STAT 341: Final Take Home Assessment

Sketch Solutions and Marking Rubric

MARKING

Questions are worth 1 or 5 points.

- Questions worth 1 point are trivial; 1 point is awarded if the answer is correct and 0 points otherwise.
- Questions worth 5 points (that are attempted) are to be marked as follows:
  - 5: answer is entirely correct
  - 4: answer is mostly correct
  - 3: answer is partially correct
  - 2: answer is mostly incorrect
  - 1: answer is entirely incorrect, but an attempt was made
- A question that is not attempted it worth 0 points.

QUESTION 1 [5 POINTS]

By searching the web, find a public dataset that constitutes a population. For this data, provide the following:

- A description of the data (define what is a unit and two variate(s) that have been recorded)
- A justification for why the dataset is indeed a population (as opposed to a sample)
- A URL to access the data

COMMENTS & SOLUTIONS

For full points the student must have sufficiently justified *why* the described dataset is a population. Common point deductions are as follows:

- URL missing: -1
- Data description missing: -1
- Population justification missing: -2
- Population justifcation inadequate: -1

QUESTION 2 [15 POINTS]

The Horvitz-Thompson estimate of a population total, calculated over a sample $\mathcal{S} = \{y_1, y_2, \ldots, y_n\}$, is given by

$$a_{HT}(\mathcal{S}) = \sum_{u \in \mathcal{S}} \frac{y_u}{\pi_u}.$$

(a) [5] Determine whether the Horvitz-Thompson estimate $a_{HT}(\mathcal{S})$ is location invariant, location equivariant, or neither.

(b) [5] Determine whether the Horvitz-Thompson estimate $a_{HT}(\mathcal{S})$ is scale invariant, scale equivariant, or neither.

(c) [5] Derive and sketch the sensitivity curve $SC\left(y; a_{HT}(\mathcal{S})\right)$ for the Horvitz-Thompson estimate, given a sample $\mathcal{S} = \{y_1, y_2, \ldots, y_{n-1}\}$.

COMMENTS & SOLUTIONS

(a) Let $S^* = \{x_1, x_2, \ldots, x_n\} \equiv \{y_1 + b, y_2 + b, \ldots, y_n + b\}$ for some $b \in \mathbb{R}$. Then

$$\begin{aligned} a_{HT}(\mathcal{S}^*) &= \sum_{u \in \mathcal{S}^*} \frac{x_u}{\pi_u} \\ &= \sum_{u \in \mathcal{S}} \frac{y_u + b}{\pi_u} \\ &= \left(\sum_{u \in \mathcal{S}} \frac{y_u}{\pi_u}\right) + \left(\sum_{u \in \mathcal{S}} \frac{b}{\pi_u}\right) \\ &= a_{HT}(\mathcal{S}) + b \times \sum_{u \in \mathcal{S}} \pi_u^{-1} \end{aligned}$$

Thus $a_{HT}(\mathcal{S}^*) \neq a_{HT}(\mathcal{S}) + b$ because $\sum_{u \in \mathcal{S}} \pi_u^{-1} \neq 1$ and so the Horvitz-Thompson estimate is neither location invariant nor location equivariant.

(b) Let $S^* = \{x_1, x_2, \ldots, x_n\} \equiv \{m \times y_1, m \times y_2, \ldots, m \times y_n\}$ for some $m \in \mathbb{R}$. Then

$$\begin{aligned} a_{HT}(\mathcal{S}^*) &= \sum_{u \in \mathcal{S}^*} \frac{x_u}{\pi_u} \\ &= \sum_{u \in \mathcal{S}} \frac{m \times y_u}{\pi_u} \\ &= m \times \sum_{u \in \mathcal{S}} \frac{y_u}{\pi_u} \\ &= m \times a_{HT}(\mathcal{S}) \end{aligned}$$

Thus the Horvitz-Thompson estimate is scale equivariant.

(c) Let $\pi_1^*, \pi_2^*, \ldots, \pi_{n-1}^*, \pi_n^*$ be the inclusion probabilities for $y_1, y_2, \ldots, y_{n-1}, y$. Note that $\pi_1^*, \pi_2^*, \ldots, \pi_{n-1}^*$ may be different from $\pi_1, \pi_2, \ldots, \pi_{n-1}$ (the inclusion probabilities of $y_1, y_2, \ldots, y_{n-1}$ when the variate y is not included in the sample). Thus

$$a_{HT}(\mathcal{S} \cup \{y\}) = \sum_{u \in \mathcal{S} \cup \{y\}} \frac{y_u}{\pi_u^*} = \sum_{u \in \mathcal{S}} \frac{y_u}{\pi_u^*} + \frac{y}{\pi_n^*}$$

and so

$$\begin{aligned} SC(y) &= n\left[a_{HT}(\mathcal{S} \cup \{y\}) - a_{HT}(\mathcal{S})\right] \\ &= n\left[\sum_{u \in \mathcal{S}} \frac{y_u}{\pi_u^*} + \frac{y}{\pi_n^*} - \sum_{u \in \mathcal{S}} \frac{y_u}{\pi_u}\right] \\ &= n \sum_{u \in \mathcal{S}} \left(\frac{1}{\pi_u^*} - \frac{1}{\pi_u}\right) y_u + \left(\frac{n}{\pi_n^*}\right) y \end{aligned}$$

which is a linear function of y with positive slope and non-zero intercept. The sketch of $SC(y)$ versus y should reflect this.

