HW3

Sam Olson

Problem 1:

A researcher wants to estimate the total number of patients discharged from hospitals in Iowa in January 2019. It is known that there are $N = 145$ hospitals in Iowa and the researcher obtains a list of all $N = 145$ hospitals in Iowa from administrative data. The list contains the number of inpatient beds in each hospital in the population. She (= the researcher) decides to select a sample using probability proportional to size sampling with replacement. The size variable (x_i) is the number of inpatient beds in the hospital. The total of x_i for all 145 hospitals in Iowa is $T_x = \sum_{i=1}^{N} x_i = 13,785$ inpatient beds. She selects a probability proportional to size sample with replacement with three independent draws and draw probability proportional to x_i. She collects the number of patients discharged in January 2019 for the sampled hospitals. The table below contains the data for the hospitals obtained in the three draws.

Draw	Hospital ID	Number of beds (x_i)	Number of patients (y_i)
1	46	250	754
2	88	100	321
3	113	450	1362

1.

What is the estimate of the total number of patients discharged in January 2019 from the population of hospitals in this region?

Answer

The estimate of the total number of patients discharged is:

$$\hat{T}_y = \frac{1}{3}\sum_{j=1}^{3}\frac{y_{i_j}}{\pi_{i_j}} = \frac{1}{3}\left(\frac{754}{250/13785} + \frac{321}{100/13785} + \frac{1362}{450/13785}\right) \approx 42{,}534.74$$

2.

Provide a 95% confidence interval for the total number of patients discharged in January 2019. Show intermediate steps.

Answer

$$\hat{V}(\hat{T}_y) = \frac{1}{n(n-1)}\sum_{j=1}^{3}\left(\frac{y_{i_j}}{\pi_{i_j}} - \hat{T}_y\right)^2 \approx 759{,}188.4$$

$$\hat{SE}(\hat{T}_y) = \sqrt{759{,}188.4} \approx 871.41$$

$$\text{CI}_{95\%} = \hat{T}_y \pm 1.96 \cdot \hat{SE} \approx (40{,}826.78,\ 44{,}242.70)$$

3.

Estimate the average number of patients discharged per hospital in January 2019 and provide a corresponding standard error. Show intermediate steps.

Answer

$$\hat{\bar{Y}} = \frac{\hat{T}_y}{N} = \frac{42{,}534.74}{145} \approx 293.34$$

$$\hat{SE}(\hat{\bar{Y}}) = \frac{\hat{SE}(\hat{T}_y)}{N} = \frac{871.41}{145} \approx 6.01$$

The average number of patients discharged per hospital is 293.34 ± 6.01.

Problem 2:

A city block is divided into 100 blocks from which 5 blocks are selected with replacement and with probability proportional to the number of households enumerated in a previous census. Within each sampled block, the average household income and the average household size (=number of people in the household) are obtained from the sampled blocks. The following table presents a summary of information obtained from the sample blocks.

Block	Block Size	Average Household income ($\times 10^{-3}$)	Average Household size
1	50	30	2
2	60	70	4
3	47	80	5
4	50	50	4
5	70	60	4

1.

What is the estimated average household income and its estimated variance?

Answer

$$\hat{Y} = \frac{1}{n}\sum_{k=1}^{n} \bar{y}_{a_k} = \frac{30+70+80+50+60}{5} = 58 \times 10^3\$$$

$$\hat{V}(\hat{Y}) = \frac{1}{n(n-1)}\sum_{k=1}^{n}\left(\bar{y}_{a_k} - \hat{Y}\right)^2 = \frac{(30-58)^2 + \cdots + (60-58)^2}{5 \times 4} = 74 \times 10^6\$^2$$

$$SE(\hat{Y}) = \sqrt{74} \times 10^3 \approx 8.60 \times 10^3\$$$

2.

What is the estimated per capita income (= income per person) and its estimated variance? (You may need to use a Taylor linearization.)

