STAT 341: Final Take Home Assessment

DUE: Monday April 20 by 11:59pm EST

INSTRUCTIONS

Your assessment must be submitted by the due date listed at the top of this document, and it must be submitted electronically via Crowdmark. Please ensure that your submission meets the following criteria:

- Like an in-person examination, **your solutions must be handwritten**. Physical handwritten solutions on paper may be scanned or photographed and uploaded. Digital handwritten solutions from a tablet are also fine.

- **NOTE:** It is your responsibility to ensure that your solutions can be **easily** read. If a solution cannot be **easily** read it simply will not be marked. This means:

  (i) Scans/photos are of high resolution
  (ii) Scans/photos are sufficiently bright
  (iii) Your handwriting is legible.

- As per the directive above, any calculation I require you to do must be done by hand. You may use R to verify your solution, but your solution should not contain any R code or output.

- Incorrect answers may receive partial credit if your work is shown. An incorrect answer with no work shown will receive 0 points.

- Solutions for different questions must begin on different pages.

- Different parts of the same question must be *clearly* identified and responded to in the order they were asked.

- Numeric answers should be rounded to four decimal places (unless the answer is exact to fewer than four decimal places).

MARKING

Questions are worth 1 or 5 points.

- Questions worth 1 point are trivial; you will be awarded 1 point if your answer is correct and 0 points otherwise.
- Questions worth 5 points (that are attempted) will 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

ACADEMIC INTEGRITY STATEMENT

Due to the increased risk of academic dishonesty in the context of take home assessments, you must print, sign, and upload this page as a declaration of the academic integrity of your submission.

My signature belows signifies that I _____ (print your name) declare each of the following statements to be true:

- The work I submit here is entirely my own.
- I have not used any unauthorized aids.
- I have not discussed and will not discuss the contents of this assessment with anyone until after the submission deadline.
- I am aware that misconduct related to final assessments can result in significant penalties, including failing the course and suspension (as covered in Policy 71: https://uwaterloo.ca/secretariat/policies-procedures-guidelines/policy-71).

Signature: _____

Date: _____

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

Some places you might consider looking:

- Kaggle
- UCI Machine Learning Repository
- r/datasets
- data.gov
- KDnuggets

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 $\mathcal{S} = \{y_1, y_2, \ldots, y_{n-1}\}$.

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.

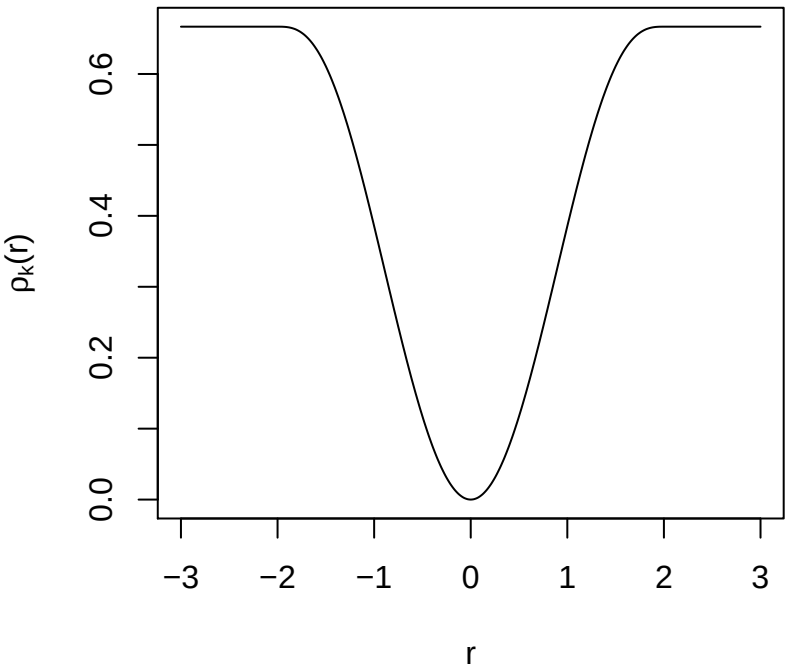
Another objective function that similarly facilitates robust regression is the **Tukey Biweight objective function**:

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

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

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

This function, for $k = 2$, is visualized below.



(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.

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 $\pi_{uv} = P(u \in \mathcal{S}, v \in \mathcal{S})$.

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.

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.

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.

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

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

(b) [5] Prove that

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

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).

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.