# California State Polytechnic University Pomona

# STA 4320 – Midterm Exam II

Prof. He Jiang

Oct 2024

**Name**: ________________________________________

**Student ID Number**: ____________________________

---

This exam contains 7 pages (including this cover page) and 4 questions. Total amount of points is 40.

You have 50 minutes to complete this exam.

You must write your name and student ID number on the top of this page.

You must write your solution in the space provided.

You may use one page of letter-sized written notes, single or double sided.

You may use a scientific calculator (no graphing ones).

You must show your steps in free response questions (unless the question said otherwise); no points will be given for answers that are not supported.

Please write your solutions clearly and legibly; no credit will be given for unclear or illegible solutions.

If you encounter a long decimal, round it to at least 4 digits after the decimal point. For example round $\pi = 3.14159265\ldots$ to 3.1416.

Some R code that might be helpful in the exam:

```
qnorm(0.975) [1] 1.9600
qt(0.975, 13) [1] 2.1604        qt(0.975, 14) [1] 2.1448        qt(0.975, 15) [1] 2.1314
pf(4.3890, 2, 499) [1] 0.9871        pf(4.3890, 4, 499) [1] 0.9983
pf(181.4, 2, 499) [1] 1        pf(181.4, 4, 499) [1] 1
```

1. The Boston dataset[1] consists of housing values of $n = 506$ suburbs of Boston, along with other characteristics for each suburb.

   We would like to investigate the relationship between medv, $Y$ variable, median value of owner-occupied homes, in unit of thousand dollars, with following variables:

   chars, $X_1$, a categorical variable indicating whether the location is next to the Charles River, with 0 indicating no and 1 indicating yes.

   lstat, $X_2$, a numerical variable indicating the proportion of households with lower income, in unit of percentages.

   We want to first investigate the underlying model without interaction:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

   The resulting R output for the command lm(medv $\sim$ chas + lstat, data=Boston) is given below:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.09412    0.56067  60.809  < 2e-16
chas         4.91998    1.06939   4.601 5.34e-06
lstat       -0.94061    0.03804 -24.729  < 2e-16
```

   We then want to investigate the underlying model with the interaction term:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

   The resulting R output for the command lm(medv $\sim$ chas * lstat, data=Boston) is given below:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.76716    0.57018  59.222  < 2e-16
chas         9.82513    2.10320   4.672 3.84e-06
lstat       -0.91498    0.03897 -23.478  < 2e-16
chas:lstat  -0.43288    0.16017  -2.703  0.00711
```

[1] ISLR2 package.

(a) (2 points) In the first model (without interaction), what is the meaning of 4.91998 (in the second row, first column)?

(Given the proportion of households with lower income (and the intercept) in the model), median housing prices next to the Charles River are on average 4.91998 thousand dollars higher than those not next to Charles River.

(b) (2 points) What is an important limitation of the first model (without interaction), when we are trying to investigate the relationship between medv and lstat for the two groups (one next to Charles River and the other not next to Charles River)?

We are assuming the slope (the decrease in medv resulting from the increase in lstat) being the same for houses next to Charles River and houses not next to Charles River.

(c) (2 points) In the second model (with interaction), for houses next to Charles River, what is the average decrease in price (in unit of thousand dollars) resulting from the increase of one unit (in unit of percentages) in the proportion of households with lower income? Please give support to your answer.

For houses next to the River, $X_1 = 1$, so the estimated slope is:

$$\hat{\beta}_2 + \hat{\beta}_3 = -0.91498 - 0.43288 = -1.34786$$

Therefore for houses next to Charles River, the average decrease in price resulting from the increase of one unit in the proportion of households with lower income is 1.34786 thousand dollars.

(d) (2 points) In the second model (with interaction), considering the relationship between medv and lstat, is there a significant difference between houses next to the Charles River and houses not next to the Charles River? Please formalize this into a hypothesis test.

(Given all other variables and the intercept in the model:)

$\mathrm{H}_0 : \beta_3 = 0$ vs $H_1 : \beta_3 \neq 0$

(e) (2 points) At the significance level of $\alpha = 0.05$, please give a conclusion to the previous hypothesis testing in the context of the current question.

The appropriate p value is 0.00711. Therefore we reject $H_0$ and conclude that there is significant difference regarding the slope between houses next to Charles River and houses not next to Charles River.

2. Colinearity poses serious problems to least squares regression. Consider the Boston dataset. Here we would like to investigate the relationship between medv, $Y$ variable, median value of owner-occupied homes, in unit of thousand dollars, with the following numerical variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

tax, $X_1$, full-value property tax rate per $10,000$ dollars.

crim, $X_2$, per capita crime rate by town.

rad, $X_3$, index of accessibility to radial highways.

