HW2

Sam Olson

Problem 1 (20 pt)

A city has a total of 100,000 dwelling units, of which 35,000 are houses, 45,000 are apartments, and 20,000 are condominiums. A stratified sample of size $n = 1000$ is selected using proportional allocation (and rounding the sample sizes to the nearest integer). The three strata are houses ($h = 1$), apartments ($h = 2$), and condominiums ($h = 3$). The table below gives the estimates of the mean energy consumption per dwelling unit for the three strata and the corresponding standard errors.

Stratum (h)	Estimated Mean Energy Consumption ($\bar{y}_h$) (kWh per dwelling unit)	Estimated Standard Error ($\hat{SE}(\bar{y}_h)$)
House ($h = 1$)	915	4.84
Apartments ($h = 2$)	641	2.98
Condominium ($h = 3$)	712	7.00

1.

Estimate the total energy consumption for the full population of 100,000 dwelling units.

Answer

```
nh <- c(35000, 45000, 20000)
barY <- c(915, 641, 712)
tStr <- sum(nh * barY)
tStr
```

```
## [1] 75110000
```

$$\hat{T}_{str} = \sum_{h \in H} N_h \bar{y}_h = 915(35,000) + 641(45,000) + 712(20,000) = 75,110,000$$

Units of kWh per dwelling unit

2.

Estimate the standard error of the estimator used in (1).

Answer

```
nh2 <- nh^2
seBarY <- c(4.84, 2.98, 7.00)
seHatT <- sqrt(sum(nh2 * (seBarY^2)))
seHatT
```

```
## [1] 257447.4
```

$$SE(\hat{T}_{str}) = \sqrt{\text{Var}\left(\sum_{h \in H} N_h \bar{y}_h\right)} = \sqrt{\left(\sum_{h \in H} \text{Var}\left(N_h \bar{y}_h\right)\right)} = \sqrt{\left(\sum_{h \in H} N_h^2 \text{Var}\left(\bar{y}_h\right)\right)}$$

$$SE(\hat{T}_{str}) = \sqrt{(35,000^2)(4.84^2) + (45,000^2)(2.98^2) + (20,000^2)(7.00^2)} = 257,447.4$$

Units of kWh per dwelling unit

3.

What would be the sample size if the optimal allocation is to be used (under $n = 1000$) for this population? Assume that the survey costs are the same for each stratum.

Hint: Use the following steps:

a)

What is the sample size n_h for each stratum under proportional allocation?

Answer

$$\frac{n}{N} = \frac{1,000}{100,000} = 0.01$$

Under proportional allocation, we just take the above sampling rate multiplied by the population of the strata, i.e.:

$$n_h = N_h \times 0.01$$

Corresponding to:

$$n_1 = 350$$

$$n_2 = 450$$

$$n_3 = 200$$

b)

Note that:

$$\hat{SE}(\bar{y}_h) = \sqrt{\frac{1}{n_h}\left(1 - \frac{n_h}{N_h}\right)s_h^2}$$

Thus, you can obtain s_h^2.

Answer To obtain s_h^2, note that by definition:

$$SE(\bar{y}_h) = \sqrt{\left(\frac{1}{n_h} - \frac{1}{N_h}\right)S_h^2}$$

```
littlenH <- c(350, 450, 200)

# should be better with notation, but I love camel case
bigSh <- (seBarY^2)/(littlenH^-1 - nh^(-1))
sqrt(bigSh)
```

```
## [1] 91.00427 63.53381 99.49367
```

We want the population-level standard errors by strata. To that end:

$$\sqrt{\left(\frac{1}{350} - \frac{1}{35000}\right) S_1^2} = 4.84$$

$$\sqrt{\left(\frac{1}{450} - \frac{1}{45000}\right) S_2^2} = 2.98$$

$$\sqrt{\left(\frac{1}{200} - \frac{1}{20000}\right) S_3^2} = 7.00$$

Solving for S_1, S_2, S_3 respectively:

$$S_1 \approx 91.00, \quad S_2 \approx 63.53, \quad S_3 \approx 99.49$$

c)

Apply Neyman allocation (optimal allocation) using s_h in place of S_h.

Answer Combining everything from the above together, we use the Neyman allocation formula, i.e.:

$$n_h = \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} n$$

where $n = 1000$.

```
sh <- c(91, 63.53, 99.49)

neyman <- round(((nh * sh)/ sum(nh *sh))*1000,
                digits = 0)
neyman
```

```
## [1] 396 356 248
```

```
sum(neyman)
```

```
## [1] 1000
```

Explicitly,

$$n_1 = \frac{35,000(91.00)}{35,000(91.00) + 45,000(63.53) + 20,000(99.49)} \cdot 1000 \approx 396$$

$$n_2 = \frac{45,000(63.53)}{35,000(91.00) + 45,000(63.53) + 20,000(99.49)} \cdot 1000 \approx 356$$

$$n_3 = \frac{10,000(99.49)}{35,000(91.00) + 45,000(63.53) + 20,000(99.49)} \cdot 1000 \approx 248$$

$$n_1 \approx 396 \quad n_2 \approx 356, \quad n_3 \approx 248$$

Where we round to the nearest integer, and validate by checking the sum of samples is 1,000 as expected.

4.

What would be the estimated standard error of the total estimator under the optimal allocation in (3)? Compare it with the answer in (2). Which one is smaller?

