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

For this problem, we use the PPS estimator (with replacement). As we are given the "draw probabilities", the estimator is of the form:

$$\hat{T}_y = \frac{1}{n} \sum_{j=1}^{n} \frac{y_{i_j}}{p_{i_j}}$$

With $n = 3$.

Calculating using the information provided, our estimate is:

$$\hat{T}_y = \frac{1}{3} \sum_{j=1}^{3} \frac{y_{i_j}}{p_{i_j}} = \frac{1}{3} \left(\frac{754}{250/13785} + \frac{321}{100/13785} + \frac{1362}{450/13785} \right) \approx 42,516$$

```r
denom <- 13785
p_ij <- c((250/denom), (100/denom), (450/denom))
y_ij <- c(754, 321, 1362)
n <- 3

tHat <- (1/n) * sum(y_ij / p_ij)
tHat
```

```
## [1] 42516
```

Using the above, we estimate the total number of patients discharged in January 2019 from the population of hospitals in this region is 42,516.

Note: When I manually calculated this (greater precision) the estimate came out to 42,517.87, which is close but nonetheless different than the above. This also has implications for the resulting confidence interval, though only slightly. For simplicity, the answers provided will align with the R output.

2.

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

Answer

We center our confidence interval on the point estimate calculated in the prior step.

We then need to find the critical value, which for a 95% confidence interval is roughly 1.96 (assuming normally distributed mean/average via CLT).

Then, we have to calculate our Standard Error. To begin with, we calculate the variance using the estimated variance formula for PPS, of the form:

$$\hat{V}(\hat{T}_y) = \frac{1}{n(n-1)} \sum_{j=1}^{3} \left(\frac{y_{i_j}}{p_{i_j}} - \hat{T}_y \right)^2 \approx 753{,}357.8$$

```r
variance_est <- (1/(n*(n-1))) * sum(((y_ij/p_ij) - tHat)^2)
variance_est
```

```
## [1] 753357.8
```

With the variance estimate, we then get the standard error and combine with our typical confidence interval formula. Giving us:

$$\hat{SE}(\hat{T}_y) = \sqrt{753{,}357.8} \approx 867.96$$

Taken together, our 95% confidence interval is of the form:

$$\text{CI}_{95\%} = \hat{T}_y \pm 1.96 \cdot \hat{SE} \approx (40{,}814.79,\ 44{,}217.21)$$

So, our 95% confidence interval for the total number of patients discharged in January 2019 is (40,814.79, 44,217.21).

```r
SE <- sqrt(variance_est)
SE
```

```
## [1] 867.9619
```

```r
CI_lower <- tHat - 1.96 * SE
CI_upper <- tHat + 1.96 * SE
c(CI_lower, CI_upper)
```

```
## [1] 40814.80 44217.21
```

3.

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

Answer

$$\hat{\bar{Y}} = \frac{\hat{T}_y}{N} = \frac{42{,}516}{145} \approx 293.21$$

$$\hat{SE}(\hat{\bar{Y}}) = \frac{\hat{SE}(\hat{T}_y)}{N} = \frac{867.96}{145} \approx 5.99$$

Explicitly, without rounding (hand calculation), we have: The average number of patients discharged per hospital is 293.21 with Standard Error 5.99.

```r
N <- 145
tHat/N
```

```
## [1] 293.2138
```

```r
SE/N
```

```
## [1] 5.985944
```

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

For this problem, we again use the PPS estimator (with replacement, again). However, as we are estimating an average, and do not have the draw probabilities, we use the following formula for our point estimate:

$$\hat{\bar{Y}} = \frac{1}{n}\sum_{k=1}^{n} \bar{y}_k = \frac{30 + 70 + 80 + 50 + 60}{5} = 58$$

And for the variance formula, we have:

$$\hat{V}(\hat{\bar{Y}}) = \frac{1}{n(n-1)}\sum_{k=1}^{n} (\bar{y}_k - \hat{\bar{Y}})^2 = \frac{(30-58)^2 + \cdots + (60-58)^2}{20} = 74$$

Taken together, the estimated average household income is $\$58 \times 10^3$ and its estimated variance is $\$74 \times 10^3$.

Note: I believe there is a typo in the table provided, and that $\times 10^3$ should be used instead of $\times 10^{-3}$. This assumption will be carried into the following problems when giving units of the estimates.