QUESTION 3 [11 POINTS]

In class we talked about *robust regression* as an outlier-resistant means to estimate $\boldsymbol{\theta} = (\alpha, \beta)^T$ in the context of the following simple linear regression model

$$y_u = \alpha + \beta x_u + r_u.$$

We did so using the *Huber objective function* which behaved like the least squares objective function for small (in magnitude) values of r_u but which was less sensitive than the least squares objective function for large (in magnitude) values of r_u.

Another objective function that similarly facilitates robust regression is the **Tukey Biweight objective function**:

$$\rho(\boldsymbol{\theta}; \mathcal{P}) = \sum_{u \in \mathcal{P}} \rho_k(r_u)$$

where $\boldsymbol{\theta} = (\alpha, \beta)^T$, $r_u = y_u - \alpha - \beta x_u$ and

$$\rho_k(r) = \begin{cases} \frac{r^2}{2} - \frac{r^4}{2k^2} + \frac{r^6}{6k^4} & \text{for } |r| \leq k \\ \frac{k^2}{6} & \text{for } |r| > k \end{cases}$$

(a) [5] Determine the vector $\boldsymbol{\psi}(\boldsymbol{\theta}; \mathcal{P})$ and matrix $\boldsymbol{\psi}'(\boldsymbol{\theta}; \mathcal{P})$. Show your work.

(b) [1] In terms of $\boldsymbol{\theta} = (\alpha, \beta)^T$ and the data, write the equation that the Newton-Raphson method is designed to solve.

(c) [5] In point form, describe the Newton-Raphson algorithm (in terms of $\boldsymbol{\theta} = (\alpha, \beta)^T$ and the data). Define any notation that you introduce.

COMMENTS & SOLUTIONS

(a)

$$\begin{aligned}
\boldsymbol{\psi}(\boldsymbol{\theta}; \mathcal{P}) &= \begin{bmatrix} \frac{\partial \rho(\boldsymbol{\theta}; \mathcal{P})}{\partial \alpha} \\ \\ \frac{\partial \rho(\boldsymbol{\theta}; \mathcal{P})}{\partial \beta} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{u \in \mathcal{P}} \frac{\partial \rho_k(r_u)}{\partial \alpha} \\ \\ \sum_{u \in \mathcal{P}} \frac{\partial \rho_k(r_u)}{\partial \beta} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{u \in \mathcal{P}} \frac{\partial \rho_k(r_u)}{\partial r_u} \times \frac{\partial r_u}{\partial \alpha} \\ \\ \sum_{u \in \mathcal{P}} \frac{\partial \rho_k(r_u)}{\partial r_u} \times \frac{\partial r_u}{\partial \beta} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{u \in \mathcal{P}} \frac{\partial \rho_k(r_u)}{\partial r_u} \times (-1) \\ \\ \sum_{u \in \mathcal{P}} \frac{\partial \rho_k(r_u)}{\partial r_u} \times (-x_u) \end{bmatrix} \\
&= \begin{bmatrix} -\sum_{u \in \mathcal{P}} \frac{\partial \rho_k(r_u)}{\partial r_u} \\ \\ -\sum_{u \in \mathcal{P}} \frac{\partial \rho_k(r_u)}{\partial r_u} \times x_u \end{bmatrix} \equiv \begin{bmatrix} \psi_1 \\ \\ \psi_2 \end{bmatrix}
\end{aligned}$$

where

$$\frac{\partial \rho_k(r)}{\partial r} = \begin{cases} r - \frac{2r^3}{k^2} + \frac{r^5}{k^4} & \text{if } |r| \leq k \\ 0 & \text{if } |r| > k \end{cases}$$

And

$$\boldsymbol{\psi}'(\boldsymbol{\theta};\mathcal{P}) = \begin{bmatrix} \frac{\partial \psi_1}{\partial \alpha} & \frac{\partial \psi_1}{\partial \beta} \\ \frac{\partial \psi_2}{\partial \alpha} & \frac{\partial \psi_2}{\partial \beta} \end{bmatrix}$$

where

$$\frac{\partial \psi_1}{\partial \alpha} = -\sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial \alpha \partial r_u} = -\sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial r_u^2} \times \frac{\partial r_u}{\partial \alpha} = -\sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial r_u^2} \times (-1) = \sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial r_u^2}$$

and

$$\frac{\partial \psi_1}{\partial \beta} = -\sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial \beta \partial r_u} = -\sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial r_u^2} \times \frac{\partial r_u}{\partial \beta} = -\sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial r_u^2} \times (-x_u) = \sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial r_u^2} \times x_u$$

and

$$\frac{\partial \psi_2}{\partial \alpha} = -\sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial \alpha \partial r_u} \times x_u = -\sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial r_u^2} \times \frac{\partial r_u}{\partial \alpha} \times x_u = -\sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial r_u^2} \times (-1) \times x_u = \sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial r_u^2} \times x_u$$

and

$$\frac{\partial \psi_2}{\partial \beta} = -\sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial \beta \partial r_u} \times x_u = -\sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial r_u^2} \times \frac{\partial r_u}{\partial \beta} \times x_u = -\sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial r_u^2} \times (-x_u) \times x_u = \sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial r_u^2} \times x_u^2.$$

