{.05,.22}$	1.23 $_{.64,2.37}$	**.12** $_{.06,.24}$
20-39	**.22** $_{.11,.46}$	1.16 $_{.72,1.87}$	**.22** $_{.11,.76}$
40-64	**.30** $_{.15,.59}$	1.17 $_{.72,1.90}$	**.25** $_{.13,.45}$
Seat Position (vs. Driver)			
Front Row	1.14 $_{.70,1.85}$	.76 $_{.55,1.04}$	.95 $_{.63,1.44}$
Rear Row	.11 $_{.01,1.13}$	**.22** $_{.06,.83}$	**.08** $_{.02,.34}$
Restraint Use (vs. Belted)			
Belted, not 3pt	1.74 $_{.31,9.55}$	**5.00** $_{1.04,24.07}$	**8.62** $_{1.43,51.85}$
Unrestrained	**3.50** $_{1.89,6.47}$	1.24 $_{.89,1.72}$	**3.79** $_{2.48,5.78}$
Delta-V (vs. $<$15 kph)			
15-35 kph	6.81 $_{.53,87.69}$	1.62 $_{.26,10.33}$	**7.64** $_{1.05,55.61}$
35+ kph	**51.7** $_{3.71,720.8}$	2.32 $_{.37,14.55}$	**66.6** $_{7.93,559.0}$
Model Year (vs. $<$1998)			
1998-2002	**1.84** $_{1.36,2.49}$	1.31 $_{.90,1.90}$	**1.63** $_{1.12,2.35}$
2003+	**1.74** $_{1.20,2.55}$	1.06 $_{.61,1.82}$	1.84 $_{.64,5.28}$
Vehicle (vs. cars)			
Pickups	1.10 $_{.52,2.32}$	1.51 $_{.96,2.39}$	1.05 $_{.65,1.70}$
Vans	.76 $_{.45,1.27}$	**3.13** $_{1.55,6.36}$	1.00 $_{.59,1.70}$
SUVs	.93 $_{.40,2.15}$	1.30 $_{.79,2.13}$	.88 $_{.52,1.50}$

Superpopulation

- Focus on modeling of of $f(\mathbf{Y}|\mathbf{X};\Theta)$.
  - Project results from model to the full population if **X** known.
- Sample selection ignorable if design is ignorable: $f(\delta_{\mathbf{s}}|\mathbf{Y},\mathbf{X};\Phi) = f(\delta_{\mathbf{s}}|\mathbf{X};\Phi)$.
- But that is again typically not the case in non-probability samples.
- Partition **Y** into sample and non-sample units: $f(\mathbf{Y}|\mathbf{X};\Theta) = f(\mathbf{Y}_s|\mathbf{Y}_{\bar{s}},\mathbf{X};\Theta)f(\mathbf{Y}_{\bar{s}}|\mathbf{X};\Theta)$.
- If $f(\mathbf{Y}_s|\mathbf{Y}_{\bar{s}},\mathbf{X};\Theta) = f(\mathbf{Y}_s|\mathbf{X};\Theta)$ then model estimates from sample can be use to predict non-sampled elements.

Poststratification and Generalized Regression Estimation

- Suppose Y_i is linear in $\mathbf{X}_i$:

$$E_M(y_i) = \mathbf{x}_i^T \beta$$

  for unknown parameter β.
- Solving estimating equation for β using sample data yields least squares estimator

$$\hat{\beta} = (\mathbf{X}_s^T \mathbf{X}_s)^{-1} \mathbf{X}_s^T \mathbf{y}_s.$$

- Predict nonsampled units by $\hat{y}_i = \mathbf{x}_i^T \hat{\beta}$. A predictor of the population total t is then given by

$$\hat{t} = \sum_{i \in s} y_i + \sum_{i \in \bar{s}} \hat{y}_i = \sum_{i \in s} y_i + (\mathbf{t}_{Ux} - \mathbf{t}_{sx})^T \hat{\beta}$$

  where $\mathbf{t}_{Ux}$ corresponds to population totals for $\mathbf{X}$.

