Name:

UTeid:

M339G Predictive Analytics
University of Texas at Austin
In-Term Exam I
Instructor: Milica Čudina

Notes: This is a closed book and closed notes exam. The maximal score on this exam is 100 points. **There are many ways in which any single problem can be solved. The solutions herein are just one possible way to tackle the given problems.**

All written work handed in by the student is considered to be **their own work, prepared without unauthorized assistance.**

The University Code of Conduct

”The core values of The University of Texas at Austin are learning, discovery, freedom, leadership, individual opportunity, and responsibility. Each member of the university is expected to uphold these values through integrity, honesty, trust, fairness, and respect toward peers and community. As a student of The University of Texas at Austin, I shall abide by the core values of the University and uphold academic integrity.”

”I agree that I have complied with the UT Honor Code during my completion of this exam.”

Signature:

1.1. **DEFINITIONS.**

Problem 1.1. (10 points) Provide the definition of *bias*.

Solution: See the solutions to the first homework assignment.

Problem 1.2. (10 points) Provide the definition of the *mean-squared error* in the context of parameter estimation.

Solution: See the solutions to the first homework assignment.

1.2. **CONCEPTUAL QUESTIONS.**

Problem 1.3. (10 points) Provide an example of a situation where principal component regression would be more appropriate then multiple linear regression. You can use examples from class or variations thereof.

Solution: Solutions will vary. The salient point of any response which is to earn credit must be that it is desirable to both reduce the dimension of the set of predictors (search ”curse of dimensionality”, if you want) and to get rid of collinearity. Our example from class could be the seating position data set with a variety of highly associated biometric predictors.

Problem 1.4. (10 points) Describe the difference between classification and clustering.

Solution: Solutions will vary. The salient point of any response which is to earn credit must be that classification has a response variable available (supervised learning) whereas clustering does not (unsupervised learning).

1.3. **FREE RESPONSE PROBLEMS.** Please, explain carefully all your statements and assumptions. Numerical results or single-word answers without an explanation (even if they’re correct) are worth 0 points.

Problem 1.5. (10 points)Consider a simple linear regression fitted on 20 observations. In our usual notation, you are given the following:

- $\sum(y_i - \hat{y}_i)^2 = 10$
- $\sum(\hat{y}_i - \bar{y})^2 = 112$

Find the coefficient of determination.

Solution: You should know that

$$TSS = \sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 = RSS + \sum (\hat{y}_i - \bar{y})^2.$$

The first given fact tells us that the RSS equals 10. The second given fact tells us that $TSS - RSS$ equals 112. So,

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{TSS - RSS}{TSS} = \frac{112}{122}\,.$$

For your convenience, I am providing the proof of the above equality. You did not need to prove this fact in the exam. Proving

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

is equivalent to proving

$$\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0.$$

As we did in class, let us denote $\varepsilon_i = y_i - \hat{y}_i$ for all $i = 1, \dots, n$. Recall that the sum of residuals is equal to zero, i.e., $\sum \varepsilon_i = 0$. Also, by the least-squares condition, $\sum \varepsilon_i x_i = 0$. Then,

$$\begin{aligned}
\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum \varepsilon_i (\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \bar{x}) \\
&= \sum \varepsilon_i (\beta_1 x_i - \beta_1 \bar{x}) \\
&= \beta_1 \sum \varepsilon_i x_i - \beta_1 \bar{x} \sum \varepsilon_i = 0.
\end{aligned}$$

Problem 1.6. (15 points) *Source: An old SOA exam.*
Consider a multiple linear regression where predictor X_1 stands for the amount of precipitation in a month (in inches), predictor X_2 stands for the traffic volume in a month, and predictor X_3 stands for the indicator of whether a holiday weekend occurred during a particular month.

The response variable Y corresponds to the number of fatal car accidents.

You fit the data for the past year, and get the following coefficients

Intercept	-2.358
Precipitation	0.245
Volume	1.129
Holiday	2.334

(i) (5 points) What is your prediction for a month with no holiday weekends, with the precipitation equal to 1.5 inches, and the traffic volume equal to 4?
(ii) (5 points) How do you interpret the coefficient for *Precipitation*?
(iii) (5 points) How do you interpret the coefficient for *Holiday*?

Solution:

$$-2.358 + 0.245(1.5) + 1.129(4) = 2.5255$$

The coefficient for *Precipitation* tells us that for every additional inch of precipitation there are 0.245 more accidents per our model on average (with all else kept fixed).

The coefficient for *Holiday* tells us that months with holiday weekends have 2.334 more fatal car accidents than comparable months without holiday weekends.

Problem 1.7. (10 points) You are using $K-$nearest neighbors in a classification problem with $X = (X_1, X_2)$ as predictors and Y as the response. Here are the observed values:

x_1	0	2	5	6
x_2	1	2	4	1
y	1	2	1	2

Using $K = 3$, figure out how the above points would be classified and the misclassification error. Then, state how you would classify point $(2, 4)$.

