STAT 521: Midterm Exam   Name:

Problem 1: (30 pts)

We are interested in estimating the proportion of international students at ISU with population size $N = 10,000$. Suppose that we have a simple random sample of size $n = 100$ and the result is as follows.

	Male	Female
International	30	20
American	20	30

It is known that the population proportion of male students is 60 %.

(a) Compute the confidence interval for the proportion of international students at ISU using the post-stratified estimator with gender being the poststratum.

Solution: Post-stratification estimator

$$\hat{\bar{Y}}_{\text{post}} = 0.6 \times 0.6 + 0.4 \times 0.4 = 0.52$$

Estimated variance

$$\widehat{V}\left(\hat{\bar{Y}}_{\text{post}}\right) = \frac{1}{99} \times \left(1 - \frac{1}{100}\right) \{0.6 \times 0.6 \times 0.4 + 0.4 \times 0.4 \times 0.6\} = 0.0024$$

Thus, the 95% confidence interval is

$$\left(0.52 - 1.96 \times \sqrt{0.0024}, 0.52 + 1.96 \times \sqrt{0.0024}\right) = (0.424, 0.616)$$

(b) Compute the reduction of variance due to poststratum compared to the HT estimator in the estimation of the proportion of international students at ISU.

Solution:

$$\hat{V}\left(\widehat{\bar{Y}}_{\mathrm{HT}}\right)=\frac{1}{n}\left(1-\frac{n}{N}\right)s^2=\frac{1}{100}\times\left(1-\frac{1}{100}\right)\times\frac{100}{99}\times 0.5\times 0.5=0.0025.$$

Thus,

$$\frac{\hat{V}(\widehat{\bar{Y}}_{\mathrm{post}})}{\hat{V}(\widehat{\bar{Y}}_{\mathrm{HT}})}=\frac{0.0024}{0.0025}=0.96$$

(c) Suppose that we are interested in estimating $\theta = \theta_1 - \theta_2$, where θ_1 is the proportion of international students among male students and θ_2 is the proportion of international students among female students. Using the notation in the following table, $\theta_1 = N_{11}/N_{+1}$ and $\theta_2 = N_{12}/N_{+2}$. Using the data in Table 1, estimate θ and its variance. (May ignore the finite population correction term.)

	Sample		Population	
	Male	Female	Male	Female
International	n_{11}	n_{12}	N_{11}	N_{12}
American	n_{21}	n_{22}	N_{21}	N_{22}
Total	n_{+1}	n_{+2}	N_{+1}	N_{+2}

Solution: Point estimator is easy to compute:

$$\hat{\theta} = \hat{\theta}_1 - \hat{\theta}_2 = \frac{30}{50} - \frac{20}{50} = 0.2$$

Now,

$$\begin{aligned} V\left(\hat{\theta}_1\right) &= V\left(\frac{n_{11}}{n_{+1}}\right) \\ &\doteq V\left\{\frac{1}{nP_{+1}}\left(n_{11} - \theta_1 n_{+1}\right)\right\} \\ &= \left(\frac{1}{nP_{+1}}\right)^2 n\left\{P_{11}\left(1-P_{11}\right) - 2\theta_1 P_{11}\left(1-P_{+1}\right) + \theta_1^2 P_{+1}(1-P_{+1})\right\} \\ &= \frac{1}{nP_{+1}}\theta_1\left(1-\theta_1\right), \end{aligned}$$

where $P_{ij} = N_{ij}/N$. Thus,

$$\hat{V}\left(\hat{\theta}_1\right) = \frac{1}{50} \times 0.6 \times 0.4 = 0.0048.$$

Solution: Similarly, we can obtain

$$V\left(\hat{\theta}_2\right) = \frac{1}{nP_{+2}}\theta_2\left(1-\theta_2\right)$$

which leads to

$$\hat{V}\left(\hat{\theta}_2\right) = \frac{1}{50} \times 0.6 \times 0.4 = 0.0048.$$

Also,

$$\begin{aligned} Cov\left(\hat{\theta}_1, \hat{\theta}_2\right) &\doteq Cov\left\{\frac{1}{nP_{+1}}\left(n_{11}-\theta_1 n_{+1}\right), \frac{1}{nP_{+2}}\left(n_{12}-\theta_2 n_{+2}\right)\right\} \\ &= \frac{n}{(nP_{+1})(nP_{+2})}\left[(-P_{11}P_{12}) + \theta_1 P_{+1}P_{12} + \theta_2 P_{11}P_{+2} - \theta_1\theta_2 P_{+1}P_{+2}\right] \\ &= 0. \end{aligned}$$

Therefore,

$$\hat{V}\left(\hat{\theta}\right) \doteq \frac{1}{50} \times 0.6 \times 0.4 + \frac{1}{50} \times 0.4 \times 0.6 = 0.0096.$$

Problem 2: (20 pts)

Let x_1, x_2, x_3, x_4, x_5 be the five sample observations from SRS and we observed that $x_k = k$ in the sample. For this sample, we wish to assign the weights such that $\sum_{i=1}^{5} w_i = 1$ and $\sum_{i=1}^{5} w_i x_i = 4$. To uniquely determine w_i's, suppose that we want to minimize

$$\sum_{i=1}^{n} \left(w_i - \frac{1}{n} \right)^2$$

subject to $\sum_{i=1}^{n} w_i = 1$ and $\sum_{i=1}^{n} w_i x_i = 4$, where $n = 5$. Find the resulting weights.