Therefore

$$\boldsymbol{\psi}'(\boldsymbol{\theta};\mathcal{P}) = \sum_{u\in\mathcal{P}} \frac{\partial^2 \rho_k(r_u)}{\partial r_u^2} \begin{bmatrix} 1 & x_u \\ x_u & x_u^2 \end{bmatrix}$$

where

$$\frac{\partial^2 \rho_k(r)}{\partial r^2} = \begin{cases} 1 - \frac{6r^2}{k^2} + \frac{5r^4}{k^4} & \text{if } |r| \le k \\ 0 & \text{if } |r| > k \end{cases}$$

(b) The equation to be solved is

$$\boldsymbol{\psi}'(\boldsymbol{\theta};\mathcal{P}) = \mathbf{0}$$

which in terms of the equations derived above is

$$\begin{bmatrix} -\sum_{u\in\mathcal{P}} \frac{\partial \rho_k(r_u)}{\partial r_u} \\ -\sum_{u\in\mathcal{P}} \frac{\partial \rho_k(r_u)}{\partial r_u} \times x_u \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

(c) Given a starting value $\widehat{\boldsymbol{\theta}}_0$

- Initialize $i = 0$
- LOOP over i
  - Update the iterate
$$\widehat{\boldsymbol{\theta}}_{i+1} = \widehat{\boldsymbol{\theta}}_i - \left[\boldsymbol{\psi}'(\widehat{\boldsymbol{\theta}}_i; \mathcal{P})\right]^{-1} \boldsymbol{\psi}(\widehat{\boldsymbol{\theta}}_i; \mathcal{P})$$
  - Check convergence
    - IF coverged RETURN
    - ELSE $i = i + 1$ and repeat LOOP
- RETURN $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_i$

QUESTION 4 [15 POINTS]

Cluster sampling is a probabilistic sampling mechanism that is applicable when a population $\mathcal{P}$ can be parititioned into H clusters (i.e., sub-populations) $\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_H\}$ such that

$$\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2 \cup \cdots \cup \mathcal{P}_H \qquad \text{and} \qquad N = N_1 + N_2 + \cdots + N_H$$

where N_h is the size of cluster $h = 1, 2, \ldots, H$. In *two-stage cluster sampling* the sample $\mathcal{S}$ is obtained in two stages:

1. Randomly select (without replacement) $n_1 \leq H$ clusters
2. From each of those n_1 clusters, randomly select (without replacement) n_2 units.

The size of the sample $\mathcal{S}$ is thus $n = n_1 \times n_2$.

(a) [5] Assuming $u \in \mathcal{P}_j$, calculate the (marginal) inclusion probability $\pi_u = P(u \in \mathcal{S})$.

(b) [5] Assuming $u \in \mathcal{P}_j$ and $v \in \mathcal{P}_j$ and $u \neq v$, calculate the joint inclusion probability $\pi_{uv} = P(u \in \mathcal{S}, v \in \mathcal{S})$.

(c) [5] Assuming $u \in \mathcal{P}_j$ and $v \in \mathcal{P}_k$ and $u \neq v$, calculate the joint inclusion probability $\pi_{uv} = P(u \in \mathcal{S}, v \in \mathcal{S})$.

COMMENTS & SOLUTIONS

Note that each stage in this two-stage sampling protocol corresponds to taking simple random samples without replacement; first, a SRSWOR of clusters and second, a SRSWOR of units from within each of the selected clusters. Moreover, selections at each stage are made independently. We can exploit these two facts in the derivations that follow. In particular, we will use, without proof, SRSWOR inclusion probabilities where appropriate (students may do the same). And due to the independence, each probabability may be written as a product of the Stage 1 probability and the Stage 2 probability, as will be seen below.

(a)

$$\begin{aligned}
\pi_u &= P(u \in \mathcal{S}) \\
&= P(\mathcal{P}_j \text{ is selected} \cap u \text{ is selected}) \\
&= P(\mathcal{P}_j \text{ is selected}) \times P(u \text{ is selected} \mid \mathcal{P}_j \text{ is selected}) \\
&= \frac{n_1}{H} \times \frac{n_2}{N_j}
\end{aligned}$$

(b)

$$\begin{aligned}
\pi_{uv} &= P(u \in \mathcal{S}, v \in \mathcal{S}) \\
&= P(\mathcal{P}_j \text{ is selected} \cap u \text{ is selected} \cap v \text{ is selected}) \\
&= P(\mathcal{P}_j \text{ is selected}) \times P(u \text{ is selected} \cap v \text{ is selected} \mid \mathcal{P}_j \text{ is selected}) \\
&= \frac{n_1}{H} \times \frac{n_2(n_2 - 1)}{N_j(N_j - 1)}
\end{aligned}$$