Answer

The standard error via Neyman allocation is given by the formula:

$$SE(\hat{T}_{str}) = \sqrt{\sum_{h=1}^{H} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_h^2}$$

```
firstTerm <- nh^2 / littlenH
secondTerm <- (1 - (littlenH/nh))
thirdTerm <- sh^2
sqrt(sum(firstTerm * secondTerm * thirdTerm))
```

```
## [1] 257435.2
```

Given:

$$N_1 = 35,000, \quad N_2 = 45,000, \quad N_3 = 20,000$$

$$n_1 = 396, \quad n_2 = 356, \quad n_3 = 248$$

$$S_1 \approx 91.00, \quad S_2 \approx 63.53, \quad S_3 \approx 99.49$$

We compute:

$$\frac{N_1^2}{n_1} \left(1 - \frac{n_1}{N_1}\right) S_1^2$$

$$\frac{(35,000)^2}{396} \left(1 - \frac{396}{35,000}\right) (91.00)^2$$

$$\frac{N_2^2}{n_2} \left(1 - \frac{n_2}{N_2}\right) S_2^2$$

$$\frac{(45,000)^2}{356}\left(1-\frac{356}{45,000}\right)(63.53)^2$$

$$\frac{N_3^2}{n_3}\left(1-\frac{n_3}{N_3}\right)S_3^2$$

$$\frac{(20,000)^2}{248}\left(1-\frac{248}{20,000}\right)(99.49)^2$$

Summing these three terms and taking the square root:

Calculating:

$$T_1=\frac{(35,000)^2}{396}\left(1-\frac{396}{35,000}\right)(91.00)^2=25,326,894,797.98$$

$$T_2=\frac{(45,000)^2}{356}\left(1-\frac{356}{45,000}\right)(63.53)^2=22,776,307,940.68$$

$$T_3=\frac{(20,000)^2}{248}\left(1-\frac{248}{20,000}\right)(99.49)^2=15,766,970,443.16$$

Summing these:

$$T_1+T_2+T_3=25,326,894,797.98+22,776,307,940.68+15,766,970,443.16=63,870,173,181.82$$

Taking the square root:

$$SE_{Prop}(\hat{T}_{str})=\sqrt{63,870,173,181.82}\approx 257,435.2<257,447.4=SE_{Srs}(\hat{T}_{str})$$

SE under proportional allocation is smaller than the previously calculated SE, which is expected.

Problem 2 (10 pt)

Consider a simple random sample of size $n = 200$ from a finite population with size $N = 10,000$, measuring (X, Y), taking values on $\{(0,0),(0,1),(1,0),(1,1)\}$. The finite population has the following distribution.

	$X=1$	$X=0$	
$Y=1$	N_{11}	N_{10}	N_{1+}
$Y=0$	N_{01}	N_{00}	N_{0+}
	N_{+1}	N_{+0}	N

The population count N_{ij} are unknown.

Suppose that the realized sample has the following sample counts:

	$X=1$	$X=0$	
$Y=1$	70	30	100
$Y=0$	50	50	100
	120	80	200

1.

If it is known that $N_{+1} = N_{+0} = 5000$, how can you make use of this information to obtain a post-stratified estimator of $\theta = E(Y)$, using X as the post-stratification variable?

Answer

We can obtain a post-stratified estimator of $\theta = E(Y)$ via:

$$\hat{\theta} = W_1\bar{y}_1 + W_2\bar{y}_2 = 0.5 \cdot \left(\frac{70}{120}\right) + 0.5 \cdot \left(\frac{30}{80}\right) = 0.479$$

2.

If we are interested in estimating $\theta = P(Y = 1|X = 1)$, discuss how to estimate θ from the above sample and how to estimate its variance (Hint: Use Taylor expansion of ratio estimator to obtain the sampling variance).

Answer

For $\theta = P(Y = 1 \mid X = 1)$, we can make use of conditional probability, i.e.,

$$P(Y = 1 \mid X = 1) = \frac{P(Y = 1 \cap X = 1)}{P(X = 1))}$$

Substituting known quantities into this formula gives us:

$$\hat{\theta} = \frac{\hat{P}(X = 1, Y = 1)}{\hat{P}(X = 1)} = \frac{n_{11}}{n_{1+}} = \frac{70}{120} = 0.583$$

where $n_{1+} = n_{11} + n_{10}$.

Next, for variance estimation, we use the available hint and consider the Taylor expansion of the ratio estimator:

$$\hat{\theta} \approx \theta + \frac{1}{E(n_{1+})}(n_{11} - \theta n_{1+}) = \theta + \frac{1}{E(n_{1+})/n}\left(\frac{n_{11}}{n} - \theta\frac{n_{1+}}{n}\right) = \theta + \frac{1}{E(\hat{P}_{1+})}\left(\hat{P}_{11} - \theta\hat{P}_{1+}\right)$$

This gives us the variance of $\hat{\theta}$,

$$V(\hat{\theta}) \approx \frac{1}{P_{1+}^2}\left\{V(\hat{P}_{11}) + \theta^2 V(\hat{P}_{1+}) - 2\theta\text{Cov}(\hat{P}_{11}, \hat{P}_{1+})\right\}$$

Under SRS then, we know $V(\hat{P}_{11})$ is given by the expression:

$$V(\hat{P}_{11}) = \frac{1}{n}(1-f)P_{11}(1-P_{11})$$

And similarly:

$$V(\hat{P}_{1+}) = \frac{1}{n}(1-f)P_{1+}(1-P_{1+})$$

And also:

$$\text{Cov}(\hat{P}_{11}, \hat{P}_{1+}) = \frac{1}{n}(1-f)P_{11}(1-P_{1+})$$

Where f is defined as usual, i.e. $f = \frac{n}{N} = \frac{200}{10{,}000} = 0.02$.