Answer

First compute average household size:

$$\hat{\hat{X}} = \frac{2+4+5+4+4}{5} = 3.8 \text{ people/household}$$

Per capita income ratio:

$$\hat{\theta} = \frac{\hat{Y}}{\hat{\hat{X}}} = \frac{58}{3.8} \approx 15.26 \times 10^3\$$$

Define the linearized variable:

$$z_k = y_k - \hat{\theta} x_k$$

where: - y_k = average household income for block k (in \$10ˆ3\$) - x_k = average household size for block k

Compute z_k values:

$$\begin{aligned}
z_1 &= 30 - 15.26 \times 2 = -0.52 \\
z_2 &= 70 - 15.26 \times 4 = 8.96 \\
z_3 &= 80 - 15.26 \times 5 = 3.70 \\
z_4 &= 50 - 15.26 \times 4 = -11.04 \\
z_5 &= 60 - 15.26 \times 4 = -1.04
\end{aligned}$$

Variance of z_k:

$$s_z^2 = \frac{1}{n-1}\sum_{k=1}^{n}(z_k - \bar{z})^2 = \frac{(-0.52 - 0.012)^2 + \cdots + (-1.04 - 0.012)^2}{4} \approx 54.29$$

Final variance estimate:

$$\hat{V}(\hat{\theta}) = \left(\frac{1}{\hat{\bar{\bar{X}}}}\right)^2 \frac{s_z^2}{n} = \frac{1}{3.8^2} \cdot \frac{54.29}{5} \approx 0.752 \times 10^6 \$^2$$

$$SE(\hat{\theta}) = \sqrt{0.752} \times 10^3 \approx 0.867 \times 10^3 \$$$

Per capita income: $15.26 \pm 0.87 \times 10^3 \$$

Problem 3:

A researcher wants to estimate the average household income in a city using two-phase sampling.

Phase 1: Basic Survey

200 households are selected using simple random sampling (SRS) from 5,000 households. Collected info: the household size x_i which is the total number of adults and children in household i.

Phase 2: Detailed Income Survey

From the 200 households, 80 households are selected to Collected info: Household income y_i (\$1,000).

1.

If the second phase sample were selected using probability proportional to household size (PPS). Calculate the second-phase conditional inclusion probabilities $\pi_{i|A_1}^{(2)}$ for a household i with 2 adults and 1 child. Can you compute the overall inclusion probability for this household?

Answer This cannot be numerically evaluated unless $T_x^{(1)}$ is known.

This follows from:

The Second-Phase Inclusion Probabilities for a household with $x_i = 3$ (2 adults + 1 child):

Conditional Probability ($\pi_{i|A_1}^{(2)}$):

$$\pi_{i|A_1}^{(2)} = \frac{n_2 \cdot x_i}{T_x^{(1)}} = \frac{80 \times 3}{\sum_{j \in A_1} x_j} = \frac{240}{T_x^{(1)}}$$

where $T_x^{(1)}$ is the total household size in the Phase 1 sample.

Overall Inclusion Probability:

$$\pi_i = \underbrace{\frac{200}{5000}}_{\text{Phase 1}} \times \underbrace{\frac{240}{T_x^{(1)}}_{\text{Phase 2}}}$$

Back to the Question The researcher decide to use a simple random sample in the second phase to select the 80 households, and the summary statistics from both phases are as follows:

Phase 1 Summary Statistics

$$\bar{x}_1 = 3.2, \quad s_{x1}^2 = 2.0$$

Phase 2 Summary Statistics

$$\bar{x}_2 = 3.5, \quad s_{x2}^2 = 2.2, \quad \bar{y}_2 = 58, \quad s_{y2}^2 = 100, \quad r_{xy} = 0.6$$

2.

Estimate the mean household income using $\pi^\star$-estimator.