PS and GREG: Variance estimation

- Can rewrite $\hat{t}$ as $\sum_{i \in s} w_i y_i$ where

$$w_i = 1 + (\mathbf{t}_{Ux} - \mathbf{t}_{sx})^T (\mathbf{X}_s^T \mathbf{X}_s)^{-1} \mathbf{x}_i.$$

  - Corresponds to the generalized regression estimator (GREG) (Deville and Sarndal 1992).
  - If $\mathbf{X}$ is categorical, $\hat{t}$ corresponds to the poststratified estimator: $\hat{t}^{PS} = \sum_{h=1}^{H} N_h \overline{y}_{sh}$.
- In many cases, however, the availability of control totals may be somewhat or very limited, especially to allow the critical assumption $f(\mathbf{Y}_s | \mathbf{Y}_{\overline{s}}, \mathbf{X}; \Theta) = f(\mathbf{Y}_s |, \mathbf{X}; \Theta)$ to be made.
- In this case, replace $\mathbf{t}_{Ux}$ with $\mathbf{t}_{Bx}$, where $\mathbf{t}_{Bx}$ is obtained from a "benchmark" probability survey (Dever and Valliant 2016).

Model Assisted Calibration

- The weights w_i in GREG can be viewed as the weights that minimize $\sum_{i\in s}(w_i - 1)^2$ subject to the constraint that $\sum_{i\in s} w_i \mathbf{x}_i = \mathbf{t}_{Ux}$ or $\sum_{i\in s} w_i \mathbf{x}_i = \mathbf{t}_{Bx}$.
- Model assisted calibration (Wu and Sitter 2001) replaces the latter constraint with $\sum_{i\in s} w_i \hat{y}_i = \sum_{i\in U} \hat{y}_i$.
- This yields

$$\hat{t}^{MA} = \sum_{i\in s} y_i + (\sum_{i\in U} \hat{y}_i - N/n_s \sum_{i\in s} \hat{y}_i)\hat{\beta}^{MC}$$

where $\hat{\beta}^{MC} = \frac{(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{(\hat{y}_i - \bar{\hat{y}})^2}$.

Estimated control LASSO calibration

- In many cases we may want to use a large vector of potential control totals, particulary if we are obtaining them from a benchmark probability survey.
- In this case, rather that obtaining $\hat{\beta}$ by least squares, use adaptive LASSO a more robust estimation procedure (Chen et al. 2018).

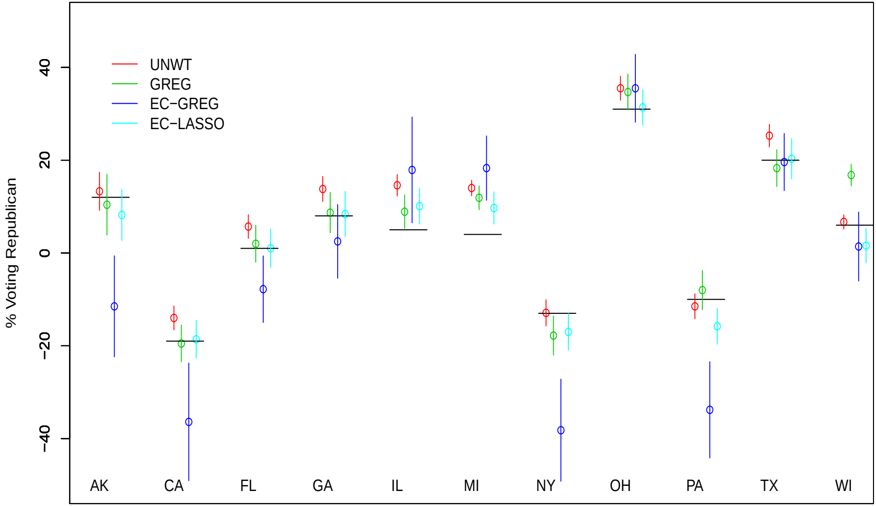
$$\hat{\beta} = \underset{\beta}{argmin} \left(\sum_{i \in s_A} \left(y_i - \mathbf{x}_i^T \beta \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \left| \hat{\beta}_j^{MLE} \right|^{-\gamma} \right).$$

- Drives parameters associated with weak predictors to 0 by penalizing covariates with large effect sizes in favor of lowering prediction error when the sample size is small (Zou 2006).