As given, we know, $\hat{\theta} = \hat{P_{11}}/\hat{P_{1+}} \to \theta = P_{11}/P_{1+}$. So we have all quantities known to solve directly for $V(\hat{\theta})$! Simplifying:

$$V(\hat{\theta}) = \frac{1}{n}(1-f)\frac{1}{P_{1+}^2}\left\{P_{11}(1-P_{11}) + \frac{P_{11}^2}{P_{1+}^2}P_{1+}(1-P_{1+}) - 2\frac{P_{11}}{P_{1+}}P_{11}(1-P_{1+})\right\}$$

$$V(\hat{\theta}) = \frac{1}{n}(1-f)\frac{1}{P_{1+}}\left\{P_{11} - \frac{P_{11}^2}{P_{1+}}\right\} = \frac{1}{n}(1-f)\frac{1}{P_{1+}}\theta(1-\theta)$$

Using the above simplification, we can turn to the estimated variance formula:

$$\hat{V}(\hat{\theta}) = (1-f)\frac{1}{n_{1+}}\hat{\theta}(1-\hat{\theta})$$

Using $f = 0.02$, $n_{1+} = 120$, and $\hat{\theta} = 70/120$, we then have:

```
f <- 0.02
n1Plus <- 120
hatTheta <- 70/120

(1 - f) * (1 / n1Plus) * hatTheta * (1 - hatTheta)
```

```
## [1] 0.001984954
```

$$\hat{V}(\hat{\theta}) = 0.001985$$

Problem 3 (10 pt)

Suppose that we have a finite population of $(Y_{hi}(1), Y_{hi}(0))$ generated from the following superpopulation model:

$$\begin{pmatrix} Y_{hi}(0) \\ Y_{hi}(1) \end{pmatrix} \sim \left[\begin{pmatrix} \mu_{h0} \\ \mu_{h1} \end{pmatrix}, \begin{pmatrix} \sigma_{h0}^2 & \sigma_{h01} \\ \sigma_{h01} & \sigma_{h1}^2 \end{pmatrix} \right] \tag{1}$$

for $i = 1, \ldots, N_h$ and $h = 1, \ldots, H$. Instead of observing $(Y_{hi}(0), Y_{hi}(1))$, we observe $T_{hi} \in \{0, 1\}$ and

$$Y_{hi} = T_{hi} Y_{hi}(1) + (1 - T_{hi}) Y_{hi}(0)$$

The parameter of interest is the average treatment effect:

$$\tau = \sum_{h=1}^{H} W_h (\mu_{h1} - \mu_{h0}),$$

where $W_h = N_h / N$. The estimator is:

$$\hat{\tau}_{\text{sre}} = \sum_{h=1}^{H} W_h \hat{\tau}_h$$

where

$$\hat{\tau}_h = \frac{1}{N_{h1}} \sum_{i=1}^{N_h} T_{hi} Y_{hi} - \frac{1}{N_{h0}} \sum_{i=1}^{N_h} (1 - T_{hi}) Y_{hi}$$

1.

Compute the variance of $\hat{\tau}_{\text{sre}}$ using the model parameters in (1).

Answer

Under the given assumptions, by definition:

$$E(\hat{\tau}_{sre} \mid \mathcal{F}_N) = \sum_{h=1}^{H} W_h \bar{\tau}_h$$

where

$$\bar{\tau}_h = N_h^{-1} \sum_{i=1}^{N_h} \{Y_{hi}(1) - Y_{hi}(0)\}$$

Via EVVE, we can express the total variance as:

$$V(\hat{\tau}_{sre}) = V\{E(\hat{\tau}_{sre} \mid \mathcal{F}_N)\} + E\{V(\hat{\tau}_{sre} \mid \mathcal{F}_N)\}$$

For the first term, we note that:

$$V\{E(\hat{\tau}_{sre} \mid \mathcal{F}_N)\} = V\{\sum_{h}^{H} W_h \bar{\tau}_h\}$$

Also, noting the lecture slides, we know the second term of this expression too!

$$V(\hat{\tau}_{sre} \mid \mathcal{F}_N) = \sum_{h=1}^{H} W_h^2 \frac{1}{N_h} \left(\frac{N_{h0}}{N_{h1}} S_{h1}^2 + \frac{N_{h1}}{N_{h0}} S_{h0}^2 + 2S_{h01} \right)$$

Substituting these into our initial total variance formula:

$$V(\hat{\tau}_{sre}) = V\left\{ \sum_{h=1}^{H} W_h \bar{\tau}_h \right\} + E\left\{ \sum_{h=1}^{H} W_h^2 \frac{1}{N_h} \left(\frac{N_{h0}}{N_{h1}} S_{h1}^2 + \frac{N_{h1}}{N_{h0}} S_{h0}^2 + 2S_{h01} \right) \right\}$$

As given from our model, we know our variance-covariance matrix, such that we have a known formula for the firm term in our total variance equation:

$$V\left\{ \sum_{h=1}^{H} W_h \bar{\tau}_h \right\} = \sum_{h=1}^{H} W_h^2 \frac{1}{N_h} (\sigma_{h1}^2 + \sigma_{h0}^2 - 2\sigma_{h01})$$

Taking expectation, note that only the assignment of treatments is random (all else are considered fixed), giving us:

$$E\left\{ \sum_{h=1}^{H} W_h^2 \frac{1}{N_h} \left(\frac{N_{h0}}{N_{h1}} S_{h1}^2 + \frac{N_{h1}}{N_{h0}} S_{h0}^2 + 2S_{h01} \right) \right\} = \sum_{h=1}^{H} W_h^2 \frac{1}{N_h} \left(\frac{N_{h0}}{N_{h1}} \sigma_{h1}^2 + \frac{N_{h1}}{N_{h0}} \sigma_{h0}^2 + 2\sigma_{h01} \right)$$

So we have effectively simplified both parts of our total variance formula. We will find that a number of terms cancel when we bring them all back together, specifically:

$$V(\hat{\tau}_{sre}) = \sum_{h=1}^{H} W_h^2 \frac{1}{N_h} (\sigma_{h1}^2 + \sigma_{h0}^2 - 2\sigma_{h01}) + \sum_{h=1}^{H} W_h^2 \frac{1}{N_h} \left(\frac{N_{h0}}{N_{h1}} \sigma_{h1}^2 + \frac{N_{h1}}{N_{h0}} \sigma_{h0}^2 + 2\sigma_{h01} \right) = \sum_{h=1}^{H} W_h^2 \left(\frac{\sigma_{h1}^2}{N_{h1}} + \frac{\sigma_{h0}^2}{N_{h0}} \right)$$

Noting that:

$$\frac{1}{N_h} \sigma_{h1}^2 + \frac{1}{N_h} \frac{N_{h0}}{N_{h1}} \sigma_{h1}^2 = \frac{\sigma_{h1}^2}{N_{h1}}$$

As $N_h = N_{h1} + N_{h0}$.

