Bayesian inference for sample surveys

Roderick Little and Trivellore Raghunathan

Module 9: Role of sampling weights in regression

Weighting and models

- The weights can't generally be ignored from a modeling perspective
  - Ignores different selection effects that bias estimates
- Weights are auxiliary covariates from a modeling perspective
- Design: weight the respondents
  - One size fits all *Y* variables
- Model: use weights to help predict non-sampled and non-responding values
  - Weighting adds noise for *Y*'s unrelated to weights
- The model perspective is more flexible (but potentially more work)

Weighting in multiple regression

- Model-based: standard method of estimation is ordinary least squares (OLS)
  - OLS if the residual variance is constant
  - Or WLS, weighting by the inverse of the residual variance, if the residual variance is not constant:

    $y_i \mid x_i \sim N(\beta_0 + \beta^T x_i, \sigma^2 / u_i),\ w_i \propto u_i$
- Design-based: WLS, weighting cases by inverse of probability of selection, $w_i = 1/\pi_i$
- These approaches to weighting are in general different, so which is right?
  - Much debated in the literature. See for example Brewer and Mellor (1973), DuMouchel and Duncan (1983)

Key concepts

- Superpopulation parameters: parameters included in a superpopulation model (e.g. the regression coefficients in a multiple regression model)
- Finite population quantities: population quantities defined by fitting model to the whole population, by some specified method (e.g. ordinary least squares)
- Target model: a model that defines the target quantity (or quantities) of interest
- Working model: a model used for predicting non-sampled units in the population
  - Could be Bayesian

Target model and working model 1

- Design: stratified sampling, Z = strata

  Target quantity: $\bar{Y}$

  Target model: $y_i \sim_{\text{iid}} N(\mu, \sigma^2)$

  ($\bar{Y}$ is estimate of μ from fitting this model to population)

Working model: $(y_i \mid z_i = j) \sim_{\text{iid}} N(\mu_j, \sigma^2)$

(Prediction model needs to condition on strata)

Resulting estimate of $\bar{Y}$ weights cases by their sampling weights

Target model and working model 2

- Design: stratified sampling, Z = strata

Target quantity: $\bar{Y}_j$ = mean of Y in stratum j

Target = working model: $(y_i \mid z_i = j) \sim N(\mu_j, \sigma^2)$

Estimate of $\bar{Y}_j = \bar{y}_j$, sample mean in stratum j

(Not weighted since weights are constant within strata)

Weighting in Regression

- Appropriate analysis depends on how the variables leading to the design weights enter the model of substantive interest
  - (a) all are included
  - (b) some are included, others aren't
  - (c) none are included
- Consider these distinctions for regression coefficients

Regression with sample weights

- Target model:

  $y_i \mid x_i \sim N(\beta_0 + \beta^T x_i, \sigma^2 / u_i)$, u_i known (constant for OLS)
- Target parameter: β
- Corresponding finite population parameter: B = result of fitting model to the entire population

  $z_i =$ design variables leading to sampling weights

  (stratum, size in pps sample)
- Consider three cases:
- (a) z_i included as part of x_i

  (b) z_i not a part of x_i

  (c) $z_i = (z_{i1}, z_{i2})$, z_{i1} a part of x_i, z_{i2} not a part of x_i

Regression with sample weights

(a) z_i included as part of x_i

If working model is correctly specified, then regression with weight u_i is correct – no need to include the sample weight

Design-weighted regression with weight $u_i w_i$ yields a design-consistent estimate of the target population quantity *B*. If this differs markedly from model estimate with weight u_i, this suggests model is misspecified, and assumptions need checking.

Regression with sample weights

(b) z_i not a part of x_i

Working model with weight u_i is subject to a known selection bias arising from the stratified design – only valid if this selection does not affect the target parameter estimate

Principled modeling approach is to regress y_i on x_i and z_i and then average over the distribution of z_i given x_i; e.g. if $E(y_i \mid x_i, z_i) = \gamma_0 + \gamma_1 x_i + \gamma_2 z_i$ then

$E(y_i \mid x_i) = \gamma_0 + \gamma_1 x_i + \gamma_2 E(z_i \mid x_i, \psi)$, etc.

Bayes simulation: impute draws of the non-sampled values of Y based on regression of Y on X, Z, and then fit regression of Y on X to imputed population. Repeat to simulate posterior distribution of β

Regression with sample weights

(b) z_i not a part of x_i

Pragmatic approach: design-based regression of y_i on x_i with weights $w_i u_i$

Model-based justification: assume a working model with a different regression model for y_i on x_i within each stratum defined by Z. Regression of y_i on x_i with weight $w_i u_i$ then approximates the posterior mean of β. (Little 2004, Example 11)

Regression with sample weights

(b) z_i not a part of x_i

Pragmatic approach B: compare regression of y_i on x_i with weights $w_i u_i$ with regression of y_i on x_i with weights u_i. If coefficients of interest are close, effects of selection may be ignored, leading to model-based solution.

Dumouchel and Duncan (1983): provides a test of equality of coefficients from weighted and unweighted least squares
Uses weighted least squares as a specification check on the target model

Regression with sample weights

(c) $z_i = (z_{i1}, z_{i2})$, z_{i1} a part of x_i, z_{i2} not a part of x_i

Principled modeling approach is to regress y_i on x_i and z_{i2} and then average over the distribution of z_{i2} given x_i; e.g. if $E(y_i \mid x_i, z_{i2}) = \gamma_0 + \gamma_1 x_i + \gamma_2 z_{i2}$ then

$E(y_i \mid x_i) = \gamma_0 + \gamma_1 x_i + \gamma_2 E(z_{i2} \mid x_i, \psi)$, etc.

Bayes simulation: impute draws of the non-sampled values of Y based on regression of Y on X, Z_2, and then fit regression of Y on X to imputed population. Repeat to simulate posterior distribution of β

Regression with sample weights

(c) $z_i = (z_{i1}, z_{i2})$, z_{i1} a part of x_i, z_{i2} not a part of x_i

Pragmatic approach: design-based regression of y_i on x_i with weights $w_{i2}u_i$, where w_{i2} is component of sampling weight attributable to z_{i2} (given z_{i1}).

(Weighting on w_iu_i is ok but inefficient)

Pragmatic approach B: compare regression of y_i on x_i with weights w_iu_i with regression of y_i on x_i with weights u_i. If coefficients of interest are close, effects of selection may be ignored, leading to model-based solution.

Summary

- Calibrated Bayes approach
  - Sampling weights as predictors
  - Flexible models for relationship between outcomes and sampling weights: eg penalized spline of propensity model
  - For regression, impute nonsampled cases using a target model that includes sampling weights (or stratum) as predictor; then fit target model to fitted population
- Next: cluster, multistage sampling