STAT 521: Homework 5 Solution

Problem 1:

Consider a finite population of size N with measurement $(\mathbf{x}, y)$.

We are interested in estimating $Y = \sum_{i=1}^N y_i$ using the linear estimator of the form $\hat{Y}_\omega = \sum_{i \in A} \omega_i y_i$. We wish to impose the following calibration constraints to the final weights:

$$\sum_{i \in A} \omega_i \mathbf{h}(\mathbf{x}_i) = \sum_{i=1}^N \mathbf{h}(\mathbf{x}_i) \tag{1}$$

where $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), \ldots, h_L(\mathbf{x})]$ is a known function of $\mathbf{x}$.

To uniquely determine ω_i, we consider minimizing

$$Q(\boldsymbol{\omega}) = \sum_{i \in A} \left(\omega_i - \pi_i^{-1}\right)^2 q_i$$

subject to (1) where $q_i = q(\mathbf{x}_i)$ is a known function of $\mathbf{x}_i$ and π_i is the first-order inclusion probability of unit i. Let $\hat{Y}_{\text{cal}} = \sum_{i \in A} \hat{\omega}_i y_i$ be the calibration estimator of Y using $\hat{\omega}_i$ obtained from the above calibration problem.

Answer the following questions.

1. Find the closed-form expression for $\hat{\omega}_i$.

Solution: Using Lagrange multiplier method, we minimize

$$Q(\boldsymbol{\omega}, \boldsymbol{\lambda}) = \frac{1}{2} \sum_{i \in A} \left(\omega_i - \pi_i^{-1}\right)^2 q_i - \boldsymbol{\lambda}' \left(\sum_{i \in A} \omega_i \mathbf{h}(\mathbf{x}_i) - \sum_{i=1}^N \mathbf{h}(\mathbf{x}_i) \right)$$

with respect to $\boldsymbol{\omega}$ and $\boldsymbol{\lambda}$. Now, solving $\frac{\partial}{\omega_i} Q = 0$ leads to

$$w_i = \pi_i^{-1} + \boldsymbol{\lambda}' \mathbf{h}(\mathbf{x}_i)/q_i.$$

Now, by the calibration equation, we obtain

$$\hat{\boldsymbol{\lambda}}' = \left(\sum_{i=1}^N \mathbf{h}(\mathbf{x}_i) - \sum_{i \in A} \pi_i^{-1} \mathbf{h}(\mathbf{x}_i) \right)' \left\{ \sum_{i \in A} \mathbf{h}(\mathbf{x}_i) \mathbf{h}(\mathbf{x}_i)'/q_i \right\}^{-1}.$$

Therefore, we obtain

$$\hat{\omega}_i = \pi_i^{-1} + \left(\sum_{i=1}^N \mathbf{h}(\mathbf{x}_i) - \sum_{i \in A} \pi_i^{-1} \mathbf{h}(\mathbf{x}_i) \right)' \left\{ \sum_{i \in A} \mathbf{h}(\mathbf{x}_i) \mathbf{h}(\mathbf{x}_i)'/q_i \right\}^{-1} \mathbf{h}(\mathbf{x}_i)/q_i.$$

2. Under the assumption of $\mathbf{h}_i'\mathbf{a} = q_i/\pi_i$ holds for some $\mathbf{a}$, show that

$$\sum_{i \in A} \frac{1}{\pi_i}\left(y_i - \mathbf{h}_i'\hat{\boldsymbol{\beta}}_h\right) = 0 \tag{2}$$

and

$$\sum_{i=1}^{N}\left(y_i - \mathbf{h}_i' B_h\right) = 0 \tag{3}$$

where $\mathbf{h}_i = \mathbf{h}\left(\mathbf{x}_i\right)$ and

$$\hat{\boldsymbol{\beta}}_h = \left\{\sum_{i \in A} \mathbf{h}(\mathbf{x}_i)\mathbf{h}(\mathbf{x}_i)'/q_i\right\}^{-1} \sum_{i \in A} \mathbf{h}(\mathbf{x}_i) y_i/q_i$$

and

$$B_h = \left\{\sum_{i=1}^{N} \pi_i \mathbf{h}(\mathbf{x}_i)\mathbf{h}(\mathbf{x}_i)'/q_i\right\}^{-1} \sum_{i=1}^{N} \pi_i \mathbf{h}(\mathbf{x}_i) y_i/q_i.$$