2.

Assuming the model parameters in (1) are known, what is the optimal sample allocation such that $\text{Var}(\hat{tau}_{Sre}$ is minimized subject to $N_h = N_{h1} + N_{h0}$ for $h = 1, ..., H$ are fixed? That is, how to choose N_{h1} and N_{h0} for a given N_h?

Answer

We want to minimize:

$$Q(N_{h1}, N_{h0}) = \frac{\sigma_{h1}^2}{N_{h1}} + \frac{\sigma_{h0}^2}{N_{h0}}$$

subject to the constraint:

$$N_h = N_{h1} + N_{h0}$$

Given this is a constrained optimization problem, we consider using Lagrange:

$$\mathcal{L}(N_{h1}, N_{h0}, \lambda) = \frac{\sigma_{h1}^2}{N_{h1}} + \frac{\sigma_{h0}^2}{N_{h0}} + \lambda(N_{h1} + N_{h0} - N_h)$$

Taking partial derivatives, setting both zero:

$$\frac{\partial \mathcal{L}}{\partial N_{h1}} = -\frac{\sigma_{h1}^2}{N_{h1}^2} + \lambda = 0$$

$$\frac{\partial \mathcal{L}}{\partial N_{h0}} = -\frac{\sigma_{h0}^2}{N_{h0}^2} + \lambda = 0$$

From the first equation:

$$\lambda = \frac{\sigma_{h1}^2}{N_{h1}^2}$$

From the second equation:

$$\lambda = \frac{\sigma_{h0}^2}{N_{h0}^2}$$

Setting these two expressions equal to one another gives us:

$$\frac{\sigma_{h1}^2}{N_{h1}^2} = \frac{\sigma_{h0}^2}{N_{h0}^2}$$

Taking the square root on both sides to isolate $\sigma_{h1}, \sigma_h 0$:

$$\frac{\sigma_{h1}}{N_{h1}} = \frac{\sigma_{h0}}{N_{h0}}$$

Now we want to isolate N_{h1} and utilize our constraint:

$$N_{h1} = N_{h0}\frac{\sigma_{h1}}{\sigma_{h0}}$$

As we are constrained by $N_h = N_{h1} + N_{h0}$, we have: $N_{h0} = N_h - N_{h1}$:

$$N_{h1} = (N_h - N_{h1})\frac{\sigma_{h1}}{\sigma_{h0}}$$

Solving for N_{h1}:

$$N_{h1}\left(1 + \frac{\sigma_{h1}}{\sigma_{h0}}\right) = N_h\frac{\sigma_{h1}}{\sigma_{h0}}$$

$$N^*_{h1} = N_h\frac{\sigma_{h1}}{\sigma_{h1} + \sigma_{h0}}$$

Noting:

$$\frac{\sigma_{h1}}{\sigma_{h1}} = 1 \text{ as by definition } \sigma_{h1}, \sigma_{h0} > 0$$

With N^*_{h1} and our contraint $N_h = N_{h1} + N_{h0}$, we can now easily find N^*_{h0}:

$$N^*_{h0} = N_h - N^*_{h1} = N_h\frac{\sigma_{h0}}{\sigma_{h1} + \sigma_{h0}}$$

Problem 4 (10 pt)

Assume that a simple random sample of size n is selected from a population of size N and (x_i, y_i) are observed in the sample. In addition, we assume that the population mean of x, denoted by $\bar{X}$, is known.

1.

Use a Taylor linearization method to find the variance of the product estimator $\frac{\bar{x}\bar{y}}{\bar{X}}$, where $(\bar{x}, \bar{y})$ is the sample mean of (x_i, y_i).

Answer

The product estimator is:

$$\hat{\theta} = \frac{\bar{x}\bar{y}}{\bar{X}}$$

Via Taylor linearization, we approximate $\hat{\theta}$ using a first-order expansion around the true population means $\bar{X}$ and $\bar{Y}$:

$$\hat{\theta} \approx \frac{\bar{Y}\bar{X} + (\bar{x} - \bar{X})\bar{Y} + (\bar{y} - \bar{Y})\bar{X}}{\bar{X}}$$

Simplifying,

$$\hat{\theta} \approx \bar{Y} + (\bar{x} - \bar{X})\frac{\bar{Y}}{\bar{X}} + (\bar{y} - \bar{Y}) \approx \bar{y} + \left(\bar{x}\frac{\bar{Y}}{\bar{X}}\right) - \bar{Y}$$

Taking variances, and noting $\text{Var}(\bar{Y}) = 0$, we have:

$$V(\hat{\theta}) \approx V(\bar{y}) + \frac{\bar{Y}^2}{\bar{X}^2}V(\bar{x}) + 2\frac{\bar{Y}}{\bar{X}}\text{Cov}(\bar{x}, \bar{y})$$

Under SRS, we know:

$$V(\bar{x}) = \frac{S_x^2}{n}\left(1 - \frac{n}{N}\right)$$

$$V(\bar{y}) = \frac{S_y^2}{n}\left(1 - \frac{n}{N}\right)$$

$$\text{Cov}(\bar{x}, \bar{y}) = \frac{S_{xy}}{n}\left(1 - \frac{n}{N}\right)$$

Substituting the above relations into our variance formula gives us:

$$V(\hat{\theta}) \approx \left(\frac{S_y^2}{n} + \frac{\bar{Y}^2}{\bar{X}^2}\frac{S_x^2}{n} + 2\frac{\bar{Y}}{\bar{X}}\frac{S_{xy}}{n}\right)\left(1 - \frac{n}{N}\right)$$

2.