(c)

$$\begin{aligned}
\pi_{uv} &= P(u \in \mathcal{S}, v \in \mathcal{S}) \\
&= P(\mathcal{P}_j \text{ is selected} \cap u \text{ is selected} \cap \mathcal{P}_k \text{ is selected} \cap v \text{ is selected}) \\
&= P(\mathcal{P}_j \text{ is selected} \cap \mathcal{P}_k \text{ is selected}) \times P(u \text{ is selected} \mid \mathcal{P}_j \text{ is selected}) \times P(v \text{ is selected} \mid \mathcal{P}_k \text{ is selected}) \\
&= \frac{n_1(n_1-1)}{H(H-1)} \times \frac{n_2}{N_j} \times \frac{n_2}{N_k}
\end{aligned}$$

QUESTION 5 [11 POINTS]

Suppose that $\mathcal{S} = \{9, 21\}$ is a *simple random sample without replacement* from a population $\mathcal{P}$ of size $N = 3$.

(a) [5] Calculate the Horvitz-Thompson estimate of the population average.

(b) [5] Calculate the standard error associated with the estimate from part (a).

(c) [1] Calculate an approximate 95% confidence interval for the true population average.

COMMENTS & SOLUTIONS

Note that the scenario of this question is a SRSWOR of size $n = 2$ from a population of size $N = 3$. Therefore relevant inclusion probabilities below will be

$$\pi_u = \frac{n}{N} = \frac{2}{3} \qquad \text{and} \qquad \pi_{uv} = \frac{n(n-1)}{N(N-1)} = \frac{2 \times 1}{3 \times 2} = \frac{1}{3}$$

(a)

$$\begin{aligned} a_{HT}(\mathcal{S}) &= \sum_{u \in \mathcal{S}} \frac{y_u}{\pi_u} \\ &= \frac{9/3}{2/3} + \frac{21/3}{2/3} \\ &= 15 \end{aligned}$$

(b)

$$\begin{aligned} \widehat{Var}\left[\widetilde{a}_{HT}(\mathcal{S})\right] &= \sum_{u \in \mathcal{S}} \sum_{v \in \mathcal{S}} \left(\frac{\pi_{uv} - \pi_u \pi_v}{\pi_{uv}}\right) \left(\frac{y_u}{\pi_u}\right) \left(\frac{y_v}{\pi_v}\right) \\ &= \sum_{u \in \mathcal{S}} (1 - \pi_u) \left(\frac{y_u}{\pi_u}\right)^2 + \sum_{u \in \mathcal{S}} \sum_{v \neq u \in \mathcal{S}} \left(1 - \frac{\pi_u \pi_v}{\pi_{uv}}\right) \left(\frac{y_u}{\pi_u}\right) \left(\frac{y_v}{\pi_v}\right) \\ &= \left(1 - \frac{2}{3}\right) \frac{(9/3)^2}{(2/3)^2} + \left(1 - \frac{2}{3}\right) \frac{(21/3)^2}{(2/3)^2} + 2 \left(1 - \frac{(2/3)^2}{1/3}\right) \left(\frac{9/3}{2/3}\right) \left(\frac{21/3}{2/3}\right) \\ &= \frac{27}{4} + \frac{147}{4} - \frac{63}{2} \\ &= 12 \end{aligned}$$

And therefore $SE\left[\widetilde{a}_{HT}(\mathcal{S})\right] = \sqrt{\widehat{Var}\left[\widetilde{a}_{HT}(\mathcal{S})\right]} = \sqrt{12} = 3.4641$

(c) An *approximate* 95% confidence interval for the true population average is therefore

$$a_{HT}(\mathcal{S}) \pm 2 \times SE\left[\widetilde{a}_{HT}(\mathcal{S})\right] = 15 \pm 2 \times \sqrt{12} = [8.0718, 21.9282]$$

QUESTION 6 [7 POINTS]

Interest lies in comparing two sub-populations $\mathcal{P}_1 = \{1, 4\}$ and $\mathcal{P}_2 = \{2, 3, 5\}$ by way of a permutation test.

(a) [1] State the null hypothesis H_0 associated with this test.

(b) [1] Using the discrepancy measure $D(\mathcal{P}_1, \mathcal{P}_2) = |\bar{y}_1 - \bar{y}_2|$, calculate the observed discrepancy.

(c) [5] By considering *all permutations*, calculate the p-value associated with this test.

COMMENTS & SOLUTIONS

(a)

$$H_0 : \mathcal{P}_1 \text{ and } \mathcal{P}_2 \text{ are sampled from the same population}$$

Alternatively, something along the lines of $\mathcal{P}_1$ and $\mathcal{P}_2$ being indistinguishable would also be okay. However, the null hypothesis should *not* be stated in terms of the equality of sub-population attributes. That would be worthh 0 points.