Some R command and result (rounded to 4 decimal places) that might be helpful are provided:

```
summary( lm(crim ~ tax + rad, data = Boston) )$r.sq [1] 0.3923
summary( lm(tax ~ crim + rad, data = Boston) )$r.sq [1] 0.8288
summary( lm(rad ~ tax + crim, data = Boston) )$r.sq [1] 0.8422
```

(a) (2 points) What is one significant problem to least squares regression when the data matrix $\mathbf{X}$ has colinearity issues?

1) It is difficult or impossible to know which independent variable has a relationship with the response variable, i.e. it is hard to make interpretations.

2) Estimates of the $\hat{\beta}$ vector becomes less accurate.

3) The computation of $\hat{\beta}$ becomes difficult due to $\mathbf{X}^T\mathbf{X}$ being non-invertible.

(b) (2 points) An intuitive way to analyze colinearity is to look at the correlation matrix of the independent (i.e. $X$) variables. What is the problem with only using the correlation matrix of the independent (i.e. $X$) variables to analyze colinearity in the context of the current question?

Correlation matrix only gives relationship between two variables, but there could be the situation that one of the variable is correlated with a linear combination of the other two variables.

(c) (4 points) Compute the VIF (variance inflation factor) for $X_1$, i.e. the variable tax.

$$\mathsf{VIF}(\hat{\beta}_1) = 1/(1 - R^2_{X_1|X_{-1}}) = 1/(1 - 0.8288) = 5.8411$$

(d) (2 points) If the VIF computed above for tax is higher than the pre-set threshold value[2], what is one thing we could do to the regression model to alleviate colinearity issues?

We can remove some of the colinear independent variables, for example, remove tax.

Alternatively, we can create a new variable in terms of the linear combination of the colinear independent variables.

[2] This value is usually 5 or 10.

3. Multiple choice (no explanations required). Select one answer for each of the following parts. Please clearly **write** the letter of your choice in the bracket at the end of each question[3].

   (a) (3 points) In the context of resampling methods with applications to regression, in the usual general setting, which of the following is not an advantage of Leave One Out Cross Validation(LOOCV) over a single training validation split?

   ( B )

   A: LOOCV has no randomness

   B: LOOCV is computationally faster

   C: LOOCV uses a larger proportion of the available data to train the model

   D: LOOCV leads to a significant reduction in the over-estimation of the error

   (b) (3 points) Which of the following is not an advantage of Least Squares regression over K Nearest Neighbors(KNN) regression?

   ( A )

   A: When we want to fit a model with relaxed assumptions

   B: When we want to fit a model that can be computed quickly

   C: When the dimension of the data is high (relative to the sample size)

   D: When we want to focus on the interpretation of the regression model and understand the relationship between the response and the predictors

   (c) (4 points) Consider a dataset with the predictors in $\mathbb{R}^2$ of size $n = 4$, with predictors matrix and response vector given below:

   $$\mathbf{X} = \begin{bmatrix} 0 & 0 \\ 5 & 0 \\ 5 & 5 \\ 0 & 5 \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} 6 \\ 7 \\ 8 \\ 9 \end{bmatrix}$$

   A new data point is located at $\mathbf{x_0} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$. When $K = 2$, and using the Euclidean distance as the distance measure, evaluate the K Nearest Neighbors(KNN) regression prediction for this data point, i.e. find $\hat{f}(\mathbf{x_0})$.

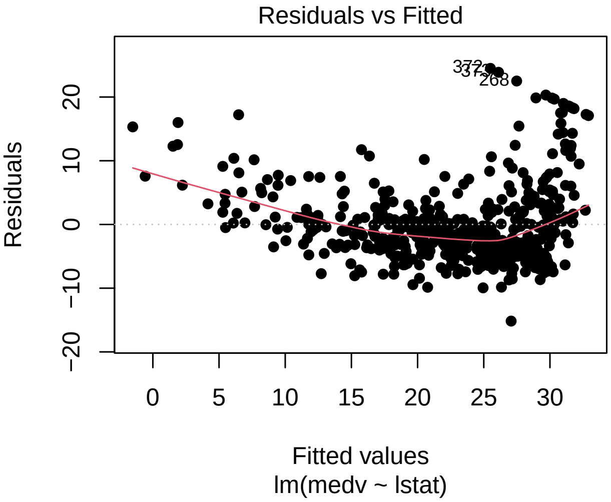
   ( B )

   A: 6 B: 6.5 C: 7 D: 7.5

[3] Please note that if you only circle the correct solution without writing the letter in the given space, 1 point per part will be deducted.

