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:

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

where π_i is the first-order inclusion probability of unit i.

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

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

Show that $\hat{\theta}_P$ is asymptotically equivalent to

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

where $p_i^* = p(x_i; \beta_0^*, \beta_1^*)$.

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?

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 (given a fixed value of expected sample size). You may assume Poisson sampling.

Problem 2: (30 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.

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

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

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 $\bar{y}_{reg}$.
2. Under what conditions, we have $V\left(\bar{y}_{reg}\right) < V\left(\bar{y}_2\right)$? Answer the question in terms of the sample sizes.
3. Discuss how you can obtain a consistent estimator for the variance of $\bar{y}_{reg}$ from the two samples.
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.