(b)

$$\begin{aligned} d_{obs} &= D(\mathcal{P}_1, \mathcal{P}_2) \\ &= |\overline{y}_1 - \overline{y}_2| \\ &= |\frac{1+4}{2} - \frac{2+3+5}{3}| \\ &= |-\frac{5}{6}| \\ &= 5/6 \\ &= 0.8333 \end{aligned}$$

(c) There are exactly $\binom{5}{2} = \binom{5}{3} = 10$ unique shufflings of these two sub-populations. Each of these rearrangments and their corresponding discrepancy measure values are shown in the table below.

$\mathcal{P}^*_{1,1} = \{1,2\}$	$\mathcal{P}^*_{2,1} = \{3,4,5\}$	$D_1 = 5/2$
$\mathcal{P}^*_{1,2} = \{1,3\}$	$\mathcal{P}^*_{2,2} = \{2,4,5\}$	$D_2 = 5/3$
$\mathcal{P}^*_{1,3} = \{1,4\}$	$\mathcal{P}^*_{2,3} = \{2,3,5\}$	$D_3 = 5/6$
$\mathcal{P}^*_{1,4} = \{1,5\}$	$\mathcal{P}^*_{2,4} = \{2,3,4\}$	$D_4 = 0$
$\mathcal{P}^*_{1,5} = \{2,3\}$	$\mathcal{P}^*_{2,5} = \{1,4,5\}$	$D_5 = 5/6$
$\mathcal{P}^*_{1,6} = \{2,4\}$	$\mathcal{P}^*_{2,6} = \{1,3,5\}$	$D_6 = 0$
$\mathcal{P}^*_{1,7} = \{2,5\}$	$\mathcal{P}^*_{2,7} = \{1,3,4\}$	$D_7 = 5/6$
$\mathcal{P}^*_{1,8} = \{3,4\}$	$\mathcal{P}^*_{2,8} = \{1,2,5\}$	$D_8 = 5/6$
$\mathcal{P}^*_{1,9} = \{3,5\}$	$\mathcal{P}^*_{2,9} = \{1,2,4\}$	$D_9 = 5/3$
$\mathcal{P}^*_{1,10} = \{4,5\}$	$\mathcal{P}^*_{2,10} = \{1,2,3\}$	$D_{10} = 5/2$

The p-value is the fraction of these D values greater than or equal to $d = 5/6$:

$$\begin{aligned} p\text{-value} &= \frac{1}{10}\sum_{i=1}^{10} I_{[5/6,\infty)}(D_i) \\ &= \frac{1+1+1+0+1+0+1+1+1+1}{10} \\ &= \frac{8}{10} \\ &= 0.8 \end{aligned}$$

QUESTION 7 [12 POINTS]

Suppose that the sample $\mathcal{S} = \{1, 2, 3\}$ is selected from a population $\mathcal{P}$ and interest lies in calculating a bootstrap-based confidence interval, so the following $B = 10$ bootstrap samples are obtained.

$$\begin{array}{ll} \mathcal{S}_1^\star = \{1,1,1\} & \mathcal{S}_6^\star = \{2,3,3\} \\ \mathcal{S}_2^\star = \{1,1,2\} & \mathcal{S}_7^\star = \{3,3,3\} \\ \mathcal{S}_3^\star = \{1,2,2\} & \mathcal{S}_8^\star = \{1,1,3\} \\ \mathcal{S}_4^\star = \{2,2,2\} & \mathcal{S}_9^\star = \{1,3,3\} \\ \mathcal{S}_5^\star = \{2,2,3\} & \mathcal{S}_{10}^\star = \{1,2,3\} \end{array}$$

(a) [5] Using the naive normal theory approach, calculate an 80% confidence interval for the population variance

$$a(\mathcal{P}) = \frac{\sum_{u \in \mathcal{P}} (y_u - \overline{y})^2}{N}.$$

(b) [1] Using the percentile method, calculate an 80% confidence interval for the population variance

$$a(\mathcal{P}) = \frac{\sum_{u \in \mathcal{P}} (y_u - \overline{y})^2}{N}.$$

(c) [1] Using the percentile method, calculate an 80% confidence interval for the population standard deviation

$$a(\mathcal{P}) = \sqrt{\frac{\sum_{u \in \mathcal{P}} (y_u - \overline{y})^2}{N}}.$$

(d) [5] In point form, describe the bootstrap-t method for confidence interval calculation for some population attribute $a(\mathcal{P})$. Define any notation that you introduce.

COMMENTS & SOLUTIONS

(a) The formula for an 80% naive normal theory confidence interval for $a(\mathcal{P})$ is

$$a(\mathcal{S}) \pm 1.2815 \times \widehat{SD}_\star \left[\widetilde{a}(\mathcal{S})\right]$$

where $a(\mathcal{S})$ is the attribute calculated on the sample, 1.2815 is the critical value from the $N(0,1)$ that provides 80% confidence and $\widehat{SD}_\star \left[\widetilde{a}(\mathcal{S})\right]$ is the bootstrap standard deviation. The attribute calculated on the sample is

$$a(\mathcal{S}) = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = \frac{2}{3}$$

In order to calculate the bootstrap standard deviation we must first calculate the attribute on all of the bootstrap samples. Each of these variances is shown in the table below.