Find the condition that this product estimator has a smaller variance than the sample mean $\bar{y}$.

Answer

For $\hat{\theta}$ to be more efficient than $\bar{y}$, we require:

$$V(\hat{\theta}) < V(\bar{y})$$

As we already calculated $V(\hat{\theta})$, this means:

$$\frac{S_y^2}{n} + \frac{\bar{Y}^2}{\bar{X}^2}\frac{S_x^2}{n} + 2\frac{\bar{Y}}{\bar{X}}\frac{S_{xy}}{n} < \frac{S_y^2}{n}$$

We then just need to arrange terms to get at our desired conclusion. To that end we have:

$$\frac{\bar{Y}^2}{\bar{X}^2}\frac{S_x^2}{n} + 2\frac{\bar{Y}}{\bar{X}}\frac{S_{xy}}{n} < 0$$

$$\frac{\bar{Y}^2}{\bar{X}^2}S_x^2 + 2\frac{\bar{Y}}{\bar{X}}S_{xy} < 0$$

$$\frac{\bar{Y}}{\bar{X}}\left(\frac{\bar{Y}}{\bar{X}}S_x^2 + 2S_{xy}\right) < 0$$

For $\frac{\bar{Y}}{\bar{X}} > 0$:

$$\frac{\bar{Y}}{\bar{X}}S_x^2 + 2S_{xy} < 0$$

Giving us:

$$2S_{xy} < -\frac{\bar{Y}}{\bar{X}}S_x^2$$

And finally:

$$S_{xy} < -\frac{\bar{Y}}{2\bar{X}}S_x^2$$

So we need the random variables X and Y to be negatively correlated and meet the above condition to ensure greater efficiency for $\hat{\theta}$ comapred to $\bar{y}$.

3.

Prove that if the population covariance of x and y is zero, then the product estimator is less efficient than $\bar{y}$.

Answer

If the population covariance $S_{xy} = 0$, the variance formula for the product estimator $\hat{\theta}$ detailed in part 1 simplifies to:

$$V(\hat{\theta}) = V(\bar{y}) + \frac{\bar{Y}^2}{\bar{X}^2} V(\bar{x})$$

We know that $\frac{\bar{Y}^2}{\bar{X}^2} V(\bar{x}) > 0$, because each term in the equation is positive. Using this, we then know:

$$V(\hat{\theta}) > V(\bar{y})$$

Thus, when x and y are uncorrelated ($S_{xy} = 0$), the product estimator $\hat{\theta}$ is less efficient than $\bar{y}$. This also follows a similar argument using the results from part 2 as well.

Problem 5 (10 pt)

In a population of 10,000 businesses, we want to estimate the average sales $\bar{Y}$. For that, we sample $n = 100$ businesses using simple random sampling. Furthermore, we have at our disposal the auxiliary information "number of employees", denoted by x, for each business. It is known that $\bar{X} = 50$ in the population. From the sample, we computed the following statistics:

- $\bar{y}_n = 5.2 \times 10^6$ (average sales in the sample)
- $\bar{x}_n = 45$ employees (sample mean)
- $s_y^2 = 25 \times 10^{10}$ (sample variance of y_k)
- $s_x^2 = 15$ (sample variance of x_k)
- $r = 0.8$ (sample correlation coefficient between x and y)

Answer the following questions:

1.

Compute a 95% confidence interval for $\bar{Y}$ using the ratio estimator.

Answer

By definition, the ratio estimator for the population mean sales is given by:

$$\hat{\bar{Y}}_R = \bar{y}_n \frac{\bar{X}}{\bar{x}_n}$$

Calculating:

```
n <- 100
bary <- 5.2e6
barx <- 45
barX <- 50

barYratio <- bary * barX / barx
barYratio
```

```
## [1] 5777778
```

$$\hat{\bar{Y}}_R = (5.2 \cdot 10^6) \cdot \frac{50}{45} = 5.778e6$$

The variance of the ratio estimator is approximated by the formula:

$$V(\hat{\bar{Y}}_R) \approx \bar{Y}^2 \left(\frac{1}{n}\right) \left(\frac{s_y^2}{\bar{y}_n^2} + \frac{s_x^2}{\bar{x}_n^2} - 2r \frac{s_y}{\bar{y}_n} \frac{s_x}{\bar{x}_n}\right)$$

Calculating:

```
r <- 0.8
barY <-
sy2 <- 25e10
sx2 <- 15
barY <- bary * barX / barx

involved <- (sy2 / bary^2) + (sx2/barx^2) - (2 * r * (sqrt(sy2) / bary) * (sqrt(sx2) / barx))
varRatio <- barY^2 * (1/n) * involved
varRatio
```

```
## [1] 1139018550
```

$$V(\hat{\bar{Y}}_R) = (5.778 \cdot 10^6)^2 \cdot \frac{1}{100}\left(\frac{25 \cdot 10^{10}}{(5.2 \cdot 10^6)^2} + \frac{15}{45^2} - 2(0.8)\frac{5 \cdot 10^5}{5.2 \cdot 10^6}\frac{3.873}{45}\right) = 1.139018550e9 \approx 1.139e9$$

From this we can calculate the standard error, and use that value to then calculate the 95% confidence interval. Doing so:

```
interval <- qnorm(0.975)
ue <- barYratio + (interval * sqrt(varRatio))
le <- barYratio - (interval * sqrt(varRatio))
interval <- c(le, ue)
interval
```

```
## [1] 5711630 5843925
```

$$\hat{\bar{Y}}_R \pm 1.96 \cdot \sqrt{V(\hat{\bar{Y}}_R)} = 5.778e6 \pm \sqrt{1.139e9} \rightarrow (5.719e6, 5.844e6)$$

2.