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:

$$\hat{\bar{X}} = \frac{2 + 4 + 5 + 4 + 4}{5} = 3.8$$

Giving us a point estimate of on average 3.8 people per household. We need to convert this into a ratio though, so we then have:

$$\hat{\theta} = \frac{\hat{Y}}{\hat{\hat{X}}} = \frac{58}{3.8} \approx 15.26$$

Giving us the ratio $\$15.26 \times 10^3$

Via linearizaton, we define the variable z (linarized variable) as:

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

Note: This method of linearization is somewhat different than the method we've typically used in class. However, I believe it does match the ultimate results.

That being said, we compute z_k values for each of the five blocks by:

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

Computing the variance using the values of the linearized variable:

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

Converting back to an estimated variance of θ, we have:

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

Giving us an estimated per capita income of $15.26(\times 10^3)$ with estimated variance $0.752(\times 10^3)$.

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 $\pi^{(2)}_{i|A_1}$ for a household i with 2 adults and 1 child. Can you compute the overall inclusion probability for this household?

Answer We cannot explicitly compute the overall inclusion probability π_i unless we know the total household size $T_x^{(1)} = \sum_{j=1}^{200} x_j$ from Phase 1. And this typically is something we don't know, as it would mean we already have data from all Phase 1 households. I believe this is a fundamental limitation of two-phase designs and why estimators such as π^* estimator/regression estimator are used for two-phase sampling designs.

Explicitly, the above conclusion comes from:

Under the Two-Phase Sampling design, the second-phase inclusion probability for a household with $x_i = 3$ (2 adults + 1 child) is given by:

$$\pi^{(2)}_{i|A_1} = \frac{n \cdot x_i}{T_x^{(1)}}$$

Using all known quantities, up to this point, this simplifies to:

$$\pi^{(2)}_{i|A_1} = \frac{80 \cdot 3}{\sum_{j \in A_1} x_j} = \frac{240}{T_x^{(1)}}$$

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

The overall inclusion probability is then given by:

$$\pi_i = \pi^{(1)} \cdot \pi^{(2)}_{i|A_1} = \frac{200}{5000} \cdot \left(80 \cdot \frac{x_i}{T_x^{(1)}}\right)$$

At most, we may simplify to:

$$\pi_i = \frac{16000 x_i}{5000 T_x^{(1)}}$$

However, as noted, we still require knowing $T_x^{(1)}$ explicitly to calculate the overall inclusion probability.

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

The formula for the $\pi^\star$-estimator is of the form:

$$\hat{Y}^* = \sum_{i \in A_2} \frac{y_i}{\pi_i^*}$$

But we need to calculate:

$$\pi_i^* = \pi_i^{(1)} \cdot \pi_{i|A_1}^{(2)}$$

To that end, for any household i, the Phase 1 inclusion probability is:

$$\pi_i^{(1)} = \frac{n_1}{N} = \frac{200}{5000} = 0.04$$

Given Phase 1 sample, the conditional Phase 2 inclusion probability is:

$$\pi_{i|A_1}^{(2)} = \frac{n_2}{n_1} = \frac{80}{200} = 0.4$$

Taken together, we have:

$$\pi_i^* = \pi_i^{(1)} \cdot \pi_{i|A_1}^{(2)} = 0.04 \cdot 0.4 = 0.016$$

Returning to the original expression for the π^* estimator:

$$\hat{Y}^* = \sum_{i \in A_2} \frac{y_i}{\pi_i^*} = \frac{1}{0.016} \sum_{i \in A_2} y_i$$

One more issue: We need to derive $\sum_{i \in A_2} y_i$ using the Phase 2 sample mean, $\bar{y}_2 = 58$, by:

$$\sum_{i \in A_2} y_i = n_2 \cdot \bar{y}_2 = 80 \cdot 58 = 4{,}640$$

Thus:

$$\hat{Y}^* = \frac{4,640}{0.016} = 290,000$$

However, that quantity is the total estimate. To get to the average household estimate, we divide by the total population size $N = 5,000$, taken from Phase 1:

$$\bar{y}_{\pi^*} = \frac{\hat{Y}^*}{N} = \frac{290,000}{5,000} = 58$$

Given units are in \$1,000, our π^* estimator provides a mean income estimate of \$58,000.