$a_1^\star = a(\mathcal{S}_1^\star) = 0$	$a_6^\star = a(\mathcal{S}_6^\star) = 2/9$
$a_2^\star = a(\mathcal{S}_2^\star) = 2/9$	$7_1^\star = a(\mathcal{S}_7^\star) = 0$
$a_3^\star = a(\mathcal{S}_3^\star) = 2/9$	$8_1^\star = a(\mathcal{S}_8^\star) = 8/9$
$a_4^\star = a(\mathcal{S}_4^\star) = 0$	$a_9^\star = a(\mathcal{S}_9^\star) = 8/9$
$a_5^\star = a(\mathcal{S}_5^\star) = 2/9$	$a_{10}^\star = a(\mathcal{S}_{10}^\star) = 2/3$

The average of these ten values is $\overline{a}^\star = \frac{1}{3}$. The bootstrap standard deviation can now be calculated:

$$\begin{aligned}\widehat{SD}_\star\left[\widetilde{a}(\mathcal{S})\right] &= \sqrt{\frac{\sum_{b=1}^{10}\left(a(\mathcal{S}_b^\star)-\overline{a}^\star\right)^2}{9}} \\ &= \sqrt{\frac{10}{81}} \\ &= 0.3514\end{aligned}$$

Substituting all of the necessary pieces into the formula above yields the 80% confidence interval for $a(\mathcal{P})$:

$$a(\mathcal{S}) \pm 1.2815 \times \widehat{SD}_\star\left[\widetilde{a}(\mathcal{S})\right] = \frac{2}{3} \pm 1.2815 \times \sqrt{\frac{10}{81}} = [0.2614, 1.1170]$$

(b) The percentile-based confidence interval is determined by finding appropriate quantiles from the bootstrap distribution $a_1^\star, a_2^\star, \ldots, a_{10}^\star$. Because we want an 80% CI we must determine $Q_a(0.1)$ and $Q_a(0.9)$, the 10^{th} and 90^{th} quantiles.

$$0 \quad 0 \quad 0 \quad \frac{2}{9} \quad \frac{2}{9} \quad \frac{2}{9} \quad \frac{2}{9} \quad \frac{2}{3} \quad \frac{8}{9} \quad \frac{8}{9}$$

The ten values are arranged above from smallest to largest, the 10^{th} and 90^{th} quantiles are easily found to be the first and last numbers, i.e., $Q_a(0.1) = 0$ and $Q_a(0.9) = 8/9$. Thus the 80% CI is given by

$$[0, 8/9] = [0, 0.8889].$$

(c) Since percentile method CI's are equivariant to 1:1 transformations the 80% CI for $SD(\mathcal{P}) = \sqrt{Var(\mathcal{P})}$ is

$$[\sqrt{0}, \sqrt{8/9}] = [0, 0.9428].$$

(d)

- For a sample $\mathcal{S}$
  - Calculate $a(\mathcal{S})$
  - Obtain B bootstrap samples $\mathcal{S}_1^\star, \mathcal{S}_2^\star, \ldots, \mathcal{S}_B^\star$
- For each bootstrap sample
  - Calculate $a(\mathcal{S}_b^\star)$ and $\widehat{SD}\left[a(\mathcal{S}_b^\star)\right]$ and

$$z_b = \frac{a(\mathcal{S}_b^\star) - a(\mathcal{S})}{\widehat{SD}\left[a(\mathcal{S}_b^\star)\right]}$$

    Note that calculating $\widehat{SD}\left[a(\mathcal{S}_b^\star)\right]$ will require either knowing a closed-form expression for this quantity or the double bootstrap.
- From the values $z_1, z_2, \ldots, z_B$ determine

$$c_{lower} = Q_z(p/2) \text{ and } c_{upper} = Q_z(1-p/2)$$

- Calculate the bootstrap standard deviation $\widehat{SD}_\star\left[\widetilde{a}(\mathcal{S})\right]$

- The $(1-p)\times 100\%$ bootstrap-t confidence interval for $a(\mathcal{P})$ is

$$\left[a(\mathcal{S}) - c_{upper} \times \widehat{SD}_\star\left[\widetilde{a}(\mathcal{S})\right], a(\mathcal{S}) - c_{lower} \times \widehat{SD}_\star\left[\widetilde{a}(\mathcal{S})\right]\right]$$

Note that a student's bullet points do not need to match mine exactly, but they should follow a logical order and describe how each component of the CI is determnined.

QUESTION 8 [10 POINTS]

In class we saw that the average prediction squared error ($APSE$)

$$APSE(\mathcal{P}, \widetilde{\mu}) = \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{u \in \mathcal{P}} (y_u - \widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u))^2$$

could be decomposed into three interpretable components. In this question, you are going to prove each step. Your notation should follow that of the notes and each simplification must be justified mathematically.