Compute a 95% confidence interval for $\bar{Y}$ using the regression estimator based on the simple linear regression of y on x (with intercept).

Answer

For the regression estimator, we use the known formula for the population mean given by:

$$\hat{\bar{Y}}_{reg} = \bar{y}_n + b(\bar{X} - \bar{x}_n)$$

where the estimated slope b is given by:

$$b = r\frac{s_y}{s_x}$$

```
b <- r * sqrt(sy2) / sqrt(sx2)
b
```

```
## [1] 103279.6
```

Substituting the values:

$$b = 0.8 \cdot \frac{5 \cdot 10^5}{3.873} = 1.033e5$$

Thus, the regression estimator is:

```
barYreg <- bary + b * (barX - barx)
barYreg
```

```
## [1] 5716398
```

$$\hat{\bar{Y}}_{reg} = \bar{y}_n + b(\bar{X} - \bar{x}_n) = (5.2e6) + (1.033e5)(50 - 45) = 5.717e6$$

The variance of the regression estimator is then given by the formula:

$$V(\hat{\bar{Y}}_{reg}) = \frac{s_y^2}{n}(1 - r^2)$$

```
vYreg <- (sy2 / n) * (1 - r^2)
vYreg
```

```
## [1] 9e+08
```

Evaluating this quantity:

$$V(\hat{\bar{Y}}_{reg}) = \frac{25 \cdot 10^{10}}{100}(1 - 0.64) = 9e9$$

Again, we can then calculate the standard error, and use this value to calculate the 95% confidence interval:

```r
interval <- qnorm(0.975)
ue2 <- barYreg + (interval * sqrt(vYreg))
le2 <- barYreg - (interval * sqrt(vYreg))
interval2 <- c(le2, ue2)
interval2
```

```
## [1] 5657599 5775197
```

$$\hat{\bar{Y}}_{reg} \pm 1.96 \cdot \sqrt{V(\hat{\bar{Y}}_{reg})} \rightarrow (5.658e6, 5.775e6)$$

As a note, the interval constructed using the regression estimator is more precise than the one constructed using the ratio estimator (smaller margin of error).

Problem 6 (10 pt)

Under the setup of Chapter 6, Part 1 lecture, prove the last two equalities on page 23:

$$\text{Cov}\left(\frac{1}{N_1}\sum_{i=1}^{N} T_i e_i(1), \frac{1}{N_0}\sum_{i=1}^{N}(1-T_i)\mathbf{x}_i'\mathbf{B}_0|\mathcal{F}_N\right) = 0 \tag{1}$$

$$\text{Cov}\left(\frac{1}{N_0}\sum_{i=1}^{N}(1-T_i) e_i(0), \frac{1}{N_0}\sum_{i=1}^{N}(1-T_i)\mathbf{x}_i'\mathbf{B}_0|\mathcal{F}_N\right) = 0 \tag{2}$$

Answer

Setup, a note on known quantities and what they refer to

- T_i is the treatment indicator for unit i (1 received treatment, 0 is control).
- $N_1 = \sum_{i=1}^{N} T_i$ is the number (size of sample) of treated units.
- $N_0 = \sum_{i=1}^{N}(1-T_i)$ is the number (size of sample) of control units.
- $e_i(1)$ and $e_i(0)$ are error terms for treatment and control, respectively.
- $\mathbf{x}_i$ is the vector of covariates for unit i (including intercept).
- $\mathbf{B}_0$ is a fixed coefficient vector.

Proof

Since treatment assignments T_i are independent of errors and covariates, we have:

For treatment:

$$E[T_i e_i(1)|\mathcal{F}_N] = \pi_i e_i(1)$$

And for control:

$$E[(1-T_i)e_i(0)|\mathcal{F}_N] = (1-\pi_i)e_i(0)$$

Additionally, for treatment:

$$E[T_i\mathbf{x}_i'\mathbf{B}_0|\mathcal{F}_N] = \pi_i\mathbf{x}_i'\mathbf{B}_0$$

And similarly, for control:

$$E[(1-T_i)\mathbf{x}_i'\mathbf{B}_0|\mathcal{F}_N] = (1-\pi_i)\mathbf{x}_i'\mathbf{B}_0$$

where $\pi_i = P(T_i = 1)$ (first order inclusion probability of being assigned Treatment).