Solution: By the definition of $\hat{\boldsymbol{\beta}}_h$, we have

$$\sum_{i \in A}\left(y_i - \mathbf{h}_i'\hat{\boldsymbol{\beta}}_h\right)\mathbf{h}_i/q_i = \mathbf{0}.$$

Thus, we get

$$\sum_{i \in A}\left(y_i - \mathbf{h}_i'\hat{\boldsymbol{\beta}}_h\right)\underbrace{\mathbf{h}_i'\mathbf{a}}_{=q_i/\pi_i}\frac{1}{q_i} = \mathbf{0}'\mathbf{a},$$

which proves (2). Similarly, by the definition of B_h, we have

$$\sum_{i=1}^{N} \pi_i\left(y_i - \mathbf{h}_i' B_h\right)\mathbf{h}_i \frac{1}{q_i} = \mathbf{0}.$$

Thus, we get

$$\sum_{i=1}^{N} \pi_i\left(y_i - \mathbf{h}_i' B_h\right)\underbrace{\mathbf{h}_i'\mathbf{a}}_{=q_i/\pi_i}\frac{1}{q_i} = \mathbf{0}'\mathbf{a},$$

which proves (3).

3. Show this: If $\mathbf{h}_i'\mathbf{a} = q_i/\pi_i$ holds for some $\mathbf{a}$, then $\hat{Y}_{\text{cal}} = \sum_{i \in A}\hat{\omega}_i y_i$ is equivalent to the projection estimator of the form

$$\hat{Y}_{\text{proj}} = \sum_{i=1}^{N} \mathbf{h}_i'\hat{\boldsymbol{\beta}}_h.$$

Solution: By the solution to Q1, we obtain

$$\hat{Y}_{\text{cal}} = \sum_{i \in A} \hat{\omega}_i y_i = \sum_{i \in A} \pi_i^{-1} y_i + \left(\sum_{i=1}^{N} \mathbf{h}_i - \sum_{i \in A} \pi_i^{-1} \mathbf{h}(\mathbf{x}_i) \right)' \hat{\boldsymbol{\beta}}_h$$

By (2), we have

$$\sum_{i \in A} \pi_i^{-1} y_i - \sum_{i \in A} \pi_i^{-1} \mathbf{h}_i' \hat{\boldsymbol{\beta}}_h = 0$$

and so

$$\hat{Y}_{\text{cal}} = \sum_{i=1}^{N} \mathbf{h}_i' \hat{\boldsymbol{\beta}}_h.$$

4. Show this: If (2) holds, then we have

$$\hat{Y}_{\text{cal}} = Y + \sum_{i \in A} \frac{1}{\pi_i} \eta_i + \left(\sum_{i=1}^{N} \mathbf{h}_i - \sum_{i \in A} \frac{1}{\pi_i} \mathbf{h}_i \right)' \left(\hat{\boldsymbol{\beta}} - B_h \right)$$

where

$$\eta_i = y_i - \mathbf{h}_i' B_h \tag{4}$$

and

$$B_h = \left\{ \sum_{i=1}^{N} \pi_i \mathbf{h}(\mathbf{x}_i) \mathbf{h}(\mathbf{x}_i)' / q_i \right\}^{-1} \sum_{i=1}^{N} \pi_i \mathbf{h}(\mathbf{x}_i) y_i / q_i.$$