4. Multiple choice (no explanations required). Select one answer for each of the following parts. Please clearly **write** the letter of your choice in the bracket at the end of each question[4].

(a) (3 points) We consider the relationship between medv and lstat in the Boston dataset. When fitting a simple linear regression by least squares (with degree 1), the Residual vs Fitted values plot is given. Which of the following is the most likely situation based on this plot? ( C )



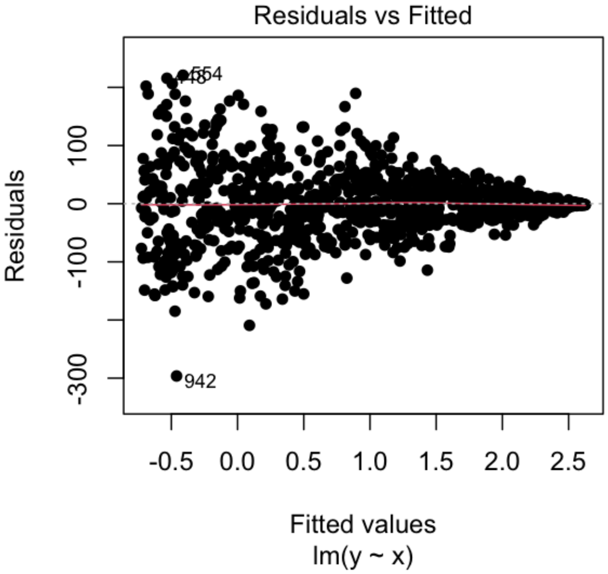
A: This plot shows the normality assumption on the error terms being violated

B: This plot shows the independence assumption on the error terms being violated

C: This plot shows the expectation of 0 assumption on the error terms being violated

D: This plot shows that all of the relevant assumptions on the error terms for least squares regression are being satisfied

(b) (3 points) We are given the Residual vs Fitted values plot for a simulated data set consisting of numerical data of size $n = 1,000$. The response is denoted as $Y$ and the independent variable is a one-dimensional variable $X$. This plot clearly shows that one of the assumptions for the error terms for least squares regression has been violated. Which of the following is the best way to remedy(correct) this problem? ( D )



[4] Please note that if you only circle the correct solution without writing the letter in the given space, 1 point per part will be deducted.

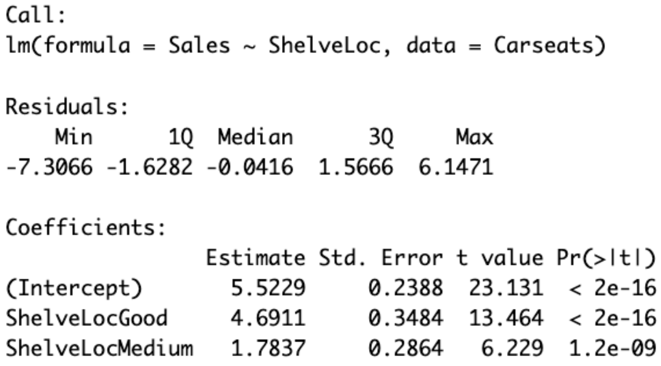
A: Keep $Y$ the same, and change $X$ to $X^2$

B: Keep $Y$ the same, and change $X$ to $\log(X)$

C: Keep $X$ the same, and change $Y$ to $Y^2$

D: Keep $X$ the same, and change $Y$ to $\log(Y)$

(c) (4 points) The Carseats dataset[5] contains sales of child car seats at $n = 400$ different stores. We would like to investigate the relationship between Sales, $Y$ variable, unit sales (in thousands) at each location, with ShelveLoc, $X$, a categorical variable consisting of 3 levels: Bad, Medium, and Good. These levels indicate the quality of the shelving location for the car seats at each site. The R output for the regression is given below, note that R automatically chose ShelveLocGood and ShelveLocMedium in the output:



```
Call:
lm(formula = Sales ~ ShelveLoc, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.3066 -1.6282 -0.0416  1.5666  6.1471 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.5229     0.2388  23.131  < 2e-16
ShelveLocGood     4.6911     0.3484  13.464  < 2e-16
ShelveLocMedium   1.7837     0.2864   6.229  1.2e-09
```

Based on the R output, which of the following statements is wrong?

( B )

A: On average, good shelve locations yield 10.214 thousand units of sale.

B: On average, medium shelve locations yield 1.7837 thousand units of sale.

C: At the significance level of $\alpha = 0.05$, there is significant evidence that sales for bad shelve locations are greater than 0 thousand units.

D: At the significance level of $\alpha = 0.05$, there is significant evidence that sales between medium shelve locations and bad shelve locations are different.

[5] ISLR2 package.