Hint: Draw a picture in the plane of (x_{i1}, x_{i2}) *for* $i = 1, 2, 3, 4$.

Solution: With the neighbourhood of size 3, we get the following predictions:

x_1	0	2	5	6
x_2	1	2	4	1
y	1	1	2	2

We did fine: the misclassification error rate is $1/2$. The new point would be classified as 1.

Problem 1.8. (10 points) *Source: MAS-I, Spring 2018.*
You are given the following information about an insurance policy:

(i) The probability of a policy renewal $p(X)$ follows a logistic model with an intercept and one explanatory variable.
(ii) $\hat{\beta}_0 = 5$
(iii) $\hat{\beta}_1 = -0.65$

Calculate the fitted odds of renewal at $x = 5$.

Solution: In the logistic model, the odds are $e^{\beta_0+\beta_1 x}$. So, the coefficients given above yield

$$e^{5-0.65(5)} = 5.754603$$

1.4. **MULTIPLE CHOICE QUESTIONS.**

Problem 1.9. (5 points) Here is an example of a statistical problem.

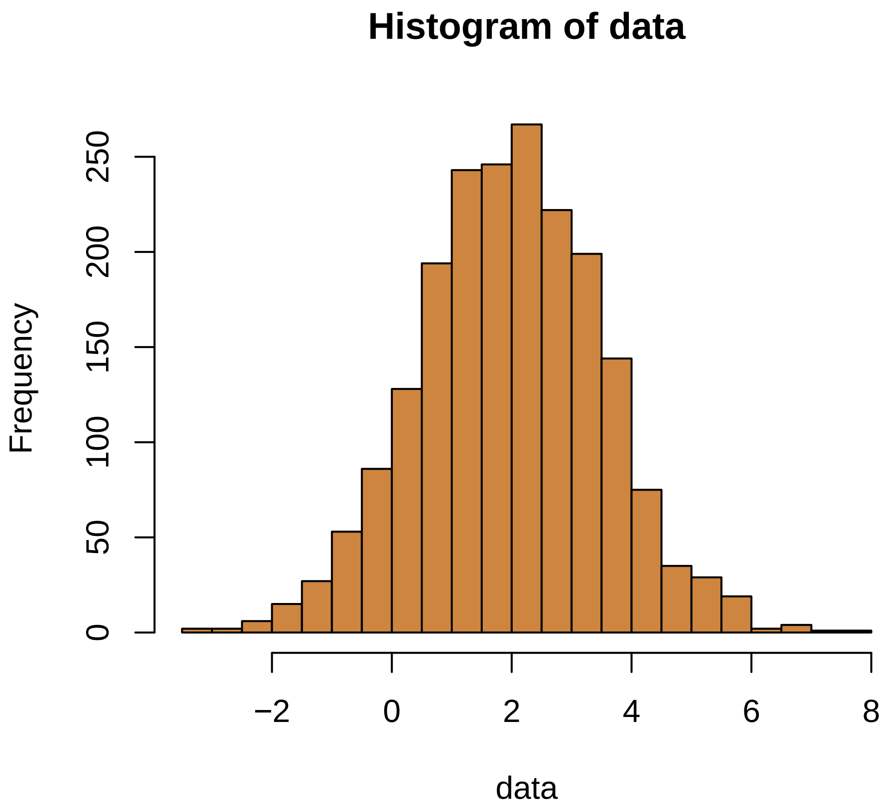
You are trying to model the length of a family's vacation based on these predictors: type of vacation (mountains, beach, culture), and the number of family members.

Which of the following procedures is the most applicable in this case?

(a) Multiclass logistic regression.
(b) Simple linear regression.
(c) Multiple linear regression with two categorical predictors.
(d) Multiple linear regression with one categorical and one numerical predictor.
(e) None of the above techniques apply in this case.

Solution: (d)

Problem 1.10. (5 points) You have a sample of size 252 from a distribution that you know from past experience looks like this:



Your task is to estimate its mean. Of the following, what procedure(s) would be acceptable in this case? Choose **all** that apply.

(a) A 95% bootstrap confidence interval using quantiles.

(b) Using the 't.test' command in **R**.

(c) A $2SE$ bootstrap confidence interval.

(d) The standard $z-$procedure 95%-confidence interval.

(e) None of the above.

Solution: (a, b, c, d)

Problem 1.11. (5 points) *Source: SRM Sample Questions.*
Consider the following statements:

I The proportion of variance explained by an additional principal component never decreases as more principal components are added.

II The cumulative proportion of variance explained never decreases as more principal components are added.

III Loading vectors of different principal components are orthogonal.

IV Scree plots help us determine the number of principal components to use.

Which of the statements is **FALSE**?

(a) I only

(b) II only

(c) III only

(d) I, II, and III

(e) None of the above.

Solution: (a)
By the PCA algorithm, I is **FALSE**. In fact, it's the opposite. Statement II is **TRUE** as are statements III and IV.