Solution: Note that

$$\begin{aligned}
\hat{Y}_{\text{cal}} &= \sum_{i \in A} \pi_i^{-1} y_i + \left(\sum_{i=1}^{N} \mathbf{h}_i - \sum_{i \in A} \pi_i^{-1} \mathbf{h}(\mathbf{x}_i) \right)' \hat{\boldsymbol{\beta}}_h \\
&= \sum_{i \in A} \pi_i^{-1} y_i + \left(\sum_{i=1}^{N} \mathbf{h}_i - \sum_{i \in A} \pi_i^{-1} \mathbf{h}(\mathbf{x}_i) \right)' B_h \\
&+ \left(\sum_{i=1}^{N} \mathbf{h}_i - \sum_{i \in A} \pi_i^{-1} \mathbf{h}(\mathbf{x}_i) \right)' \left(\hat{\boldsymbol{\beta}}_h - B_h \right) \\
&= \sum_{i=1}^{N} \mathbf{h}_i' B_h + \sum_{i \in A} \pi_i^{-1} \eta_i + \left(\sum_{i=1}^{N} \mathbf{h}_i - \sum_{i \in A} \pi_i^{-1} \mathbf{h}(\mathbf{x}_i) \right)' \left(\hat{\boldsymbol{\beta}}_h - B_h \right).
\end{aligned}$$

By (3), we obtain

$$\sum_{i=1}^{N} \mathbf{h}_i' B_h = \sum_{i=1}^{N} y_i$$

and the result follows.

5. Now, suppose that we have a superpopulation model with $Y_i \mid \mathbf{x}_i \sim (m(\mathbf{x}_i), q(\mathbf{x}_i)\sigma^2)$. Show that $E(\eta_i \mid \mathbf{X}) = 0$ if $\mathbf{h}_i$ includes $m(\mathbf{x}_i)$ in the sense that $\mathbf{h}_i'\boldsymbol{\alpha} = m(\mathbf{x}_i)$ for some $\boldsymbol{\alpha}$. [Hint: Show that $E_\zeta(B_h) = \boldsymbol{\alpha}$.]

Solution: Note that

$$
\begin{aligned}
E\left(B_h \mid \mathbf{X}\right) &= \left\{\sum_{i=1}^{N} \pi_i \mathbf{h}(\mathbf{x}_i)\mathbf{h}(\mathbf{x}_i)'/q_i\right\}^{-1} \sum_{i=1}^{N} \pi_i \mathbf{h}(\mathbf{x}_i) E(Y_i \mid \mathbf{X})/q_i \\
&= \left\{\sum_{i=1}^{N} \pi_i \mathbf{h}(\mathbf{x}_i)\mathbf{h}(\mathbf{x}_i)'/q_i\right\}^{-1} \sum_{i=1}^{N} \pi_i \mathbf{h}(\mathbf{x}_i)\mathbf{h}(\mathbf{x}_i)'\boldsymbol{\alpha}/q_i \\
&= \boldsymbol{\alpha}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
E\left(\eta_i \mid \mathbf{X}\right) &= E\left(Y_i \mid \mathbf{X}\right) - \mathbf{h}_i' E\left(B_h \mid \mathbf{X}\right) \\
&= m(\mathbf{x}_i) - \mathbf{h}_i'\boldsymbol{\alpha} \\
&= 0.
\end{aligned}
$$

6. If the model is

$$y_i = m(x_i) + e_i$$

with $e_i \sim (0, q(\mathbf{x}_i)\sigma^2)$ and $\mathbf{h}_i'\boldsymbol{\alpha} = m(\mathbf{x}_i)$ for some $\boldsymbol{\alpha}$, then

$$AV\left(\sum_{i\in A} \frac{1}{\pi_i}\eta_i\right) \cong \sum_{i=1}^{N} \left(\frac{1}{\pi_i} - 1\right) q(\mathbf{x}_i)\sigma^2$$