Answer

$$\bar{y}_{\pi^\star} = \bar{y}_2 + \underbrace{(\bar{x}_1 - \bar{x}_2)}_{\text{Adjustment}} \cdot \underbrace{\frac{s_{xy}}{s_{x2}^2}}_{\text{Coefficient}}$$

Covariance:

$$s_{xy} = r_{xy} \cdot s_{x2} \cdot s_{y2} = 0.6 \times \sqrt{2.2} \times 10 \approx 8.9$$

Adjustment term:

$$\frac{8.9}{2.2} \approx 4.045$$

Final estimate:

$$58 + (3.2 - 3.5) \times 4.045 \approx 56.79 \text{ (\$1,000)}$$

3.

Calculate the approximate variance of the $\pi^\star$-estimate.

Answer

$$\text{Var}(\bar{y}_{\pi^\star}) \approx \frac{1}{n_2} s_{y2}^2 (1 - r_{xy}^2) = \frac{1}{80} \times 100 \times 0.64 = 0.8$$

4.

Calculate the regression estimator of the mean household income using household size as the covariate.

Answer

The regression estimator is:

$$\bar{y}_{\text{reg}} = \bar{y}_2 + (\bar{x}_1 - \bar{x}_2) \cdot b \quad \text{where } b = \frac{s_{xy}}{s_{x2}^2} \approx 4.045$$

$$\bar{y}_{\text{reg}} = 58 - 1.2135 = 56.79 \text{ (\$1,000)}$$

Note: Same numerical value as $\pi^\star$-estimator in this case.

5.

Calculate the approximate variance of the regression estimator.

Answer

$$\mathrm{Var}(\bar{y}_{\mathrm{reg}}) = \frac{1}{n_2} s_{y2}^2 (1 - r_{xy}^2) = 0.8$$

6.

What advantage does the regression estimator have over the $\pi^\star$-estimate?

Answer

Flexibility:
- Regression does not assume the relationship passes through the origin (unlike ratio estimators).
- Can incorporate multiple covariates if needed.

Efficiency:
- Exploits correlation structure more effectively (here, $r_{xy} = 0.6$).
- Generally achieves lower variance when x and y are strongly correlated.

Robustness:
- Less sensitive to violations of the ratio model assumptions.

Note: The identical results here occur because the regression reduces to the ratio estimator when the intercept is zero. In practice, regression is preferred unless a ratio relationship is explicitly justified.

Problem 4:

A health researcher is studying the effect of a new drug treatment ($T = 1$) versus a control ($T = 0$) on patient blood pressure reduction (Y). Because treatment was not randomly assigned, the researcher uses observational data and applies causal inference methods.

The data below summarize 10 patients:

	Treatment	Blood Pressure	Age		
ID	(T)	Change (Y)	(X)	Propensity Score $\hat{\pi}(X)$	$\hat{Q}(X,1)$, $\hat{Q}(X,0)$
1	1	-12	55	0.7	-11, -6
2	1	-10	60	0.6	-12, -7
3	1	-13	50	0.8	-10, -5
4	1	-15	65	0.5	-14, -8
5	0	-5	55	0.7	-11, -6
6	0	-6	60	0.6	-12, -7
7	0	-7	50	0.8	-10, -5
8	0	-9	65	0.5	-14, -8
9	1	-11	58	0.65	-11, -6
10	0	-8	62	0.55	-12, -7

1.

Calculate the IPW estimate of the average blood pressure change for the treated and control groups.

Answer

The Inverse Probability Weighted (IPW) estimates for each group are given by:

$$\bar{Y}_{IPW}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{\pi}(X_i)} \quad , \quad \bar{Y}_{IPW}^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)}$$

Treated Group ($T = 1$):

$$\bar{Y}_{IPW}^{(1)} = \frac{1}{10}\left(\frac{-12}{0.7} + \frac{-10}{0.6} + \frac{-13}{0.8} + \frac{-15}{0.5} + \frac{-11}{0.65}\right) = -9.50 \text{ mmHg}$$

Control Group ($T = 0$):

$$\bar{Y}_{IPW}^{(0)} = \frac{1}{10}\left(\frac{-5}{0.3} + \frac{-6}{0.4} + \frac{-7}{0.2} + \frac{-9}{0.5} + \frac{-8}{0.45}\right) = -10.05 \text{ mmHg}$$

2.

