Weighting methods for Nonresponse

Week 8 (8.5)

Stat 260, St. Clair

Assume MAR mechanism

Missing at Random given covariates (MAR)

- ϕ_i depends on x_i but not y_i (given x_i)

$$\phi_i(x_i) = P(R_i = 1 \mid x_i) = P(R_i = 1 \mid x_i, y_i)$$

- *ignorable* nonresponse: "Model" ϕ given x, then nonresponse can be ignored

Weighting methods for nonresponse

Weighting class adjustment

- use known covariates from the selected **sample**
- e.g. urban/rural status of all selected units

Poststratification

- use known covariates from the **population**
- e.g. urban/rural proportions in the population

Covariates

- measurable variables we think are associated with nonresponse
- here: must be categorical variables
- within each combination of covariates:
  - ϕ_i is the same
  - respondents and nonrespondents have similar responses y_i

Weighting class adjustment

Basic idea:

- $w_i = 1/\pi_i$ is the sampling weight
  - based on full sample of $n = n_R + n_M$ units
- Probability a unit is selected AND responds:

$P(\text{unit } i \text{ select and responds})$

$= P(\text{unit } i \text{ selected}) \times P(\text{responds} \mid \text{selected})$

$= \pi_i \times \phi_i$

→ Respondent incl. prob.

Weighting class adjustment

Basic idea:

- $w_i = 1/\pi_i$ is the sampling weight
  - based on full sample of $n = n_R + n_M$ units
- The weighting class adjusted weights: ⇒ for respondents

$$\tilde{w}_i = \frac{1}{\pi_i \phi_i} = \frac{w_i}{\phi_i}$$

Weighting class adjusted estimates

$$\hat{t}_{wc} = \sum_{\text{respondents}} \tilde{w}_i y_i, \qquad \hat{\bar{y}}_{wc} = \frac{\sum_{\text{respondents}} \tilde{w}_i y_i}{\sum_{\text{respondents}} \tilde{w}_i}$$

↳ est. of # obs. units

Weighting class adjustment

Estimating the nonresponse probability ϕ

- Divide entire sample into known classes based on x covariates
- For all responding units in class c

$$\hat{\phi}_c = \frac{\sum_i w_i \text{ for responding units in class } c}{\sum_i w_i \text{ for all sampled units in class } c} \approx \frac{\hat{N}_{RC}}{\hat{N}_C}$$

- Weight for all responding units i in class c

$$\tilde{w}_i = \frac{w_i}{\hat{\phi}_c}$$

Example: Weighting class adjustment

- SRS of $n = 100$ likely voters from a population of $N = 1000$
- Response: Trump voter
- Parameter p: proportion for Trump
- Covariate: Education (college grad or not)
- Assume: nonresponse only depends on education

Example: Weighting class adjustment

$w_i = \frac{N}{n} = \frac{1000}{100} = 10$ SRS

class	n	n_R	$\hat{p}$	$\hat{\phi}_c$	$1/\hat{\phi}_c$	$\tilde{w}_c$
college	34	17	0.35	1/2	2	20
not	66	18	0.50	.27	3.7	36.67
all	100	35	0.43	----------	----------	----------

What are the weighting class adjusted weights?

College

$\hat{\phi}_{coll} = \frac{\sum_{coll,R} w_i}{\sum_{coll} w_i} = \frac{\sum_{coll,R} 10}{\sum_{coll} 10} = \frac{17 \times 10}{34 \times 10} = \frac{1}{2}$ $\quad \frac{1}{\hat{\phi}_{coll}} = 2$

$\tilde{w}_{coll} = \frac{w_i}{\hat{\phi}_{coll}} = \frac{10}{1/2} = 20$ ⇒ Each Responding college grad represents 20 people in pop.

Not College

$\hat{\phi}_{not} = \frac{\sum_{not,R} 10}{\sum_{not} 10} = \frac{18 \times 10}{66 \times 10} \approx .27$

$\tilde{w}_c = \frac{10}{.27} \approx 36.67$

each Respond. Not college represents 36.67 people in pop.