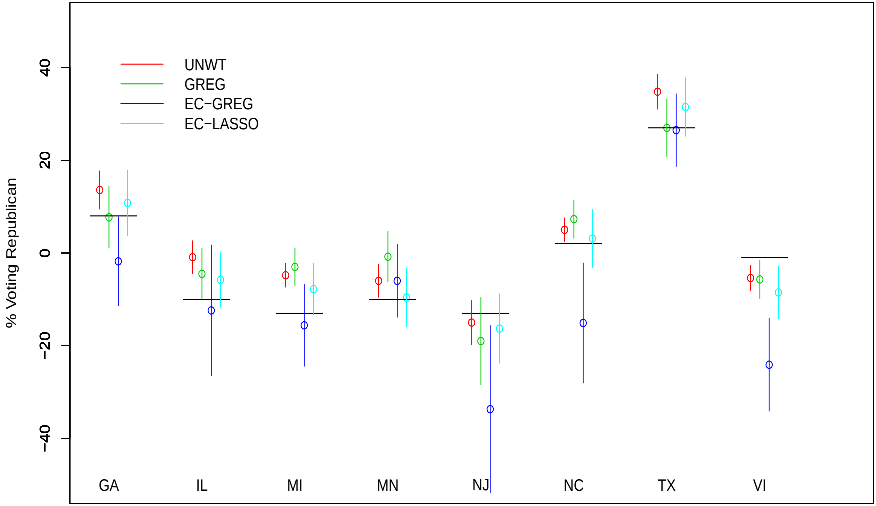
Predicting 2014 Senate and Governors Races

- Users who completed a SurveyMonkey poll in October 2014 were sometimes asked voting preferences in Senate and governor races.
- Restricted to likely voters with a Democratic or Republican candidate: 33,199 gubernatorial voters and 28,686 Senatorial voters.
- Benchmark sample: Pew Research probability sample of likely voters 1,094 gubernatorial voters and 656 Senatorial voters.
    - Common covariates: age, gender, race, education, religion, religious attendance, approval of Obama, party preference.
- Consider
    - Unadjusted.
    - Calibrated to state-level measures from probability survey.
    - Model assisted-calibration using GREG.
    - Model assisted-calibration using LASSO.

Results for Governors Races



Results for Senate Races



Results: Bias, RMSE, and Coverage

Method	Mean Bias	Mean RMSE	80% Coverage
		Governor	
Unweighted	+4.1	5.2	36% (4/11)
GREG	+1.9	5.2	64% (7/11)
EC-GREG	-7.0	15.0	36% (4/11)
EC-LASSO	-0.5	4.7	64% (7/11)
		Senate	
Unweighted	+4.0	6.0	12% (1/8)
GREG	+2.4	6.4	38% (3/8)
EC-GREG	-9.0	12.2	50% (4/8)
EC-LASSO	+1.0	5.1	50% (4/8)

Hierarchical Models

- Returning back to our poststratified estimator

$$\hat{t}^{PS} = \sum_{h=1}^{H} N_h \bar{y}_{sh} \text{ or } \hat{\bar{Y}}^{PS} = \sum_{h=1}^{H} P_h \bar{y}_{sh}$$

- Holt and Smith (1979) suggested dealing with instabilities in the estimation of $\bar{y}_{sh}$ by use of a hierachical model

$$\bar{y}_{sh} \mid \mu_h \sim N(\mu_h, \sigma^2/n_h),\ \mu_h \sim N(\mu, \tau^2).$$

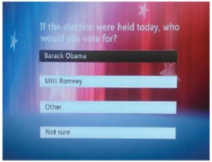
  - The mean estimator is given by $\sum_{h=1}^{H} P_h \hat{\mu}_h$, where

$$\hat{\mu}_h = E(\mu_h \mid y) = \frac{\tau^2}{\sigma^2/n_h + \tau^2}\bar{y}_h + \frac{\sigma^2/n_h}{\sigma^2/n_h + \tau^2}\bar{y}.$$

- Elliott and Little (2000): exchangeable priors oversmooth when σ^2 and τ^2 were approximately equal.
  - More structured priors (autoregressive or spline on ordered weights) had much better performance with respect to coverage and mean square error.

"Mr. P" and the 2012 Presidential Election

- Wang et al. (2015) used this hierarchical model approach, termed multilevel regression and prediction (MRP) to obtain estimates of voting behavior in the 2012 US Presidential election.
  - Sample of 350,000 Xbox users, empaneled 45 days prior to the election.
- Used detailed highly predictive covariates about voting behavior:
  - Sex, race, age, education, state, party ID, political ideology, and reported 2008 vote.
  - 176,256 cells.