Compute the DIME estimate of average treatment effect (ATE) without considering the propensity scores. Is this in general a good estimate for ATE? Briefly explain your reasoning.

Answer

The DIME (Difference in Means Estimator) is:

$$\widehat{ATE}_{DIME} = \bar{Y}_T - \bar{Y}_C$$

With - Treated group: $\bar{Y}_T = \frac{-12-10-13-15-11}{5} = -12.2$ - Control group: $\bar{Y}_C = \frac{-5-6-7-9-8}{5} = -7.0$

$$\widehat{ATE}_{DIME} = \underbrace{\frac{-12-10-13-15-11}{5}}_{\bar{Y}_T=-12.2} - \underbrace{\frac{-5-6-7-9-8}{5}}_{\bar{Y}_C=-7.0} = -5.2 \text{ mmHg}$$

No, this is not a generally good estimate — this estimate is not generally reliable because treatment was not randomly assigned, so the groups may differ systematically in confounders like age. The DIME estimator can be biased in observational studies.

3.

Calculate the IPW estimate of the ATE.

Answer

$$\widehat{ATE}_{IPW} = -9.50 - (-10.05) = 0.55 \text{ mmHg}$$

4.

Calculate the optimal AIPW estimate of the ATE.

Answer

The Augmented IPW (AIPW) estimator is:

$$\widehat{ATE}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i(Y_i - \hat{Q}(X_i,1))}{\hat{\pi}(X_i)} + \hat{Q}(X_i,1) - \frac{(1-T_i)(Y_i - \hat{Q}(X_i,0))}{1-\hat{\pi}(X_i)} - \hat{Q}(X_i,0)\right]$$

```r
df <- data.frame(
  Trt = c(1,1,1,1,0,0,0,0,1,0),  # Changed from T to Trt
  Y = c(-12,-10,-13,-15,-5,-6,-7,-9,-11,-8),
  pi = c(0.7,0.6,0.8,0.5,0.7,0.6,0.8,0.5,0.65,0.55),
  Q1 = c(-11,-12,-10,-14,-11,-12,-10,-14,-11,-12),
  Q0 = c(-6,-7,-5,-8,-6,-7,-5,-8,-6,-7)
)

# AIPW contribution per unit
df$aipw <- with(df,
  Trt*(Y - Q1)/pi + Q1 - (1 - Trt)*(Y - Q0)/(1 - pi) - Q0
)

round(mean(df$aipw), 2)
```

```
## [1] -4.75
```

5.

What is the advantage of using AIPW over IPW?

Answer

The Augmented Inverse Probability Weighting (AIPW) estimator combines both the propensity score model and the outcome regression model, making it doubly robust (which is covered in more detail in the next part.

The Augmented Inverse Probability Weighting (AIPW) estimator provides two key advantages over standard IPW:

1. Double Robustness Property
   - Produces consistent estimates if either: The propensity score model $\hat{\pi}(X)$ is correctly specified, or The outcome models $\hat{Q}(X, 1)$ and $\hat{Q}(X, 0)$ are correctly specified
2. Improved Efficiency
   - Reduces variance compared to IPW by combining weighting with regression adjustment
   - More stable with extreme propensity scores (near 0 or 1)

6.

Explain why AIPW is called doubly robust.

Answer

The AIPW estimator is referred to as doubly robust because AIPW is consistent if either the outcome model or the propensity model is correct. Specif:

- It produces a consistent estimate of the ATE if either:
  - The propensity score model is correctly specified, or
  - The outcome regression model is correctly specified

This property ensures that even if one of the two models is misspecified, the estimator still remains valid as long as the other is correct.