Example: Weighting class adjustment

$$\hat{p}_{wc} = \frac{\sum_R \tilde{w}_i y_i}{\sum_R \tilde{w}_i}$$

class	n	n_R	$\hat{p}$	$\hat{\phi}_c$	$1/\hat{\phi}_c$	$\tilde{w}_c$
college	34	17	0.35	0.5	2	20
not	66	18	0.50	0.27	3.7	36.67
all	100	35	0.43	-----	-----	----

What is the weighting class adjusted proportion?

$$\sum_R \tilde{w}_i = \sum_{coll,R} 20 + \sum_{Not,R} 36.67 = 17(20) + 18(36.67) = 1000$$

$$\sum_R \tilde{w}_i y_i = \sum_{coll,R} 20\, y_i + \sum_{Not,R} 36.67\, y_i = 20(6) + 36.67(9) = 450$$

$y_i = 1$ for $17(.35) \approx 6$ $\quad$ $y_i = 1$ for $18(\frac{1}{2}) = 9$

$y_i = 0$ for 11

$$\hat{p}_{wc} = \frac{450}{1000} = .45 \quad >> \quad \hat{p} = .43$$

Exercise 5.15: Weighting class adjustment

- Exercise 15: Treated a one-stage sample of teachers as two-stage design due to nonresponse.
- One stage sample of size n from N clusters.
- Due to nonresponse, we only have m_i respondents from the M_i units in cluster i

R full

- What are the weighting class adjusted weights?

* covariate = cluster => define "classes"

One stage: $w_{ij} = \frac{N}{n}$

cluster i

$$\phi_i = \frac{\sum_{R,i} w_{ij}}{\sum_{i} w_{ij}} = \frac{\sum_{j=1}^{m_i} \frac{N}{n}}{\sum_{j=1}^{M_i} \frac{N}{n}} = \frac{m_i \frac{N}{n}}{M_i \frac{N}{n}} = \frac{m_i}{M_i}$$

(Response rate in cluster i)

$$\tilde{w}_{ij} = \frac{w_{ij}}{\phi_i} = \frac{N/n}{m_i/M_i} = \frac{N}{n} \times \frac{M_i}{m_i}$$

weighting class adjustment = two-stage cluster stage.

Poststratification

Basic idea:

- Only consider respondents
- w_i are sampling weight
  - could be weight class adjusted, or
  - could be based on responding sample of n_R units
- Divide population into poststratum of size N_h
- For unit i in poststratum h:

$$w_i^* = \frac{w_i}{\sum_{i \in \text{stratum } h} w_i / N_h}$$

$$\frac{\sum_{R, i \in h} w_i}{N_h} = \frac{\hat{N}_{R,h}}{N_h} \quad \text{e.g.} \quad \frac{\hat{N}_{Rh}}{N_h} = \frac{1}{2}$$

$$w_i^* = 2 w_i$$

Poststratification

$$w_i^* = \frac{w_i}{\sum_{i \in \text{stratum } h} w_i / N_h}$$

- What will $\sum_{i \in \text{stratum } h} w_i / N_h$ be if nonresponse is *not* related to the poststratification covariates?

$$\frac{\sum_{R,h} w_i}{N_h} \approx 1 \Rightarrow w_i^* \approx w_i$$

- What will $\sum_{i \in \text{stratum } h} w_i / N_h$ be if nonresponse in stratum h is larger than in other stratum?

$$\frac{\sum_{R,h} w_i}{N_h} < 1 \Rightarrow w_i^* > w_i$$

Poststratification

$$w_i^* = \frac{w_i}{\sum_{i \in \text{stratum } h} w_i / N_h}$$

What is w_i^* for a SRS?

$w_i = \frac{N}{n_R}$

n_{Rh} = # respondents in poststratum h

$$\frac{\sum_{R,h} w_i}{N_h} = \frac{\sum_{R,h} \frac{N}{n_R}}{N_h} = \frac{n_{Rh} \frac{N}{n_R}}{N_h} = \frac{n_{Rh}}{N_h} \cdot \frac{N}{n_R}$$

$$w_i^* = \frac{N/n_R}{\frac{n_{Rh}}{N_h} \times \frac{N}{n_R}} = \frac{N_h}{n_{Rh}} \Rightarrow$$

exactly the "weights" used in earlier poststrat in ch. 4

Other methods for nonresponse

- Weighting methods can get complex very fast, depending on the number of classes/poststrata
    - concern when sample sizes are small or weights really big
- Other methods include
    - Raking
    - Regression modeling