(a) [5] Prove that

$$\begin{aligned} \tfrac{1}{N_{\mathcal{S}}} \textstyle\sum_{j=1}^{N_{\mathcal{S}}} \tfrac{1}{N} \textstyle\sum_{u \in \mathcal{P}} (y_u - \widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u))^2 \quad &= \quad \tfrac{1}{N_{\mathcal{S}}} \textstyle\sum_{j=1}^{N_{\mathcal{S}}} \tfrac{1}{N} \textstyle\sum_{u \in \mathcal{P}} (y_u - \tau(\mathbf{x}_u))^2 \\ &\quad + \tfrac{1}{N_{\mathcal{S}}} \textstyle\sum_{j=1}^{N_{\mathcal{S}}} \tfrac{1}{N} \textstyle\sum_{u \in \mathcal{P}} (\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u) - \tau(\mathbf{x}_u))^2 \end{aligned}$$

(b) [5] Prove that

$$\begin{aligned} \tfrac{1}{N_{\mathcal{S}}} \textstyle\sum_{j=1}^{N_{\mathcal{S}}} \tfrac{1}{N} \textstyle\sum_{u \in \mathcal{P}} (\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u) - \tau(\mathbf{x}_u))^2 \quad &= \quad \tfrac{1}{N_{\mathcal{S}}} \textstyle\sum_{j=1}^{N_{\mathcal{S}}} \tfrac{1}{N} \textstyle\sum_{u \in \mathcal{P}} (\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u) - \overline{\widehat{\mu}}(\mathbf{x}_u))^2 \\ &\quad + \tfrac{1}{N_{\mathcal{S}}} \textstyle\sum_{j=1}^{N_{\mathcal{S}}} \tfrac{1}{N} \textstyle\sum_{u \in \mathcal{P}} (\overline{\widehat{\mu}}(\mathbf{x}_u) - \tau(\mathbf{x}_u))^2 \end{aligned}$$

COMMENTS & SOLUTIONS

(a) It will be useful to define some functions and notations *a priori*. First, let us recognize that because $\mathcal{P}$ is finite, there are a finite number of distinct $\mathbf{x}$ values, which we will call $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K$. The population $\mathcal{P}$ can then be partitioned into disjoint groups of units that all share the same $\mathbf{x}$ values:

$$\mathcal{P} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots \cup \mathcal{A}_K$$

where $\mathcal{A}_k = \{u \in \mathcal{P} | \mathbf{x}_u = \mathbf{x}_k\}$ for $k = 1, 2, \ldots, K$. Recognizing this allows us to rewrite sums over $u \in \mathcal{P}$ as sums over $u \in \mathcal{A}_k$, $k = 1, 2, \ldots, K$.

Let us also define $\tau(\mathbf{x}_k)$ to be the conditional average

$$\tau(\mathbf{x}_k) = \frac{1}{n_k} \sum_{u \in \mathcal{A}_k} y_u$$

where n_k is the number of units in $\mathcal{A}_k$. Let us now prove what is required.

$$\begin{aligned} \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{u \in \mathcal{P}} (y_u - \widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u))^2 \quad &= \quad \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{u \in \mathcal{P}} \left[(y_u - \tau(\mathbf{x}_u)) - (\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u) - \tau(\mathbf{x}_u)) \right]^2 \\ &= \quad \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{u \in \mathcal{P}} (y_u - \tau(\mathbf{x}_u))^2 - 2 \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{u \in \mathcal{P}} (y_u - \tau(\mathbf{x}_u))(\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u) - \tau(\mathbf{x}_u)) \\ &\quad + \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{u \in \mathcal{P}} (\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u) - \tau(\mathbf{x}_u))^2 \end{aligned}$$

The identity is proven if we can show that the cross product term $\frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{u \in \mathcal{P}} (y_u - \tau(\mathbf{x}_u))(\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u) - \tau(\mathbf{x}_u)) = 0$. We do so by rewriting the sum over $u \in \mathcal{P}$ as sums over $u \in \mathcal{A}_k$, $k = 1, 2, \ldots, K$ as follows.

$$\begin{aligned}\frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}}\frac{1}{N}\sum_{u\in\mathcal{P}}(y_u-\tau(\mathbf{x}_u)) &= \frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}}\frac{1}{N}\sum_{k=1}^{K}\sum_{u\in\mathcal{A}_k}(y_u-\tau(\mathbf{x}_k))(\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_k)-\tau(\mathbf{x}_k))\\ &= \frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}}\frac{1}{N}\sum_{k=1}^{K}(\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_k)-\tau(\mathbf{x}_k))\sum_{u\in\mathcal{A}_k}(y_u-\tau(\mathbf{x}_k))\\ &= 0\end{aligned}$$

since $\sum_{u\in\mathcal{A}_k}(y_u-\tau(\mathbf{x}_k))=0$ because

$$\tau(\mathbf{x}_k)=\frac{1}{n_k}\sum_{u\in\mathcal{A}_k}y_u$$

(b) It will be useful to define some functions and notations *a priori*. First, let us define

$$\overline{\widehat{\mu}}(\mathbf{x})=\frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}}\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x})$$

as the average predictor function, averaged over all of the samples taken. Let us now prove what is required.