3.

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

Answer

Although the $\pi^\star$-estimator is exactly defined in the lecture slides, Slide 10 of Ch11p1, we do not have all necessary information for an exact calculation, i.e., we do need the individual y_i values, and for this calculation (unlike the point estimate) it is not sufficient to use the average. For that reason, we approximate its variance by deconstructing the variance formula into its Phase 1 and Phase 2 components.

To that end, we start with the closed form equation for the variance of the π^* estimate:

$$V(\hat{Y}^*) = V_1\left(\sum_{i\in A_1}\frac{y_i}{\pi_i^{(1)}}\right) + E_1\left(\sum_{i\in A_1}\sum_{j\in A_1}\Delta^{(2)}_{ij|A_1}\frac{y_i}{\pi_i^*}\frac{y_j}{\pi_j^*}\right)$$

Where:

First Phase Variance:

$$V_1\left(\sum_{i\in A_1}\frac{y_i}{\pi_i^{(1)}}\right)$$

Second Phase Variance:

$$E_1\left(\sum_{i\in A_1}\sum_{j\in A_1}\Delta^{(2)}_{ij|A_1}\frac{y_i}{\pi_i^*}\frac{y_j}{\pi_j^*}\right)$$

And where:

$$\Delta^{(2)}_{ij|A_1} = \pi^{(2)}_{ij|A_1} - \pi^{(2)}_{i|A_1}\pi^{(2)}_{j|A_1}$$

For Phase 1, we have SRS of $n_1 = 200$ from $N = 5000$

And for Phase 2, we have SRS of $n_2 = 80$ from $n_1 = 200$

We may simplify the two variance components as follows:

Phase 1:

$$V_1 = N^2 \left(1 - \frac{n_1}{N}\right) \frac{S_y^2}{n_1}$$

Phase 2

$$E_1 \left[V_2\right] = N^2 \left(\frac{1}{n_2} - \frac{1}{n_1}\right) \frac{S_y^2}{n_1}$$

Taken together, we have a closed form variance approximation that can be calculated using known values, using:

$$\text{Var}(\hat{Y}^*) \approx N^2 \left[\left(\frac{1}{n_1} - \frac{1}{N}\right) + \left(\frac{1}{n_2} - \frac{1}{n_1}\right)\right] s_{y2}^2$$

Where $S_y^2 \approx s_{y2}^2$

Calculating:

$$\left(\frac{1}{n_1} - \frac{1}{N}\right) = \frac{1}{200} - \frac{1}{5000} = 0.005 - 0.0002 = 0.0048$$

$$\left(\frac{1}{n_2} - \frac{1}{n_1}\right) = \frac{1}{80} - \frac{1}{200} = 0.0125 - 0.005 = 0.0075$$

The total variance is then given by:

$$5000^2 \cdot (0.0048 + 0.0075) \cdot 100 = 30,750,000$$

And the variance of the estimate is given by:

$$\text{Var}(\bar{y}_{\pi^*}) = \text{Var}\left(\frac{\hat{Y}^*}{N}\right) = \frac{30{,}750{,}000}{5000^2} = 1.23$$

Given units are in \$1,000, our variance estimate is approximately $\$1,230$.

4.

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

Answer

The regression estimator formula is of the form:

$$\bar{y}_{\text{reg}} = \bar{y}_2 + (\bar{x}_1 - \bar{x}_2) \cdot b$$

Where:

$$b = \frac{s_{xy}}{s_{x2}^2} \approx 4.045$$

All other quantities are known values, so we may calculate:

$$\bar{y}_{\text{reg}} = 58 - 1.2135 = 56.79$$

Given units are in \$1,000, our regression estimator point estimate is $\$56,790$, which is close but not exactly our estimate from the π^* estimator.

5.

Calculate the approximate variance of the regression estimator.