Solution: Note that $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i = 3$ and $\sum_{i=1}^{n} (x_i - \bar{x})^2 = 10$. Thus, the final weight is the regression weight applied to $\mu_x = 4$. Thus,

$$w_i = \frac{1}{n} + (4 - \bar{x}) \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} (x_i - \bar{x})$$

which leads to

$$w_k = 0.2 + (k - 3)/10,$$

for $k = 1, 2, 3, 4, 5$.

Problem 3: (30 pts)

Consider the following potential outcome model:

$$
\begin{aligned}
Y_i(1) &= \mathbf{x}_i'\boldsymbol{\beta}_1 + e_i(1) \\
Y_i(0) &= \mathbf{x}_i\boldsymbol{\beta}_0 + e_i(0)
\end{aligned}
$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are unknown parameters, $e_i(1)$ and $e_i(0)$ are independent of $\mathbf{x}_i$ and $E\{e_i(1)\} = E\{e_i(0)\} = 0$. We obtain a realization of finite population of size N from the above model and observe $(\mathbf{x}_i, T_i, Y_i)$ from the sample selected by completely randomized experiment, where

$$
Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0).
$$

We are interested in estimating $\theta = E\{Y(1) - Y(0) \mid T = 1\}$, which is often called the average treatment effect on the treated (ATT). Note that the finite-population parameter for θ is

$$
\bar{\theta}_N = \frac{1}{N_1} \sum_{i=1}^{N} T_i \left\{ Y_i(1) - Y_i(0) \right\}.
$$

We are interested in using a weighted version of the estimator given by

$$
\hat{\theta}_\omega = \sum_{i=1}^{N} T_i \omega_{1i} Y_i - \sum_{i=1}^{N} (1 - T_i) \omega_{0i} Y_i. \tag{1}
$$

Answer the following questions.

(a) Find the conditions on the weights (ω_{1i} and ω_{0i}) such that $\hat{\theta}_\omega$ in (1) is model-unbiased for $\bar{\theta}_N$.

Solution: Using the superpopulation model, we can express

$$
\begin{aligned}
\hat{\theta}_\omega - \bar{\theta}_N &= \sum_{i=1}^{N} T_i \omega_{1i} Y_i(1) - \sum_{i=1}^{N} (1 - T_i) \omega_{0i} Y_i(0) - N_1^{-1} \sum_{i=1}^{N} T_i Y_i(1) + N_1^{-1} \sum_{i=1}^{N} T_i Y_i(0) \\
&= \left(\sum_{i=1}^{N} T_i \omega_{1i} \mathbf{x}_i - N_1^{-1} \sum_{i=1}^{N} T_i \mathbf{x}_i \right)' \boldsymbol{\beta}_1 - \left(\sum_{i=1}^{N} (1 - T_i) \omega_{0i} \mathbf{x}_i - N_1^{-1} \sum_{i=1}^{N} T_i \mathbf{x}_i \right)' \boldsymbol{\beta}_0 \\
&+ \sum_{i=1}^{N} T_i \omega_{1i} e_i(1) - N_1^{-1} \sum_{i=1}^{N} T_i e_i(1) - \left\{ \sum_{i=1}^{N} (1 - T_i) \omega_{0i} e_i(0) - N_1^{-1} \sum_{i=1}^{N} T_i e_i(0) \right\}
\end{aligned}
$$

Solution: Thus,

$$E\left\{\hat{\theta}_\omega - \bar{\theta}_N \mid \mathbf{X}, \mathbf{T}\right\} = \left(\sum_{i=1}^N T_i\omega_{1i}\mathbf{x}_i - N_1^{-1}\sum_{i=1}^N T_i\mathbf{x}_i\right)' \boldsymbol{\beta}_1 - \left(\sum_{i=1}^N (1-T_i)\omega_{0i}\mathbf{x}_i - N_1^{-1}\sum_{i=1}^N T_i\mathbf{x}_i\right)' \boldsymbol{\beta}_0$$

which gives

$$\sum_{i=1}^N T_i\omega_{1i}\mathbf{x}_i - N_1^{-1}\sum_{i=1}^N T_i\mathbf{x}_i = \mathbf{0}$$

and

$$\sum_{i=1}^N (1-T_i)\omega_{0i}\mathbf{x}_i - N_1^{-1}\sum_{i=1}^N T_i\mathbf{x}_i = \mathbf{0}$$

as a sufficient condition for model-unbiasedness of $\hat{\theta}_\omega$.

