STAT 521: Take-Home Final Exam   Name:

Problem 1: (30 pts)

Suppose that Y is a binary random variable (taking either 1 or 0) and we are interested in estimating $\theta = P(Y = 1)$, the population proportion of $Y = 1$. We assume that x_i are available throughout the finite population but y_i are observed only in the sample.

To incorporate the auxiliary information, we consider the following logistic regression model

$$P(Y = 1 \mid x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} := p(x; \beta_0, \beta_1)$$

and estimate (β_0, β_1) by solving the following weighted score equation:

$$\sum_{i \in A} \frac{1}{\pi_i} \{y_i - p(x_i; \beta_0, \beta_1)\} (1, x_i) = (0, 0).$$

Once $(\hat{\beta}_0, \hat{\beta}_1)$ is computed from the above formula, we use the following projection estimator.

$$\hat{\theta}_P = \frac{1}{N} \sum_{i=1}^{N} \hat{p}_i,$$

where

$$\hat{p}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}$$

1. Let (β_0^*, β_1^*) be the finite-population quantity that satisfies

$$\sum_{i=1}^{N} \{y_i - p(x_i; \beta_0^*, \beta_1^*)\} (1, x_i) = (0, 0)$$

Show that, by Taylor linearization, $\hat{\theta}_p$ is asymptotically equivalent to

$$\hat{\theta}_\ell = \frac{1}{N} \sum_{i=1}^{N} p_i^* + \frac{1}{N} \sum_{i \in A} \frac{1}{\pi_i} (y_i - p_i^*) \tag{1}$$

for some γ_0^* and γ_1^*, where $p_i^* = p(\mathbf{x}_i; \beta_0^*, \beta_1^*)$. Find the expression for γ_0^* and γ_1^*.

Solution: Define

$$\hat{\theta}_\ell(\beta_0, \beta_1) \quad = \quad \frac{1}{N}\sum_{i=1}^{N} p(x_i; \beta_0, \beta_1) + \frac{1}{N}\sum_{i \in A} \frac{1}{\pi_i}\left\{y_i - p_i(x_i; \beta_0, \beta_1)\right\}.$$

Note that

$$\hat{\theta}_P = \hat{\theta}_\ell(\hat{\beta}_0, \hat{\beta}_1)$$

by the construction of $(\hat{\beta}_0, \hat{\beta}_1)$. Also, $\hat{\theta}_P$ is asymptotically equivalent to $\hat{\theta}_\ell(\beta_0^*, \beta_1^*)$ if it satisfies

$$E\left\{\frac{\partial}{\partial \beta_0}\hat{\theta}_\ell(\beta_0^*, \beta_1^*)\right\} = 0 \tag{A.1}$$

and

$$E\left\{\frac{\partial}{\partial \beta_1}\hat{\theta}_\ell(\beta_0^*, \beta_1^*)\right\} = 0. \tag{A.2}$$

Since

$$\frac{\partial}{\partial \beta_0}\hat{\theta}_\ell(\beta_0^*, \beta_1^*) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial \beta_0}p_i(x_i; \beta_0^*, \beta_1^*) - \frac{1}{N}\sum_{i \in A}\frac{1}{\pi_i}\frac{\partial}{\partial \beta_0}p_i(x_i; \beta_0^*, \beta_1^*)$$

we can show (A.1), where the expectation is wrt the sampling mechanism for selecting A. Similarly, (A.2) can be proved.

2. Show that $\hat{\theta}_\ell$ in (1) is design unbiased for $\theta_N = N^{-1}\sum_{i=1}^{N} y_i$. How to estimate the variance of $\hat{\theta}_\ell$ from the observations in the sample?

Solution: By the sampling mechanism,

$$\begin{aligned} E(\hat{\theta}_\ell) &= \frac{1}{N}\sum_{i=1}^{N} p_i^* + E\left\{\frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i}(y_i - p_i^*)\right\} \\ &= \frac{1}{N}\sum_{i=1}^{N} p_i^* + \frac{1}{N}\sum_{i=1}^{N}(y_i - p_i^*) = N^{-1}\sum_{i=1}^{N} y_i. \end{aligned}$$

Thus, $\hat{\theta}_\ell$ is unbiased for θ_N. To estimate the variance, we can use

$$\hat{V} = \frac{1}{N^2}\sum_{i\in A}\sum_{j\in A}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{\hat{e}_i}{\pi_i}\frac{\hat{e}_j}{\pi_j}.$$

3. Compute the approximate anticipated variance of $\hat{\theta}_p$ and derive the optimal π_i (in terms of x and β) that minimizes the anticipated variance.