Answer

The approximate variance is of the form:

$$\text{Var}(\bar{y}_{reg,tp}) \approx \left(\frac{1}{n} - \frac{1}{N}\right) B' S_{xx} B + \left(\frac{1}{r} - \frac{1}{n}\right) S_{ee}$$

Where:

Phase 1 Variance:

$$\left(\frac{1}{n} - \frac{1}{N}\right) B' S_{xx} B$$

Phase 2 Variance:

$$\left(\frac{1}{r} - \frac{1}{n}\right) S_{ee}$$

And where:

$$B = \frac{S_{xy}}{S_x^2}$$

And:

$$S_{ee} = S_y^2(1 - r^2)$$

Using known quantities, we calculate:

First Term (Phase 1):

$$B' S_{xx} B = \rho^2 S_y^2 = 0.36 \cdot 100 = 36 \rightarrow \left(\frac{1}{200} - \frac{1}{5000}\right) \cdot 36 = 0.1728$$

Second Term (Phase 2):

$$\left(\frac{1}{80} - \frac{1}{200}\right) \cdot 64 = 0.48$$

Combining:

$$0.1728 + 0.48 = 0.6528$$

Given units are in \$1,000, our regression estimator estimated variance is \$652.80, which is smaller than the estimated variance of the π^* estimator.

6.

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

Answer

The regression estimator is more flexible, efficient, and robust. In that order: The regression estimator can incorporate multiple covariates, which can improve the generalizability of results. Also, because the regression estimator takes advantage of covariates, it "exploits" the correlation structure between the response and covariate(s), and as a result can be more efficient (have smaller variance) than the $\pi^\star$-estimator, especially so when the covariate(s) are strongly correlated with the response. Lastly, the regression estimator is more robust in the sense that it is less sensitive to violations of the model assumptions, i.e., it can still be consistent (in the traditional sense, i.e., converges to the true parameter) when the model is mis-specified.

For the purposes of this problem, we observe the two estimators producing similar but different point estimates, and a more efficient estimator in the regression estimator compared to the π^* estimator. So results are as expected, given we have a correlation coefficient of $r_{xy} = 0.6$ from the Phase 2 Summary Statistics.

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:

ID	Treatment (T)	Blood Pressure Change (Y)	Age (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

Treatment Group, $T = 1$: Formula:

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

Calculating:

$$\text{Numerator} = \frac{-12}{0.7} + \frac{-10}{0.6} + \frac{-13}{0.8} + \frac{-15}{0.5} + \frac{-11}{0.65} = -17.14 - 16.67 - 16.25 - 30.00 - 16.92 = -96.98$$

$$\text{Denominator} = \frac{1}{0.7} + \frac{1}{0.6} + \frac{1}{0.8} + \frac{1}{0.5} + \frac{1}{0.65} = 1.429 + 1.667 + 1.25 + 2 + 1.538 = 7.884$$

Combining

$$\bar{Y}^{(1)}_{IPW} = \frac{-96.98}{7.884} \approx -12.31$$

Control Group, $T = 0$: Formula:

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

Calculating:

$$\text{Numerator} = \frac{-5}{0.3} + \frac{-6}{0.4} + \frac{-7}{0.2} + \frac{-9}{0.5} + \frac{-8}{0.45} = -16.67 - 15.00 - 35.00 - 18.00 - 17.78 = -102.45$$

$$\text{Denominator} = \frac{1}{0.3} + \frac{1}{0.4} + \frac{1}{0.2} + \frac{1}{0.5} + \frac{1}{0.45} = 3.333 + 2.5 + 5 + 2 + 2.222 = 15.055$$

Combining terms:

$$\bar{Y}_{IPW}^{(0)} = \frac{-102.45}{15.055} \approx -6.81$$

Explicitly, in R:

```
data <- data.frame(
  ID = 1:10,
  T = c(1, 1, 1, 1, 0, 0, 0, 0, 1, 0),
  Y = c(-12, -10, -13, -15, -5, -6, -7, -9, -11, -8),
  pi_X = c(0.7, 0.6, 0.8, 0.5, 0.7, 0.6, 0.8, 0.5, 0.65, 0.55)
)

numer_treated <- sum((data$T * data$Y) / data$pi_X)
denom_treated <- sum(data$T / data$pi_X)
ipw_treated <- numer_treated / denom_treated

numer_control <- sum(((1 - data$T) * data$Y) / (1 - data$pi_X))
denom_control <- sum((1 - data$T) / (1 - data$pi_X))
ipw_control <- numer_control / denom_control

ipw_treated
```

```
## [1] -12.30166
```

```
ipw_control
```

```
## [1] -6.804428
```

Similar to Q1, there is a slight difference in the manual calculations compared to the R calculations due to a loss of precision when using the `sum` function with fractions. For this problem though, the differences are very small (negligible).