$$\begin{aligned}\frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}}\frac{1}{N}\sum_{u\in\mathcal{P}}(\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u)-\tau(\mathbf{x}_u))^2 &= \frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}}\frac{1}{N}\sum_{u\in\mathcal{P}}\left[(\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u)-\overline{\widehat{\mu}}(\mathbf{x}_u))+(\overline{\widehat{\mu}}(\mathbf{x}_u)-\tau(\mathbf{x}_u))\right]^2\\ &= \frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}}\frac{1}{N}\sum_{u\in\mathcal{P}}(\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u)-\overline{\widehat{\mu}}(\mathbf{x}_u))^2\\ &\quad+2\frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}}\frac{1}{N}\sum_{u\in\mathcal{P}}(\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u)-\overline{\widehat{\mu}}(\mathbf{x}_u))(\overline{\widehat{\mu}}(\mathbf{x}_u)-\tau(\mathbf{x}_u))\\ &\quad+\frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}}\frac{1}{N}\sum_{u\in\mathcal{P}}(\overline{\widehat{\mu}}(\mathbf{x}_u)-\tau(\mathbf{x}_u))^2\end{aligned}$$

The identity is proven if we can show that the cross product term $\frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}}\frac{1}{N}\sum_{u\in\mathcal{P}}(\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u)-\overline{\widehat{\mu}}(\mathbf{x}_u))(\overline{\widehat{\mu}}(\mathbf{x}_u)-\tau(\mathbf{x}_u))=0$. We do so by interchanging the order of the sums as follows.

$$\begin{aligned}&\frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}}\frac{1}{N}\sum_{u\in\mathcal{P}}(\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u)-\overline{\widehat{\mu}}(\mathbf{x}_u))(\overline{\widehat{\mu}}(\mathbf{x}_u)-\tau(\mathbf{x}_u))\\ &\qquad= \frac{1}{N}\sum_{u\in\mathcal{P}}(\overline{\widehat{\mu}}(\mathbf{x}_u)-\tau(\mathbf{x}_u))\frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}}(\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u)-\overline{\widehat{\mu}}(\mathbf{x}_u))\\ &\qquad= 0\end{aligned}$$

since $\sum_{j=1}^{N_{\mathcal{S}}}(\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_u)-\overline{\widehat{\mu}}(\mathbf{x}_u))=0$ because

$$\overline{\widehat{\mu}}(\mathbf{x})=\frac{1}{N_{\mathcal{S}}}\sum_{j=1}^{N_{\mathcal{S}}}\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x})$$

QUESTION 9 [10 POINTS]

(a) [5] In your own words, describe what is meant by the term **overfitting** (as it relates to this course).

(b) [5] In your own words, describe what is meant by the phrase **bias-variance trade-off** (as it relates to this course).

COMMENTS & SOLUTIONS

(a) Overfitting is the phenomen whereby a model has been made overly complex such that it has been tuned to the peculiarities of the sample data used to fit it, and the model's performance on out-of-sample predictions has deteriorated.

Generally speaking, an acceptable response is one the communicates the risk of increasing model complexity to a point where in-sample predictions are very good, but out-of-sample predictions are very poor.

(b) The bias-variance trade-off is a phenomenon that must be considered when doing model selection based on predictive accuracy. This recognizes that for uncomplicated models bias is high and variance low, but for overly complicated models bias is low and variance is high. Thus the optimal model complexity (which minimizes out-of-sample prediction error) is one that simultaneously minimizes both bias *and* variance.

Generally speaking, an acceptable response is one that communicates the need to balance bias and variance when choosing a model so as to achieve good predictive performance. Answers could, but don't need to, include a plot of prediction error / bias / variance versus model complexity. If such a plot is included, deduct points if the general patterns are not represented correctly.

QUESTION 10 [5 POINTS]

Suppose that population $\mathcal{P}$ is of the form $\{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \ldots, (y_N, \mathbf{x}_N)\}$ and you wish to predict y from $\mathbf{x}$ via a polynomial predictor function $\mu(\mathbf{x})$. Unfortunately the entire population is not available for you to study and instead you only have access to the sample $\mathcal{S}$. Using cross validation you calculate $APSE$ and determine the optimal degree for your polynomial $\widehat{\mu}(\mathbf{x})$. However, as an astute statistican, you recognize that this choice of an "optimal" degree is subject to sampling variation.

Describe how you might construct a $(1-p) \times 100\%$ confidence interval for the polynomial degree.

COMMENTS & SOLUTIONS

For the given sample $\mathcal{S}$

- Generate B bootstrap samples $\mathcal{S}_1^\star, \mathcal{S}_2^\star, \ldots, \mathcal{S}_B^\star$ by sampling with replacement from $\mathcal{S}$
- For each bootstrap sample $\mathcal{S}_b^\star$ use cross validation to estimate APSE and hence pick the polynomial degree $d^\star$ yielding the fitted polynimial $\widehat{\mu}_{\mathcal{S}_b^\star}(\mathbf{x})$
- Find the $p/2$ and $1-p/2$ quantiles of $d_1^\star, d_2^\star, \ldots, d_B^\star$ and define

$$d_{lower} = Q_d(p/2) \text{ and } d_{upper} = Q_d(1-p/2)$$

- The $(1-p) \times 100\%$ CI for polynomial degree is

$$[d_{lower}, d_{upper}]$$

Note that any version of a bootstrap CI would also work.