Solution: We have only to compute the anticipate variance of the difference estimator in (**??**). Now,

$$\begin{aligned} AV(\hat{\theta}_{\mathrm{P}}) &\cong E_\zeta V_p\left(\hat{\theta}_\ell\right) \\ &= N^{-2}E_\zeta\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\frac{e_i}{\pi_i}\frac{e_j}{\pi_j}\right\} \\ &= N^{-2}\sum_{i=1}^{N}\left(\frac{1}{\pi_i} - 1\right)V_\zeta(e_i) \\ &= N^{-2}\sum_{i=1}^{N}\left(\frac{1}{\pi_i} - 1\right)p_i(1-p_i), \end{aligned}$$

where $e_i = y_i - p_i$ and $p_i = p(x_i;\beta_0,\beta_1)$. Thus, minimizing $\sum_{i=1}^{N}\pi_i^{-1}p_i(1-p_i)$ subject to $\sum_{i=1}^{N}\pi_i = n$ leads to

$$\pi_i^* \propto \sqrt{p_i(1-p_i)}.$$

Problem 2: (20 pts)

Consider a finite population with bivariate measurement (X, Y), where both X and Y are categorical taking values in $\{0, 1\}$. From the finite population, we are interested in estimating $P = Pr(Y = 1)$. Let N_{ab} be the number of elements with $(X = a, Y = b)$ in the population, where $a = 0, 1; b = 0, 1$.

From the finite population, we select a SRS of size n and observe (x_i, y_i) in the sample. Let n_{ab} be the number of elements with $(x_i, y_i) = (a, b)$ in the sample. The HT estimator of P is $\hat{P}_{HT} = n_{+1}/n$, where $n_{+1} = n_{01} + n_{11}$.

Now, suppose that x_i are available throughout the finite population so that we know N_{1+} and N_{0+} outside the sample. To take advantage of this extra information, we consider the following estimator:

$$\hat{P}_r = \frac{1}{1 + \hat{\theta}_r}$$

where

$$\hat{\theta}_r = \frac{N_{0+}}{N_{1+}} \times \frac{n_{1+}}{n_{0+}} \times \frac{n_{+0}}{n_{+1}}.$$

Answer the following questions:

1. Show that $\hat{P}_r$ is asymptotically unbiased.

Solution: We can express

$$\hat{P}_r = f(\bar{x}, \bar{y}) = \left\{ 1 + \left(\frac{1 - \bar{X}}{\bar{X}} \right) \times \left(\frac{\bar{x}}{1 - \bar{x}} \right) \times \left(\frac{1 - \bar{y}}{\bar{y}} \right) \right\}^{-1}$$

where $(\bar{x}, \bar{y}) = n^{-1} \sum_{i \in A} (x_i, y_i)$ and $\left(\bar{X}, \bar{Y}\right) = N^{-1} \sum_{i=1}^{N} (x_i, y_i)$. Now, we can show

$$f(\bar{X}, \bar{Y}) = \left\{ 1 + \frac{1 - \bar{Y}}{\bar{Y}} \right\}^{-1} = \bar{Y} = P$$

which proves the asymptotic unbiasedness of $\hat{P}_r$.

2. Derive the asymptotic variance of $\hat{P}_r$.

Solution: Using Taylor expansion, we can obtain

$$\hat{P}_r \cong f(\bar{X}, \bar{Y}) + \frac{\partial}{\partial \bar{X}} f(\bar{X}, \bar{Y})(\bar{x} - \bar{X}) + \frac{\partial}{\partial \bar{Y}} f(\bar{X}, \bar{Y})(\bar{y} - \bar{Y}) := \hat{P}_\ell.$$

Now, since

$$f(\bar{X}, \bar{Y}) = \bar{Y}$$

$$\begin{aligned}
\frac{\partial}{\partial \bar{X}} f(\bar{X}, \bar{Y}) &= -\{f(\bar{X}, \bar{Y})\}^2 \times \frac{1-\bar{X}}{\bar{X}} \times \left(\frac{1-\bar{Y}}{\bar{Y}}\right) \frac{\partial}{\partial \bar{X}} \left(\frac{\bar{X}}{1-\bar{X}}\right) \\
&= -\bar{Y}^2 \times \frac{1-\bar{X}}{\bar{X}} \times \left(\frac{1-\bar{Y}}{\bar{Y}}\right) \times \frac{1}{(1-\bar{X})^2} \\
&= -\frac{\bar{Y}(1-\bar{Y})}{\bar{X}(1-\bar{X})}
\end{aligned}$$

and

$$\frac{\partial}{\partial \bar{Y}} f(\bar{X}, \bar{Y}) = -\{f(\bar{X}, \bar{Y})\}^2 \times \frac{\partial}{\partial \bar{Y}} \left(\frac{1-\bar{Y}}{\bar{Y}}\right) = 1.$$