where $AV\left(\sum_{i\in A} \pi_i^{-1}\eta_i\right)$ is the anticipated variance (=model expectation of the design variance) of $\sum_{i\in A} \pi_i^{-1}\eta_i$ where η_i is defined in (4). [Hint: Use

$$
\begin{aligned}
AV\left(\sum_{i\in A} \frac{1}{\pi_i}\eta_i\right) &= E_\zeta\left(\sum_{i=1}^{N}\sum_{j=1}^{N} \left(\pi_{ij} - \pi_i\pi_j\right) \frac{\eta_i}{\pi_i}\frac{\eta_j}{\pi_j}\right) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} \left(\pi_{ij} - \pi_i\pi_j\right) \frac{E_\zeta(\eta_i)}{\pi_i}\frac{E_\zeta(\eta_j)}{\pi_j} \\
&+ \sum_{i=1}^{N}\sum_{j=1}^{N} \left(\pi_{ij} - \pi_i\pi_j\right) \frac{1}{\pi_i}\frac{1}{\pi_j} Cov_\zeta\left(\eta_i, \eta_j\right)
\end{aligned}
$$

and check that the first term is equal to zero.]

Solution: Since

$$V\left(\sum_{i\in A}\frac{1}{\pi_i}\eta_i \mid \mathcal{F}\right)=\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}-\pi_i\pi_j\right)\frac{\eta_i}{\pi_i}\frac{\eta_j}{\pi_j},$$

the anticipated variance is

$$\begin{aligned}
AV\left(\sum_{i\in A}\frac{1}{\pi_i}\eta_i\right) &= E_\zeta\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}-\pi_i\pi_j\right)\frac{\eta_i}{\pi_i}\frac{\eta_j}{\pi_j}\right)\\
&= \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}-\pi_i\pi_j\right)\frac{E_\zeta(\eta_i)}{\pi_i}\frac{E_\zeta(\eta_j)}{\pi_j}\\
&+ \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}-\pi_i\pi_j\right)\frac{1}{\pi_i}\frac{1}{\pi_j}Cov_\zeta\left(\eta_i,\eta_j\right).
\end{aligned}$$

Since $E_\zeta(\eta_i)=0$, the first term is zero. Also,

$$\eta_i = y_i-\mathbf{h}_i'\boldsymbol{\alpha}+\mathbf{h}_i'\left(B_h-\boldsymbol{\alpha}\right)=y_i-\mathbf{h}_i'\boldsymbol{\alpha}+O_p(n^{-1/2})$$

as $\boldsymbol{\alpha}=E_\zeta(B_h)$ is the probability limit of B_h. Thus,

$$\begin{aligned}
AV\left(\sum_{i\in A}\frac{1}{\pi_i}\eta_i\right) &= \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}-\pi_i\pi_j\right)\frac{1}{\pi_i}\frac{1}{\pi_j}Cov_\zeta\left(\eta_i,\eta_j\right)\\
&\cong \sum_{i=1}^{N}\left(\frac{1}{\pi_i}-1\right)V\left(y_i-\mathbf{h}_i'\boldsymbol{\alpha}\mid \mathbf{x}_i\right)
\end{aligned}$$

If the model is

$$y_i=m(x_i)+e_i$$

with $e_i\sim(0,q(\mathbf{x}_i)\sigma^2)$, then we have $m(\mathbf{x}_i)=\mathbf{h}_i\boldsymbol{\alpha}$ and

$$AV\left(\sum_{i\in A}\frac{1}{\pi_i}\eta_i\right)\cong\sum_{i=1}^{N}\left(\frac{1}{\pi_i}-1\right)V\left(e_i\mid \mathbf{x}_i\right)$$

which is equal to the GJLB.

Problem 2:

Consider a finite population of size N with measurement (x, y), where $x > 0$. We consider the following two-phase sampling.