"Mr. P" and the 2012 Presidential Election

- Use factorized model to predict proportion of two-party vote:

$$logit(P(Y_i \in [\text{Obama, Romney}])) = \alpha_0 + \alpha_1(\text{state last vote share}) + \sum_{k=1}^{K} a^k_{j_k[i]}$$

$$a^k_{j_k[i]} \sim N(0, \sigma^2_a)$$

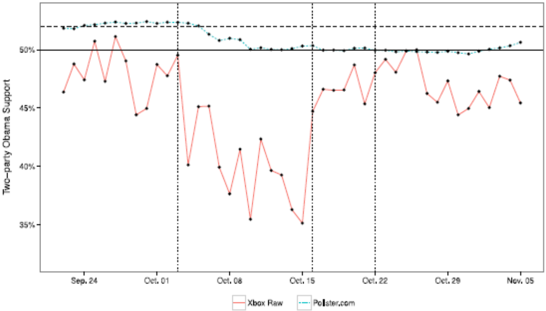
$$logit(P(Y_i \in [\text{Obama}] \mid Y_i \in [\text{Obama, Romney}])) =$$

$$\beta_0 + \beta_1(\text{state last vote share}) + \sum_{k=1}^{K} b^k_{j_k[i]}$$

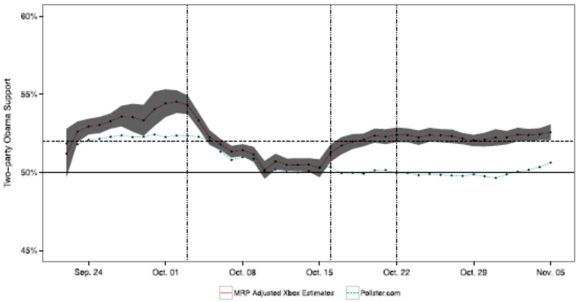
$$b^k_{j_k[i]} \sim N(0, \sigma^2_b)$$

where $j_k[i]$ indicates that the *ith* observation belongs to the *j*th category for the *k*th variable.

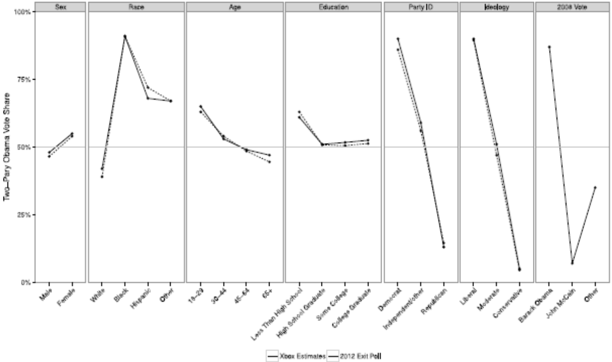
Raw Xbox Proportions vs. Tracking Polls



Multilevel Regression and Prediction vs. Tracking Polls



Xbox vs. 2012 Exit Polls in Poststrata



Advantages of Quasi-Randomization vs. Superpopulation

- Quasi-randomization has the advantage of creating a single weight for use with all analyses.
  - Convenient; parallels design-based framework, even if not strictly design-based.
  - Can go badly wrong if model is poor, and model diagnostics are not well-developed.
- Superpopulation model is more principled, but may work best with targeting a narrow set of parameters in a single analysis.
  - Fits within model-based framework.
  - Time consuming and may require higher degree of expertise to implement.

Next Steps

Certainly an open area for research!

- Propensity Scores
- Mean/Quantile Matching
- Mode effects, measurement error
- Data harmonization and alignment

Acknowledgements

- Rick Valliant, University of Michigan/University of Maryland
- Jack Chen, Survey Monkey, Inc.
- Carol Flannagan, University of Michigan

References

Chen, J.K.T., Valliant, R.L., Elliott, M.R., (2018). Model-Assisted Calibration of Non-probability Sample Survey Data Using Adaptive LASSO. *Survey Methodology*, 44:117-144.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4: 266-298.

Elliott, M. and Davis, W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: Combining data from the behavioral risk factor surveillance survey and the national health interview survey. *Journal of the Royal Statistical Society*, C54:595–609.

Elliott, M. and Little, R. J. A. (2000). Model averaging methods for weight trimming. *Journal of Official Statistics*, 16:191–209.

References

Elliott, M., Resler, A., Flannagan, C., and Rupp, J. (2010). Combining data from probability and non-probability samples using pseudo-weights. *Accident Analysis and Prevention*, 42:530–539.

Holt D., Smith T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society*, A142:33–46.

Korn, E. and Graubard, B. (1999). *Analysis of Health Surveys*. Wiley, New York.

Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77(378):237–250.

References

Rubin, D. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Smith, T. M. F. (1983). On the validity of inferences from non-random samples. *Journal of the Royal Statistical Society*, A146:394–403.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, B58:267–288.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

References

Wang, W., D. Rothschild, S. Goel, and A. Gelman (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31:980–991.

Zou, H. (2006). The adapptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.