California State Polytechnic University Pomona

STA 4320 – Midterm Exam I

Prof. He Jiang

Oct 2024

Name: ____

Student ID Number: _______________________________

This exam contains 6 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. Multiple choice (no explanations required). Select one answer for each of the following parts.

   Clearly **write** the letter of your choice in the bracket at the end of each question. 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) (3 points) In simple linear regression we considered the underlying model $Y = \beta_0 + \beta_1 X + \epsilon$, where ϵ is the error term. When trying to estimate β_0 and β_1 using the Least Squares method on a sample, which of the following is true about the assumptions for the error term ϵ?

   (B)

   A: ϵ is dependent with X

   B: ϵ follows the Normal distribution

   C: ϵ can have any expectation value

   D: ϵ can be reduced to 0 with improvements in our estimate of the model parameters

   (b) (3 points) In multiple linear regression (in the general form like we have seen in class) it is very important to be able to identify the matrix form of regression, i.e. identify the terms in $Y = X\beta + \epsilon$ where ϵ is the error vector.

   In the Advertising dataset, we would like to find the relationship between sales, the response variable, and the other 3 explanatory variables, TV, radio, newspaper. The first 3 rows of the dataset is provided.

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3

   Based on the provided data[1], what should the matrix X be?

   (D)

   A: $\begin{bmatrix} 22.1 \\ 10.4 \\ 9.3 \end{bmatrix}$ B: $\begin{bmatrix} 230.1 & 37.8 & 69.2 \\ 44.5 & 39.3 & 45.1 \\ 17.2 & 45.9 & 69.3 \end{bmatrix}$ C: $\begin{bmatrix} 230.1 & 37.8 & 69.2 & 22.1 \\ 44.5 & 39.3 & 45.1 & 10.4 \\ 17.2 & 45.9 & 69.3 & 9.3 \end{bmatrix}$ D: $\begin{bmatrix} 1 & 230.1 & 37.8 & 69.2 \\ 1 & 44.5 & 39.3 & 45.1 \\ 1 & 17.2 & 45.9 & 69.3 \end{bmatrix}$

   (c) (4 points) Consider a version of simple linear regression without the intercept, i.e. with underlying model $Y = \beta_1 X + \epsilon$, where ϵ is the error term and follows appropriate assumptions. Given a sample of size n, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, what would be our Least Squares estimator, $\hat{\beta}_1$, of the true slope coefficient (in the current slope-only model)?

   (C)

   A: $\hat{\beta}_1 = (\sum_{i=1}^n y_i)/n$

   B: $\hat{\beta}_1 = (\sum_{i=1}^n y_i)/(\sum_{i=1}^n x_i)$

   C: $\hat{\beta}_1 = (\sum_{i=1}^n x_i y_i)/(\sum_{i=1}^n x_i^2)$

   D: $\hat{\beta}_1 = \Big(\sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i \sum_{i=1}^n y_i)/n\Big) / \Big(\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2/n\Big)$

[1] In practice, of course, we need much more than 3 rows of data to fit this model via least squares.

2. No-fines concrete is beneficial in areas prone to excessive rainfall because of its excellent drainage properties. The article "Pavement Thickness Design for No-Fines Concrete Parking Lots," J. of Trans. Engr., 1995: 476–484) provided data on unit weight (denoted as x), in pcf, and porosity (denoted as y; meaning having holes through which liquid or air may pass), in percentages. We want to study how porosity (y, response variable) is related to unit weight (x, independent variable). We want to do so by using simple linear regression on the data based on least squares, i.e. fit a model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

   Below is the R result of the regression:

	Estimate	Std. Error	t value	Pr(> $\|t\|$)
(Intercept)	$\hat{\beta}_0$	4.49912	26.43	1.10e-12
x	$\hat{\beta}_1$	0.04109	-22.02	1.12e-11

Residual standard error	0.938 on 13 degrees of freedom
Multiple R-squared	0.9739, Adjusted R-squared: 0.9719
F-statistic	484.8 on 1 and 13 DF, p-value: 1.125e-11

   Here, $\sum X = \sum_{i=1}^n x_i = 1640.1$, $\sum Y = \sum_{i=1}^n y_i = 299.8$, $\sum X^2 = \sum_{i=1}^n x_i^2 = 179849.73$, $\sum Y^2 = \sum_{i=1}^n y_i^2 = 6430.06$, $\sum XY = \sum_{i=1}^n x_i y_i = 32308.59$. The sample size is $n = 15$.

   (a) (4 points) Please compute the slope estimator, i.e. compute $\hat{\beta}_1$.

$$\begin{aligned}\hat{\beta}_1 &= \Big(\sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i \sum_{i=1}^n y_i)/n\Big)/\Big(\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2/n\Big)\\ &= \Big(32308.59 - (1640.1 * 299.8)/15\Big)/\Big(179849.73 - (1640.1)^2/15\Big)\\ &= -0.9047\end{aligned}$$

   (b) (4 points) Please compute the intercept estimator, i.e. compute $\hat{\beta}_0$.