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:

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

Where:

Treated group average: $\bar{Y}_T = \frac{-12-10-13-15-11}{5} = -12.2$

And:

Control group average: $\bar{Y}_C = \frac{-5-6-7-9-8}{5} = -7.0$

$$\widehat{ATE}_{DIME} = \frac{-12-10-13-15-11}{5} - \frac{-5-6-7-9-8}{5} = -5.2$$

```
treated_outcomes <- c(-12, -10, -13, -15, -11)
control_outcomes <- c(-5, -6, -7, -9, -8)

mean_treated <- mean(treated_outcomes)
mean_control <- mean(control_outcomes)
ate_dime <- mean_treated - mean_control

ate_dime
```

```
## [1] -5.2
```

No, the DIME estimate of average treatment effect (ATE) without considering the propensity scores is generally **not** a good estimate! This estimate is problematic because it assumes treatments were randomly assigned, which is not the case! This introduces bias, as the estimate is possibly being confounded by other factors that may be relevant to the analysis (estimate of the treatment effect) such as age, gender, or other something else.

3.

Calculate the IPW estimate of the ATE.

Answer

$$\hat{\tau}_{IPW} = \bar{Y}^{(1)}_{IPW} - \bar{Y}^{(0)}_{IPW} = -12.31 - (-6.81) = -5.50$$

```
ipw_treated - ipw_control
```

```
## [1] -5.497233
```

The IPW estimate of the ATE is ≈ -5.50 (Treatment - Control).

4.

Calculate the optimal AIPW estimate of the ATE.

Answer

The optimal AIPW estimator formula is given by:

$$\hat{\tau}_{AIPW} = \hat{\mu}_{1,AIPW} - \hat{\mu}_{0,AIPW}$$

Where:

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

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

Where we use n instead of N, i.e., we swap our population N with the sample n for the purposes of estimation.

We may simplify the difference expression $\hat{\tau}$ somewhat, to get:

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

I am not going to attempt this manually. I defer strictly to the R calculation below:

```
df <- data.frame(
  Trt = c(1,1,1,1,0,0,0,0,1,0),
  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)
)

df$aipw <- df$Trt * (df$Y - df$Q1) / df$pi + df$Q1 -
           (1 - df$Trt) * (df$Y - df$Q0) / (1 - df$pi) - df$Q0

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

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

Giving us an optimal AIPW estimate of the ATE of -4.75.

Extra Validation: Individual Means, Then Difference

```
n <- nrow(df)

mu1_aipw <- with(df, mean(
  Trt * Y / pi - (Trt - pi) / pi * Q1
))

mu0_aipw <- with(df, mean(
```

```
  (1 - Trt) * Y / (1 - pi) + (Trt - pi) / (1 - pi) * Q0
))

mu1_aipw
```

```
## [1] -12.08452
```

```
mu0_aipw
```

```
## [1] -7.338889
```

```
diff <- mu1_aipw - mu0_aipw
diff
```

```
## [1] -4.745635
```

With rounding, the formula is consistent with the initial optimal AIPW estimate of the ATE of -4.75.

5.

What is the advantage of using AIPW over IPW?

Answer

The advantages the AIPW estimator has compared to IPW is the fact it is "doubly robust" and that it has possibly improved efficiency. The AIPW estimator combines both the propensity model (PM) and the outcome model (OM), making it doubly robust, covered in more detail in the next part. On the point of efficiency: The AIPW estimator reduces variance compared to IPW by combining weighting with regression adjustment, making it at least as efficient as the IPW estimator. The AIPW is also more stable with extreme propensity scores, i.e., scores near the tails, 0 or 1, due to its weighting adjustment. By comparison, the IPW estimate requires the propensity score model to be correct to ensure consistency.

6.

Explain why AIPW is called doubly robust.

Answer

The AIPW estimator is called "doubly robust" because the AIPW estimator is consistent if either the OM or the PM is correct. Specifically: The AIPW model can still be unbiased if either the OM or PM are mis-specified. This is not XOR though, at least one needs to be correctly specified!