(b) Find the optimal estimator that minimizes the model variance of $\hat{\theta}_\omega$ in (1) subject to the model-unbiasedness condition in (a). (May assume that $V\{e_i(1)\} = \sigma_1^2$ and $V\{e_i(0)\} = \sigma_0^2$.)

Solution: The model variance of $\hat{\theta}_\omega$ is

$$V\left\{\hat{\theta}_\omega \mid \mathbf{X}, \mathbf{T}\right\} = \sum_{i=1}^{N} T_i \omega_{1i}^2 \sigma_1^2 + \sum_{i=1}^{N} (1 - T_i)\omega_{0i}^2 \sigma_0^2.$$

Thus, the optimal weights are obtained by minimizing

$$Q(\omega_1, \omega_0) = \sum_{i=1}^{N} T_i \omega_{1i}^2 + \sum_{i=1}^{N} (1 - T_i)\omega_{0i}^2$$

subject to

$$\sum_{i=1}^{N} T_i \omega_{1i} \mathbf{x}_i - N_1^{-1} \sum_{i=1}^{N} T_i \mathbf{x}_i = \mathbf{0}$$

and

$$\sum_{i=1}^{N} (1 - T_i)\omega_{0i} \mathbf{x}_i - N_1^{-1} \sum_{i=1}^{N} T_i \mathbf{x}_i = \mathbf{0}.$$

The solution is

$$\hat{\omega}_{1i} = \frac{1}{N_1}$$

and

$$\hat{\omega}_{0i} = \left(N_1^{-1} \sum_{i=1}^{N} T_i \mathbf{x}_i \right)' \left(\sum_{i=1}^{N} (1 - T_i) \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \mathbf{x}_i.$$

Therefore, the optimal estimator of θ is

$$\begin{aligned}
\hat{\theta}_{\text{opt}} &= \frac{1}{N_1} \sum_{i=1}^{N} T_i Y_i - \sum_{i=1}^{N} (1 - T_i) \hat{\omega}_{0i} Y_i \\
&= \frac{1}{N_1} \sum_{i=1}^{N} T_i Y_i - \frac{1}{N_1} \sum_{i=1}^{N} T_i \mathbf{x}_i' \hat{\boldsymbol{\beta}}_0
\end{aligned}$$

where

$$\hat{\boldsymbol{\beta}}_0 = \left(\sum_{i=1}^{N} (1 - T_i) \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{N} (1 - T_i) \mathbf{x}_i Y_i.$$

(c) Show that the optimal estimator in (b) is asymptotically design unbiased for $\bar{\tau}$, where

$$\bar{\tau} = \frac{1}{N} \sum_{i=1}^{N} \{Y_i(1) - Y_i(0)\}.$$

(May assume that $\mathbf{x}$ includes 1.)

Solution: We can express

$$\hat{\theta}_{\text{opt}} = \hat{\theta}_{1,\text{opt}} - \hat{\theta}_{2,\text{opt}}$$

where $\hat{\theta}_{1,\text{opt}} = N_1^{-1} \sum_{i=1}^{N} T_i Y_i(1)$ and $\hat{\theta}_{2,\text{opt}} = N_1^{-1} \sum_{i=1}^{N} T_i \mathbf{x}_i' \hat{\boldsymbol{\beta}}_0$. By the property of CRE, we obtain

$$E\left(\hat{\theta}_{1,\text{opt}} \mid \mathcal{F}_N\right) = \frac{1}{N} \sum_{i=1}^{N} Y_i(1).$$

Also, by the definition of $\hat{\boldsymbol{\beta}}_0$, we can express

$$\hat{\theta}_{2,\text{opt}} = \frac{1}{N_1} \sum_{i=1}^{N} T_i \mathbf{x}_i' \hat{\boldsymbol{\beta}}_0 + \frac{1}{N_0} \sum_{i=1}^{N} (1 - T_i) \left(Y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_0\right).$$

Now, by the property of the CRE again, we obtain

$$\begin{aligned} E\left(\hat{\theta}_{2,\text{opt}} \mid \mathcal{F}_N\right) &\doteq \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{B}_0 + \frac{1}{N} \sum_{i=1}^{N} \left\{Y_i(0) - \mathbf{x}_i' \mathbf{B}_0\right\} \\ &= \frac{1}{N} \sum_{i=1}^{N} Y_i(0) \end{aligned}$$

Therefore, combining the two, we obtain

$$E\left\{\hat{\theta}_{\text{opt}} \mid \mathcal{F}_N\right\} \doteq \frac{1}{N} \sum_{i=1}^{N} Y_i(1) - \frac{1}{N} \sum_{i=1}^{N} Y_i(0).$$