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{\sum Y}{n} - \hat{\beta}_1 \frac{\sum X}{n} = \frac{299.8}{15} - (-0.9047)\frac{1640.1}{15} = 118.9066$$

   (c) (2 points) The last row of the output states that the p value corresponding to the F-statistic is $1.125e - 11$. Clearly state the null (H_0) and alternative (H_1) hypothesis in the context of the current question corresponding to this p value.

   H_0 : Given the intercept, $\beta_1 = 0$ v.s. H_1 : Given the intercept, $\beta_1 \neq 0$

   Alternatively H_0 : The model is $Y = \beta_0 + \epsilon$ v.s. H_1 : The model is $Y = \beta_0 + \beta_1 X + \epsilon$

3. Multiple choice (no explanations required). Select one answer for each of the following parts.

   Clearly **write** the letter of your choice in the bracket at the end of each question. 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.

   The following parts are based on the same situation and information as the previous question, i.e. on No-fines concrete.

   (a) (3 points) We are interested in knowing whether there is a significant positive relationship between unit weight and porosity. To investigate, we test (given the intercept) $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 > 0$. What is the p value for this hypothesis test?

   (B)

   A: 1.10e-12

   B: 0.56e-11

   C: 1.12e-11

   D: 2.24e-11

   (b) (3 points) What proportion of the variability in Y (porosity) can be explained using the regression model (i.e. variability in X (unit weight))?

   (D)

   A: 1.12e-11

   B: 0.0261

   C: 0.9380

   D: 0.9739

   (c) (4 points) The 95% confidence interval for the true slope coefficient β_1 is centered at the estimator $\hat{\beta}_1$, and is represented by $(\hat{\beta}_1 - \mathsf{ME}, \hat{\beta}_1 + \mathsf{ME})$, where ME stands for the margin of error. Based on information provided in this question, (some R code that might be helpful are provided on the first page of the exam), ME equals which of the following value[2]?

   (D)

   A: 0.0805

   B: 0.0876

   C: 0.0881

   D: 0.0888

[2] The question is designed this way so you do not have to use the computed value of $\hat{\beta}_1$ here to form the entire confidence interval.

4. The Boston dataset[3] 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 thousand dollars, with the following numerical explanatory variables to the right, i.e. investigate the underlying model:

   $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon$$

   The resulting R output is given below:

crim	β_1	X_1
nox	β_2	X_2
rm	β_3	X_3
age	β_4	X_4
ptratio	β_5	X_5
lstat	β_6	X_6

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 20.50369    4.26767   4.804 2.06e-06 ***
crim        -0.05782    0.03142  -1.840   0.0664 .  
nox         -5.89013    3.11992  -1.888   0.0596 .  
rm           4.40385    0.43574  10.107  < 2e-16 ***
age          0.03276    0.01299   2.521   0.0120 *  
ptratio     -0.93379    0.11958  -7.809 3.43e-14 ***
lstat       -0.56776    0.05412 -10.491  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.187 on 499 degrees of freedom
Multiple R-squared:  0.6857,    Adjusted R-squared:  0.6819
F-statistic: 181.4 on 6 and 499 DF,  p-value: < 2.2e-16
```

   Should we consider adding at least one of crim, the per capita crime rate by suburb, and nox, the nitrogen oxides concentration (parts per 10 million), to the model, given the intercept and other four variables? I.e. we would like to conduct a hypothesis test at the $\alpha = 0.05$ level of the following hypothesis:

   H_0: Given the intercept, rm, age, ptratio, lstat, $\beta_1 = \beta_2 = 0$

   H_1: Given the intercept, rm, age, ptratio, lstat, $\beta_1 \neq 0$ or $\beta_2 \neq 0$ or both $\neq 0$

[3] ISLR2 package.

(a) (2 points) Compute RSS, the residual sum of squares of the current model. Please round your answer to 4 decimal points.

$$\begin{aligned}\mathsf{RSE} &= \sqrt{\mathsf{RSS}/(n-p-1)}\\ 5.187 &= \sqrt{\mathsf{RSS}/(506-6-1)}\\ \mathsf{RSS} &= 13425.5795\end{aligned}$$

(b) (4 points) We then fit a submodel with only intercept, rm, age, ptratio, lstat, and the resulting residual sum of squares, $\mathsf{RSS}_0 = 13661.75$

Please compute the appropriate test statistic for the hypothesis test in this question.

$$F_0 = \frac{(\mathsf{RSS}_0 - \mathsf{RSS})/q}{\mathsf{RSS}/(n-p-1)} = \frac{(13661.75 - 13425.5795)/2}{13425.5795/(506-6-1)} = 4.3890$$

(c) (2 points) Compute the p value based on the above test statistic. Some R code that might be helpful are provided on the first page of the exam.

This is the F Distribution with q, $n-p-1$, i.e. with 2, 499 degrees of freedom.

$$\text{p value} = 1 - \text{pf}(4.3890,\ 2,\ 499) = 1 - 0.9871 = 0.0129$$

(d) (2 points) State a conclusion in the context of the current question.

We rejected H_0 at the $\alpha = 0.05$ level. Therefore with intercept, rm, age, ptratio, lstat in the model, we have strong evidence that at least one of crim, nox should be added to the model.