1. Phase 1: select a sample A_1 of size n_1 by SRS and observe x_i for $i \in A_1$.
2. Phase 2: From A_1, select a Poisson sample A_2 of expected sample size $n_2 (< n_1)$ with the first-order inclusion probability $\pi_i = \pi(x_i) \in (0, 1)$.

We use the following two-phase ratio estimator of $\theta = N^{-1} \sum_{i=1}^{N} y_i$:

$$\hat{\theta}_{\text{tpr}} = \frac{1}{n_1} \sum_{i \in A_1} x_i \hat{\gamma}_2$$

where

$$\hat{\gamma}_2 = \frac{\sum_{i \in A_2} \pi_i^{-1} y_i}{\sum_{i \in A_2} \pi_i^{-1} x_i}.$$

1. Show that $\hat{\theta}_{\text{tpr}}$ is asymptotically design unbiased.

Solution: Write

$$\hat{\theta}_{\text{tpr}} = \frac{1}{N} \hat{X}_1 \frac{\hat{Y}_2}{\hat{X}_2}$$

where $(\hat{X}_2, \hat{Y}_2) = (N/n_1) \sum_{i \in A_2} \pi_i^{-1} (x_i, y_i)$ and $\hat{X}_1 = (N/n_1) \sum_{i \in A_1} x_i$. Note that $\hat{X}_1$, $\hat{X}_2$, $\hat{Y}_2$ converge to X, X, and Y, respectively. Using Taylor expansion, we obtain

$$\begin{aligned}
\hat{X}_1 \frac{\hat{Y}_2}{\hat{X}_2} &\cong \frac{1}{N} X \frac{Y}{X} + \frac{1}{N} \frac{Y}{X} \left(\hat{X}_1 - X \right) + \frac{1}{N} \left(\hat{Y}_2 - Y \right) - \frac{1}{N} \frac{Y}{X} \left(\hat{X}_2 - X \right) \\
&= \frac{1}{N} \hat{Y}_2 + \frac{1}{N} \left(\hat{X}_1 - \hat{X}_2 \right) \gamma := \hat{\theta}_\ell
\end{aligned}$$

where $\gamma = Y/X$. Now, we can show that

$$E(\hat{\theta}_\ell) = Y/N.$$

2. Derive the formula for linearization variance estimator of $\hat{\theta}_{\text{tpr}}$.

Solution: From the linearization formula, we can express

$$\hat{\theta}_\ell = \frac{1}{n_1} \sum_{i \in A_1} \left\{ x_i \gamma + \frac{\delta_i}{\pi_i} (y_i - x_i \gamma) \right\}$$

where $\delta_i = 1$ if $i \in A_2$ and $\delta_i = 0$ otherwise. Thus, the linearization variance estimator is

$$\hat{V} = \frac{1}{n_1} \frac{1}{n_1 - 1} \sum_{i \in A_1} \{\hat{\eta}_i - \bar{\eta}_1\}^2$$

where $\hat{\eta}_i = x_i \hat{\gamma}_2 + (\delta_i / \pi_i)(y_i - x_i \hat{\gamma}_2)$ and $\bar{\eta}_1 = n_1^{-1} \sum_{i \in A_1} \hat{\eta}_i$.

3. Find the formula for optimal inclusion probability $\pi(x)$ for the second-phase sampling (in the sense that it minimizes the asymptotic variance).

Solution: The asymptotic variance of $\hat{\theta}_{\text{tpr}}$ is

$$V\left(\hat{\theta}_\ell\right) = V\left(n_1^{-1} \sum_{i \in A_1} y_i\right) + E\left\{n_1^{-2} \sum_{i \in A_1} \left(\frac{1}{\pi_i} - 1\right) (y_i - x_i \gamma)^2\right\}$$

where $\gamma = Y/X$. Therefore, writing

$$E\{(y_i - x_i \gamma)^2 \mid x_i\} = v(x_i)$$

the optimal $\pi^*(x_i)$ is given by

$$\pi^*(x_i) \propto \{v(x_i)\}^{1/2}.$$