STA 141A - Spring 2025 - Mock Midterm exam 2

Instructions: This midterm exam is a closed-book exam, it is scheduled for 40 minutes and written in class. Except for a pen/pencil, one hand-written cheat-sheet (both-sided), and a non-graphing calculator, no other materials are allowed. The total score is 100 points. You must explain your solutions for full credit. Partial credit can only be given if your thoughts can be followed. Make sure your name and your student ID is written on the first page.

Good luck!

Name:

Student ID:

Score

Note: Due to this course's grading scheme, the actual midterm will be more difficult than this practice one.

1. Linear regression

a. Name two ways that linear regression might not be a good fit to the data.

b. If a fitted linear model produces coefficient estimates $\hat{\beta}_0 = -1$, $\hat{\beta}_1 = 1$, what value does the model predict at $x_1 = 9$?

c. Explain the difference between R^2 and adjusted R^2.

d. What should we conclude if we see that the variance inflation factor for $\hat{\beta}_1$ is $VIF(\hat{\beta}_1) = 50$?

2. Classification

We study the effect of daily meditation (X) (in minutes per day) on concentration (Y), where we code concentration as "being focused" (class 0) and "being distracted" (class 1). We have the following data:

$$(x_1, y_1) = (5, 1), (x_2, y_2) = (4, 1), (x_3, y_3) = (11, 0), (x_4, y_4) = (9, 0), (x_5, y_5) = (6, 1).$$

a. Using the linear discriminant analysis (LDA) approach, compute the discriminant functions for both classes. Using your computed discriminant functions, explain to which class you would assign a person who meditates 4 minutes per day.

b. Using the estimates $\hat{\beta}_0 = 113.66$ and $\hat{\beta}_1 = -15.17$ from a logistic regression model fitted to this data, explain to which class you would assign a person who meditates 4 minutes per day according to the following rule: we assign anyone who meditates x minutes per day to class 1 if and only if $P(Y = 1|X = x) \geq 0.5$.

c. For the following confusion matrix, compute the prediction accuracy, the number of false negatives, and the number of false positives if we consider class 0 as "-"/"Null" and class 1 as "+"/"non-Null".

```
##      true
## pred 0 1
##    0 3 1
##    1 2 4
```

3. Bootstrap

Consider the following data set with $n = 2$ and $p = 1$:

```
data.frame(index=1:2, x=6:7, y=-6:-7)
```

```
##   index x  y
## 1     1 6 -6
## 2     2 7 -7
```

Draw all possible unique bootstrap datasets of size $n = 2$ from this dataset. For each bootstrap dataset, also state the probability of drawing it.

4. Cross-validation

We want to estimate an unknown function f that describes the response Y by $Y = f(X) + \varepsilon$, where X is the predictor and ε is an error term with $Var(\varepsilon) = 1$. We consider four different estimation approaches. Based on training data with $n = 100$ data points, the four approaches yield the respective estimates $\hat{f}_1$, $\hat{f}_2$, $\hat{f}_3$, and $\hat{f}_4$. We also computed the following values for MSE_{train}:

	$i = 1$	$i = 2$	$i = 3$	$i = 4$
Training MSE for $\hat{f}_i$	0.4	0.9	1.1	1.2

a. Suppose we want to estimate the test MSE by the validation set method. Will the training MSE of approach 1 on the validation set's training data set likely be larger, smaller, or about the same as 0.4?

b. What cross-validation methods could we instead use to estimate the test MSE, and why might we use them over the validation set method?

c. Suppose we use a method you mentioned in part b, and found that the estimated test MSE is lowest for approach 3. What does that suggest about estimator 1?

5. Clustering

Consider the following data set with $n = 4$:

```
data.frame(x1=c(9,3,1,2), x2=c(8,1,3,2))
```

```
##   x1 x2
## 1  9  8
## 2  3  1
## 3  1  3
## 4  2  2
```

Write all possible ways to partition this data into two clusters. Which partition/clustering seems like it has the smallest within-cluster variation? Justify your answer. (No need to do computations here; you can draw a picture to help you justify your answer.)