Thus,

$$\hat{P}_\ell = \bar{y} - \frac{\bar{Y}(1-\bar{Y})}{\bar{X}(1-\bar{X})} (\bar{x} - \bar{X})$$

and

$$V\left(\hat{P}_\ell\right) = V(\bar{y}) + \left\{\frac{\bar{Y}(1-\bar{Y})}{\bar{X}(1-\bar{X})}\right\}^2 V(\bar{x}) - 2\frac{\bar{Y}(1-\bar{Y})}{\bar{X}(1-\bar{X})} Cov(\bar{x}, \bar{y})$$

3. Under what conditions, $\hat{P}_r$ is more efficient than the HT estimator?

Solution: Therefore,

$$
\begin{aligned}
V(\hat{P}_\ell) < V(\bar{y}) \quad &\Longleftrightarrow \quad \left\{ \frac{\bar{Y}(1-\bar{Y})}{\bar{X}(1-\bar{X})} \right\}^2 V(\bar{x}) - 2\frac{\bar{Y}(1-\bar{Y})}{\bar{X}(1-\bar{X})} Cov(\bar{x}, \bar{y}) < 0 \\
&\Longleftrightarrow \quad \frac{\bar{Y}(1-\bar{Y})}{\bar{X}(1-\bar{X})} < 2\frac{Cov(\bar{x}, \bar{y})}{Var(\bar{x})} \\
&\Longleftrightarrow \quad Var(\bar{y}) < 2Cov(\bar{x}, \bar{y})
\end{aligned}
$$

Problem 3: (40 pts)

Assume that two independent samples are drawn from the same population. Let A_1 and A_2 be the set of the sample indices for the two SRS samples with the size n_1 and n_2, respectively. Assume that only x_i is observed in sample A_1 and x_i and y_i are observed in sample A_2. Let $\bar{x}_1 = n_1^{-1} \sum_{i \in A_1} x_i$ and $\bar{x}_2 = n_2^{-1} \sum_{i \in A_2} x_i$ be the unbiased estimators of $\bar{x}_N = N^{-1} \sum_{i=1}^N x_i$ from sample A_1 and from sample A_2, respectively. Also, $\bar{y}_2 = n_2^{-1} \sum_{i \in A_2} y_i$ is an unbiased estimator of $\bar{y}_N = N^{-1} \sum_{i=1}^N y_i$. Consider the following regression estimator

$$\bar{y}_{reg} = \bar{y}_2 + (\bar{x}_1 - \bar{x}_2) \hat{\beta}_2$$

where $\hat{\beta}_2$ is the slope β for the regression of y on x, obtained from the sample A_2.

1. Show that $\bar{y}_{reg}$ is approximately design unbiased. Compute the asymptotic variance of $\bar{y}_{reg}$.

Solution: Let

$$\begin{pmatrix} B_0 \\ B_1 \end{pmatrix} = \begin{pmatrix} N & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_i y_i \end{pmatrix}.$$

We can obtain

$$\bar{y}_{\text{reg}} = \bar{y}_2 + (\bar{x}_1 - \bar{x}_2) B_1 + (\bar{x}_1 - \bar{x}_2)\left(\hat{\beta}_2 - B_1\right)$$

Since $\bar{x}_1 - \bar{x}_2$ converges in probability to zero and $\hat{\beta}_2 - B_1$ converges in probability to zero, we can express

$$\begin{aligned} \bar{y}_{\text{reg}} &\cong \bar{y}_2 + (\bar{x}_1 - \bar{x}_2) B_1 \\ &= \frac{1}{n_2}\sum_{i \in A_2}(y_i - B_0 - B_1 x_i) + \frac{1}{n_1}\sum_{i \in A_1}(B_0 + B_1 x_i) \\ &:= \bar{y}_\ell \end{aligned}$$