Equation 1

For Equation (1), the first of the quantities to prove:

$$\text{Cov}\left(\frac{1}{N_1}\sum_{i=1}^{N} T_i e_i(1), \frac{1}{N_0}\sum_{i=1}^{N}(1-T_i)\mathbf{x}_i'\mathbf{B}_0|\mathcal{F}_N\right)$$

Using the property of linearity, we may write this as equal to:

$$\frac{1}{N_1 N_0}\sum_{i=1}^{N}\sum_{j=1}^{N}\text{Cov}\left(T_i e_i(1), (1-T_j)\mathbf{x}_j'\mathbf{B}_0|\mathcal{F}_N\right)$$

Cross terms are 0 for $i = j$, and for $i \neq j$, T_i and $(1 - T_j)$ are independent (independent treatments):

As a quick note, the cross terms being zero stems from the underlying implication of independence, i.e., generally:

$$Cov(X, Y) = E[XY] - E[X]E[Y]$$

But via independence:

$$E[XY] = E[X]E[X] \rightarrow Cov(X, Y) = 0$$

Continuing on, we are left with, for $i \neq j$,

$$\text{Cov}(T_i e_i(1), (1-T_i)\mathbf{x}_i'\mathbf{B}_0|\mathcal{F}_N) = E[T_i(1-T_i)|\mathcal{F}_N]E[e_i(1)\mathbf{x}_i'\mathbf{B}_0|\mathcal{F}_N]$$

Then, since $T_i(1 - T_i) = 0$, we may write:

$$\text{Cov}\left(\frac{1}{N_1}\sum_{i=1}^{N} T_i e_i(1), \frac{1}{N_0}\sum_{i=1}^{N}(1-T_i)\mathbf{x}_i'\mathbf{B}_0|\mathcal{F}_N\right) = E[T_i(1-T_i)|\mathcal{F}_N]E[e_i(1)\mathbf{x}_i'\mathbf{B}_0|\mathcal{F}_N] = 0$$

Equation 2

For the second equation, (2):

$$\text{Cov}\left(\frac{1}{N_0}\sum_{i=1}^{N}(1-T_i)e_i(0), \frac{1}{N_0}\sum_{i=1}^{N}(1-T_i)\mathbf{x}_i'\mathbf{B}_0|\mathcal{F}_N\right)$$

Similarly expanding the above with note of linearity, we have:

$$\frac{1}{N_0^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\text{Cov}\left((1-T_i)e_i(0), (1-T_j)\mathbf{x}_j'\mathbf{B}_0|\mathcal{F}_N\right)$$

By contrast with the prior argument used for (1), for $i \neq j$, the terms are independent and evaluate to zero. However, for $i = j$ we have:

$$\text{Cov}\left((1-T_i)e_i(0), (1-T_i)\mathbf{x}_i'\mathbf{B}_0|\mathcal{F}_N\right) = 0$$

And this relies again upon the base definition of covariance and the fact we have independence, i.e.,

$$\text{Cov}\left((1-T_i)e_i(0),(1-T_i)\mathbf{x}_i'\mathbf{B}_0|\mathcal{F}_N\right) = (1-T_i)^2\text{Cov}\left(e_i(0),\mathbf{x}_i'\mathbf{B}_0|\mathcal{F}_N\right)$$

And we have independence such that:

$$E[e_i(0)\mathbf{x}_i'\mathbf{B}_0|\mathcal{F}_N] = 0 \rightarrow \text{Cov}\left((1-T_i)e_i(0),(1-T_i)\mathbf{x}_i'\mathbf{B}_0|\mathcal{F}_N\right) = 0$$

Problem 7 (10 pt)

Under the setup of Chapter 6, Part 2 lecture:

1.

Prove Lemma 3.

Lemma 3:

Setup

Let X be a $n \times p$ matrix such that

$$X = \begin{pmatrix} \mathbf{x_1}' \\ \vdots \\ \mathbf{x_n}' \end{pmatrix}$$

and $\omega = (\omega_1, \ldots, \omega_n)'$ be an n-dimensional weight vector ($n = N_1$). Given

$$\bar{\mathbf{x}} = N^{-1} \sum_{i=1}^{N} \mathbf{x}_i',$$

and D ($p \times p$ symmetric, invertible matrix), the minimizer of

$$Q(\omega) = \gamma(\omega' X - \bar{\mathbf{x}})' D(\omega' X - \bar{\mathbf{x}}) + \omega'\omega = \gamma(X'\omega - \bar{\mathbf{x}})' D(X'\omega - \bar{\mathbf{x}}) + \omega'\omega$$

is given by

$$\hat{\omega} = (\gamma X D X' + I_n)^{-1} \gamma X D \bar{\mathbf{x}} \tag{10}$$

$$= X(X'X + \gamma^{-1} D^{-1})^{-1} \bar{\mathbf{x}} \tag{11}$$

The goal of this problem is to derive (10), where:

- X is an $n \times p$ matrix of covariates,
- ω is an n-dimensional weight vector,
- $\bar{x} = N^{-1} \sum_{i=1}^{N} x_i'$ is the population mean of the covariates,
- D is a $p \times p$ symmetric, positive definite matrix,
- γ is a scalar tuning parameter.

Answer

To derive (10), we start my minimizing the function:

$$Q(\omega) = \gamma(X'\omega - \bar{\mathbf{x}})'D(X'\omega - \bar{\mathbf{x}}) + \omega'\omega$$

In typical fashion, we take the first derivative of $Q(\omega)$ with respect to ω, i.e.:

$$\frac{dQ}{d\omega} = 2\gamma XD(X'\omega - \bar{x}) + 2\omega$$

If we take the second derivative we find:

$$\frac{dQ^2}{d\omega^2} = 2\gamma XDX' + 2I_n > 0$$

Noting that $\mathbf{XDX'}$ is positive-semidefinite, we know that $Q(\omega)$ is strictly convex, allowing us to proceed knowing the first derivative (gradient) will lead us to the global minimum.

To that end we proceed by setting the first derivative to zero, giving us the first-order condition:

$$\gamma XDX'\omega - \gamma XD\bar{\mathbf{x}} + \omega = 0$$

Rearranging terms, isolating ω gives us:

$$(\gamma XDX' + I_n)\omega = \gamma XD\bar{\mathbf{x}} \rightarrow \hat{\omega} = (\gamma XDX' + I_n)^{-1}\gamma XD\bar{\mathbf{x}}$$

This gives us equation (10):

$$\hat{\omega} = (\gamma XDX' + I_n)^{-1}\gamma XD\bar{\mathbf{x}} \tag{10}$$

I believe that is sufficient for what is being asked, i.e. that we aren't being asked to provide proof for (11), but on the off-chance that needs coverage too:

To proceed to (11), we need to utilize the Woodbury identity, allowing us to rewrite the inverse term as:

Woodbury Identity:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

Where:

- A, U, C and V are conformable matrices:

- A is $n \times n$
- C is $k \times k$
- U is $n \times k$
- V is $k \times n$

Shout Out to Wikipedia for this one.

