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.

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) Describe at least two reasons why we wish to reduce the dimensionality of the problem, for instance, why we would want to use PCA or drop some of the predictors in multiple linear regression.

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 *variance inflation factor* and explain how it is used.

Solution: See page 102 in the textbook.

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)*Source: SRM Sample Problem #18.*
For a simple linear regression model the sum of squares of the residuals is

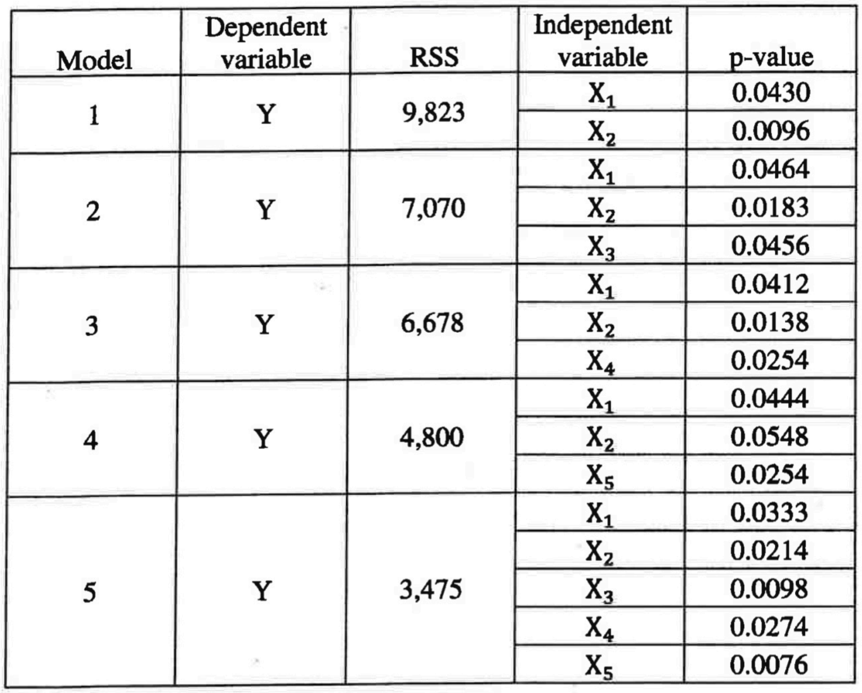
$$\sum_{i=1}^{25} e_i^2 = 230,$$

while the coefficient of determination equals 0.64. Calculate the total sum of squares for this model.

Solution: In our usual notation, we know that the coefficient of determination equals

$$R^2 = 1 - \frac{RSS}{TSS} \quad \Rightarrow \quad TSS = \frac{RSS}{1-R^2} = \frac{230}{1-0.64} = 638.8889.$$

Problem 1.6. (15 points) *Source: MAS-I, Fall 2019, Problem #39.*
An actuary has a data set with one response variable Y and five predictors $(X_1, \dots, X_5)$. She is trying to determine which subset of the predictors best fits the data and is using the **forward** selection method with no stopping rule. Here is a subset of potential models:



Model	Dependent variable	RSS	Independent variable	p-value
1	Y	9,823	X_1	0.0430
			X_2	0.0096
2	Y	7,070	X_1	0.0464
			X_2	0.0183
			X_3	0.0456
3	Y	6,678	X_1	0.0412
			X_2	0.0138
			X_4	0.0254
4	Y	4,800	X_1	0.0444
			X_2	0.0548
			X_5	0.0254
5	Y	3,475	X_1	0.0333
			X_2	0.0214
			X_3	0.0098
			X_4	0.0274
			X_5	0.0076

The procedure just selected *Model 1* as the new candidate model.

Which new variable(s) should be added to the model based on the **forward** selection method? *Justify your response!*

Solution: It should be X_5. Looking at the $p-$values, we see that X_4 and X_5 share the same $p-$value (smaller than that of X_3). However, X_5 also has a smaller RSS.

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	1	1	2	2
x_2	1	4	2	-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	1	1	2	2
x_2	1	4	2	-1
y	1	1	1	1

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

Problem 1.8. (10 points) You run a logistic regression to predict the binary response Y as it depends on a single numerical explanatory variable X. The coefficients you obtain are

You are given the following information about an insurance policy:

(i) $\hat{\beta}_0 = 10$

(ii) $\hat{\beta}_1 = -2$

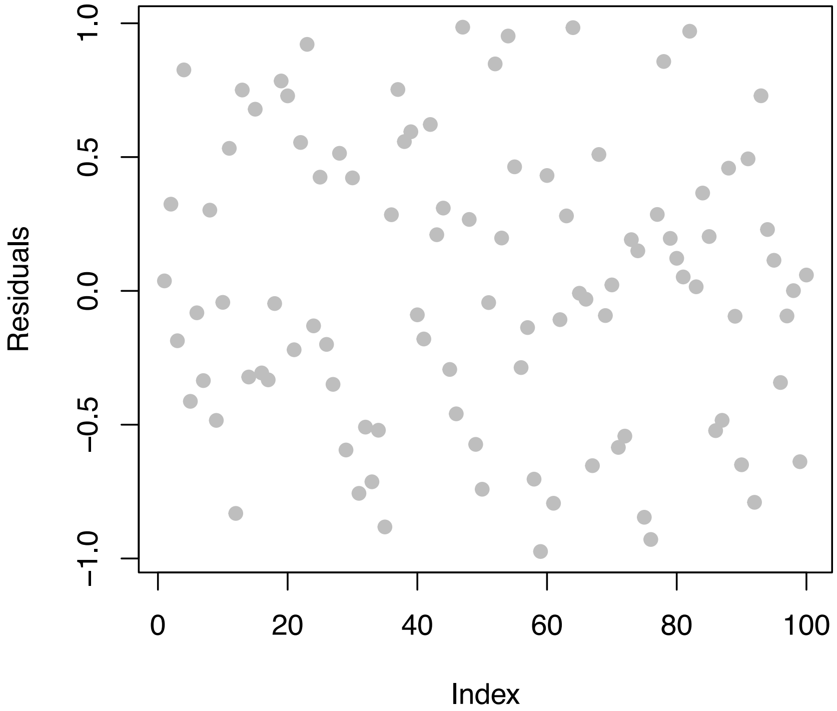
Calculate the fitted odds of success at $X = 5$. In words, how would you describe your result to a layperson?

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

$$e^{10-2(5)} = 1.$$

"Success" and "failure" are equally likely for $X = 5$.

Problem 1.9. (5 points) *Source: MAS-I Exam, Fall 2019, Problem #30.*
An actuary has a data set with one predictor variable X and one response variable Y. She divides the data set randomly into training and testing sets. The training set is used to fit an ordinary least-squares regression. In order to evaluate the fit, the actuary plots the residuals from the plot against the observations of the random variable x.



Which of the following enhancements of the model would most likely improve the fit to the testing data set?

(a) Linear spline.

(b) Polynomial regression.

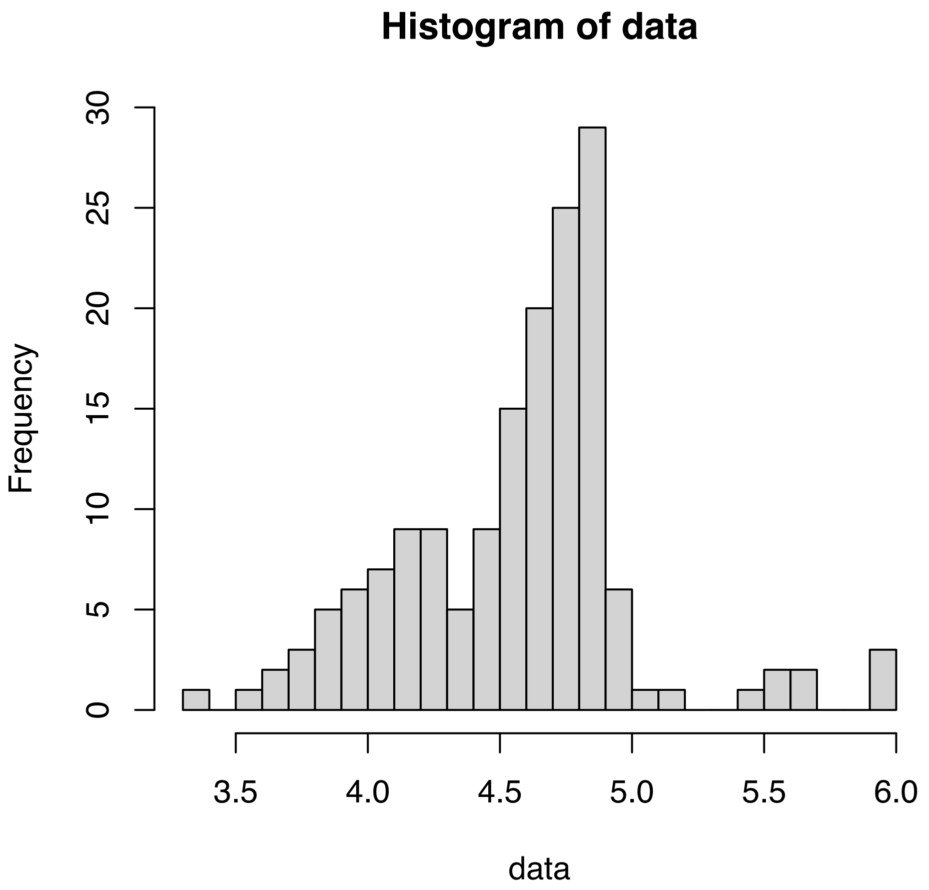
(c) Cubic spline.

(d) Natural spline.

(e) There is no evidence that any of the above techniques would improve the fit.

Solution: (e)

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 decreases as more principal components are added.

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

III The scalar product of two different loading vectors is always 1.

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: (c)
By the PCA algorithm, I is **TRUE**. Statement II is **TRUE** as is statement IV. Statement III is **FALSE**; it should be 0.