Now,

$$\begin{aligned} E(\bar{y}_\ell) &= \frac{1}{N}\sum_{i=1}^N (y_i - B_0 - B_1 x_i) + \frac{1}{N}\sum_{i=1}^N (B_0 + B_1 x_i) \\ &= \frac{1}{N}\sum_{i=1}^N y_i \end{aligned}$$

and

$$\begin{aligned} V(\bar{y}_\ell) &= V\left\{\frac{1}{n_2}\sum_{i \in A_2}(y_i - B_0 - B_1 x_i)\right\} + V\left\{\frac{1}{n_1}\sum_{i \in A_1}(B_0 + B_1 x_i)\right\} \\ &= \frac{1}{n_2}\left(1 - \frac{n_2}{N}\right)\frac{1}{N-1}\sum_{i=1}^N (y_i - B_0 - B_1 x_i)^2 + \frac{1}{n_1}\left(1 - \frac{n_1}{N}\right)\frac{1}{N-1}\sum_{i=1}^N (x_i - \bar{x}_N)^2 B_1^2 \\ &:= V_2 + V_1 \end{aligned}$$

2. Under what conditions, we have $V(\bar{y}_{reg}) < V(\bar{y}_2)$? Answer the question in terms of the sample sizes.

Solution: Write

$$\text{SST} = \sum_{i=1}^{N}(y_i - \bar{y}_N)^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 = \text{SSE} + \text{SSR}$$

where $\hat{y}_i = B_0 + B_1 x_i$. We can express

$$V(\bar{y}_\ell) = \frac{1}{n_2}\left(1 - \frac{n_2}{N}\right)\frac{1}{N-1}\text{SSE} + \frac{1}{n_1}\left(1 - \frac{n_1}{N}\right)\frac{1}{N-1}\text{SSR}$$

and

$$V(\bar{y}_2) = \frac{1}{n_2}\left(1 - \frac{n_2}{N}\right)\frac{1}{N-1}\text{SST}.$$

Solution: Thus,

$$\begin{aligned} V(\bar{y}_2) - V(\bar{y}_\ell) &= \frac{1}{n_2}\left(1 - \frac{n_2}{N}\right)\frac{1}{N-1}(\text{SST} - \text{SSE}) - \frac{1}{n_1}\left(1 - \frac{n_1}{N}\right)\frac{1}{N-1}SSR \\ &= \frac{1}{N-1}\left(\frac{1}{n_2} - \frac{1}{n_1}\right)\text{SSR}. \end{aligned}$$

Thus, $V(\bar{y}_\ell) < V(\bar{y}_2)$ if $n_1 > n_2$.

3. Discuss how you can obtain a consistent estimator for the variance of $\bar{y}_{reg}$ from the two samples.

Solution: Since $V(\bar{y}_{reg}) \cong V_1 + V_2$, we can estimate the two terms separately as follows.

$$\begin{aligned}\hat{V}_1 &= \frac{1}{n_1}\left(1-\frac{n_1}{N}\right)\frac{1}{n_1-1}\sum_{i\in A_1}(x_i-\bar{x}_1)^2\hat{B}_1\\ \hat{V}_2 &= \frac{1}{n_2}\left(1-\frac{n_2}{N}\right)\frac{1}{n_2-1}\sum_{i\in A_2}\left(y_i-\hat{B}_0-\hat{B}_1x_i\right)^2\end{aligned}$$

where

$$\begin{pmatrix}\hat{B}_0\\ \hat{B}_1\end{pmatrix} = \begin{pmatrix} n_2 & \sum_{i\in A_2} x_i \\ \sum_{i\in A_2} x_i & \sum_{i\in A_2} x_i^2\end{pmatrix}^{-1}\begin{pmatrix}\sum_{i\in A_2} y_i\\ \sum_{i\in A_2} x_iy_i\end{pmatrix}.$$

4. Express $\bar{y}_{reg}$ as a calibration estimator. That is, discuss how to express $\hat{\omega}_i$ for $\bar{y}_{reg} = \sum_{i\in A_2}\hat{\omega}_i y_i$ as the solution to the primal optimization problem of the weights.

Solution: The weight $\hat{\omega}_i$ for $\bar{y}_{reg} = \sum_{i\in A_2}\hat{\omega}_i y_i$ can be obtained by the minimizer of

$$Q(\omega) = \sum_{i\in A_2}\omega_i^2$$

subject to

$$\sum_{i\in A_2}\omega_i(1,x_i) = (1,\bar{x}_1).$$