In Application to Lemma 3:

$$\hat{\omega} = (\gamma XDX' + I_n)^{-1}\gamma XD = X(X'X + \gamma^{-1}D^{-1})^{-1}$$

And thus we have Lemma 3 in its entirety.

2.

Show that the final weight in (13) satisfies a hard calibration for $\mathbf{x}_1$:

$$\sum_{i \in A} \hat{\omega}_i \mathbf{x}_{1i} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_{1i}.$$

Setup

(9):

The implicit model is that

$$Y(1) = x_1' \beta + x_2' u + e(1) \tag{9}$$

where $u \sim (0, D_q \sigma_u^2)$ with known D_q and $e(1) \sim (0, \sigma_e^2)$.

(10-11): Given in Lemma 3

(12):

Using (11), the solution can be written as

$$\hat{\omega} = X \left(X'X + \Omega^{-1} \right)^{-1} \bar{\mathbf{x}} \tag{12}$$

where $\Omega^{-1} = \text{Diag}\{\gamma_1^{-1} D_p^{-1}, \gamma_2^{-1} D_q^{-1}\}$ and $\gamma_1 \to \infty$.

(13):

Under the mixed model setup in (9), the solution (12) can be written as

$$\hat{\omega}_i = \left(N^{-1} \sum_{i=1}^{N} \mathbf{x_i} \right)' \left\{ \sum_{i=1}^{N} T_i \mathbf{x}_i \mathbf{x}_i' + \Omega^{-1} \right\}^{-1} x_i, \tag{13}$$

Where:

- $\Omega^{-1} = \text{Diag}\{0_p, \gamma_2^{-1} D_q^{-1}\}$
- $\gamma_2 = \sigma_u^2 / \sigma_e^2$

The goal then is to take Equation (13)

$$\hat{\omega}_i = \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i \right)' \left\{ \sum_{i=1}^{N} T_i \mathbf{x}_i \mathbf{x}_i' + \Omega^{-1} \right\}^{-1} \mathbf{x}_i$$

And prove:

$$\sum_{i \in A} \hat{\omega}_i \mathbf{x}_{1i} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_{1i}$$

Answer

To start we note that the weights in equation (13) are of the form:

$$\hat{\omega}_i = \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i \right)' \left\{ \sum_{i=1}^{N} T_i \mathbf{x}_i \mathbf{x}_i' + \Omega^{-1} \right\}^{-1} \mathbf{x}_i$$

Where:

- $\Omega^{-1} = \text{Diag}\{0_p, \gamma_2^{-1} D_q^{-1}\}$
- $\gamma_2 = \sigma_u^2 / \sigma_e^2$

Using this equation, we then know:

$$\sum_{i \in A} \hat{\omega}_i \mathbf{x}_{1i} = \sum_{i \in A} \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i \right)' \left\{ \sum_{i=1}^{N} T_i \mathbf{x}_i \mathbf{x}_i' + \Omega^{-1} \right\}^{-1} \mathbf{x}_i \mathbf{x}_{1i} = \sum_{i=1}^{N} N^{-1} \left(\sum_{i=1}^{N} x_i \right)' \left(\sum_{i=1}^{N} T_i x_i x_i' + \Omega^{-1} \right)^{-1} x_i x_{i1}$$

$$\sum_{i \in A} \hat{\omega}_i \mathbf{x}_{1i} = N^{-1} \left(\sum_{i=1}^{N} x_i \right)' \left[\left(\sum_{i=1}^{N} T_i x_i x_i' + \Omega^{-1} \right)^{-1} \sum_{i=1}^{N} x_i x_{i1} \right]$$

We know that:

$$S = \sum_{i=1}^{N} x_j x_j'$$

So we may rewrite the above as:

$$\sum_{i \in A} \hat{\omega}_i \mathbf{x}_{1i} = N^{-1} \left(\sum_{j=1}^{N} x_j \right)' \left(\sum_{j=1}^{N} T_j S + \Omega^{-1} \right)^{-1} \sum_{i=1}^{N} x_i x_{i1}$$

Using generalized regression estimators we can show

$$\left(\sum_{i=1}^{N} T_j S + \Omega^{-1} \right)^{-1} S \approx I$$

For small Ω (large γ_2). For our purposes, I take the above as a given.

Using this, we then know:

$$\sum_{i=1}^{N} \hat{\omega}_i x_{i1} \approx N^{-1} \left(\sum_{j=1}^{N} x_j \right)' I \sum_{i=1}^{N} x_i x_{i1} = N^{-1} \sum_{i=1}^{N} x_{i1}'$$

Descriptively, we're saying that:

- Since $\Omega^{-1} = \text{Diag}\{0_p, \gamma_2^{-1} D_q^{-1}\}$, the term Ω^{-1} does not affect the calibration for $\mathbf{x}_{1i}$

- For large γ_2, the term $\gamma_2^{-1} D_p^{-1}$ is negligible, and the expression simplifies to:

So we conclude:

$$\sum_{i \in A} \hat{\omega}_i \mathbf{x}_{1i} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_{1i}$$

Which means our final weights $\hat{\omega}$ satisfy the hard calibration condition for $\mathbf{x}_{1i}$, as desired.

Note: Particularly for this problem, I attempted to be as precise as possible with matrix notation as used in the slides. There may be some discrepancies.