STA 206 001 FQ 2022 Final

Gregory DePaul

TOTAL POINTS

85 / 100

QUESTION 1

1 1(a) **5 / 5**

✓ **- 0 pts** Correct

QUESTION 2

2 1(b) **5 / 5**

✓ **- 0 pts** Correct

QUESTION 3

3 1(c) **5 / 5**

✓ **- 0 pts** Correct

QUESTION 4

4 1(d) **3 / 5**

✓ **- 2 pts** wrong answer but attempt to reasoning

QUESTION 5

5 2(a) **5 / 5**

✓ **- 0 pts** Correct

QUESTION 6

6 2(b) **5 / 5**

✓ **- 0 pts** Correct

QUESTION 7

7 2(c) **5 / 5**

✓ **- 0 pts** Correct

QUESTION 8

8 2(d) **2 / 5**

✓ **- 3 pts** Major mistake, wrong approach but attempt to solve the problem

QUESTION 9

9 2(e) **3 / 5**

✓ **- 2 pts** Incomplete calculation

QUESTION 10

10 3(a) **5 / 10**

✓ **- 5 pts** Major mistake, not obtain/comment correct VIF

QUESTION 11

11 3(b) **10 / 10**

✓ **- 0 pts** Correct

QUESTION 12

12 3(c) **2 / 5**

✓ **- 3 pts** Major mistake, wrong formula

QUESTION 13

13 4(a) **5 / 5**

✓ **- 0 pts** Correct

QUESTION 14

14 4(b) **5 / 5**

✓ **- 0 pts** Correct

QUESTION 15

15 4(c) **5 / 5**

✓ **- 0 pts** Correct

QUESTION 16

16 4(d) **10 / 10**

✓ **- 0 pts** Correct

QUESTION 17

17 4(e) **5 / 5**

✓ **- 0 pts** Correct

Statistics 206

Fall 2022

Final Exam: Nov. 30, 10:00am - 11:50am, TLC 3214

Print name: Greg DePaul

Print ID (all digits): 917835494

Sign name: Greg DePaul

Instructions: This is an open notes exam. No mobile device of any kind is allowed. A handheld calculator is allowed. The duration of the exam is 110 minutes which include time for distributing and collecting the exam.

The total score is **100**. You must show your work for full credit. Partial credit can only be given if your thoughts can be followed. Make sure your name is written on the first page and all the additional pages attached by yourself (if any).

You must not show this exam to anyone outside of this class or post it anywhere.

Score:

1:

2:

3:

4:

Total:

1. (20 points) Answer **true** or **false** of the following statements with regard to linear regression models **in the box** and briefly explain your answer.

   (a) The adjusted R^2 never decreases when additional X variables are added into the model.

   False

   Explanation:

   decrease in SSE may be more than offset by the loss of degrees of freedom in SSE

   (b) The summation of all elements of the hat matrix equals the sample size.

   true

   Explanation:

   $1_n^T H 1_n = 1_n^T 1_n = n$

   (c) If an X variable is uncorrelated with the response variable, then its least-squares estimated regression coefficient must be zero.

   false

   Explanation:

   only true for simple model. In multiple regression, that coefficient might be changed by multicollinearity

   (d) The residuals and the fitted values are uncorrelated whether or not the model is correct as long as the responses are uncorrelated and have equal variance.

   false

   Explanation:

   model correctness would imply that $E[e] = 0$, and therefore we would have correlation.

2. (25 points) Consider a data set with 50 cases and three variables: Y, X_1, X_2. It is given that the sample correlation between X_1 and X_2 is **0.5**, the sample correlation between Y and X_1 is **0.3**, and the sample correlation between Y and X_2 is **0.2**. Moreover, the sample mean and sample standard deviation of Y is **0.1** and **1**, respectively.

$n = 50$, $r_{12} = 0.5$, $r_{Y1} = 0.3$, $r_{Y2} = 0.2$, $\bar{Y} = 0.1$, $s\{Y\} = 1$

Consider regressing Y onto X_1 and X_2. **Calculate (a) – (e) under the standardized regression model.**

Hints: (i) Recall that in the standardized regression model, the X variables are transformed by the correlation transformation, whereas the response variable is not transformed.
(ii) Recall that for a 2×2 matrix $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, you have confirmed in homework that its inverse $M^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$, provided that $ad - bc \neq 0$

(a) the fitted regression intercept

$$\hat{\beta}_0^* = \bar{Y} = \boxed{0.1}$$

(b) the fitted regression slopes of the two X variables

$$r_{XX} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \Rightarrow r_{XX}^{-1} = \frac{1}{0.75}\begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix} = \frac{4}{3}\begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix}$$

$$r_{XY} = \begin{pmatrix} 0.3 \\ 0.2 \end{pmatrix}$$

$$\begin{pmatrix} \hat{\beta}_1^* \\ \hat{\beta}_2^* \end{pmatrix} = \sqrt{n-1}\, s_Y\, r_{XX}^{-1} r_{XY} = 7 \cdot 1 \cdot \frac{4}{3}\begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix}\begin{pmatrix} 0.3 \\ 0.2 \end{pmatrix}$$

$$= \frac{28}{3}\begin{pmatrix} 0.2 \\ 0.05 \end{pmatrix} = 4\begin{pmatrix} 1.8667 \\ 0.4667 \end{pmatrix}$$

$$S_Y = \frac{1}{n-1}\sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

(c) the total sum of squares

under the standardized model

$$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$= (n-1) S_Y$$

$$= 49 \cdot 1 = \boxed{49}$$

(d) the regression sum of squares

$$SSR = \hat{\beta}' X' X \hat{\beta}$$

$$= \begin{pmatrix} 0.1 & 1.8667 & 0.4667 \end{pmatrix} \begin{pmatrix} \frac{1}{50} & 0 & 0 \\ 0 & \frac{4}{3} & -\frac{2}{3} \\ 0 & -\frac{2}{3} & \frac{4}{3} \end{pmatrix} \begin{pmatrix} 0.1 \\ 1.8667 \\ 0.4667 \end{pmatrix}$$

$$= 3.775$$

not sure what this means.

(e) the standard errors of the fitted regression slopes

$$\sigma^2(\hat{\beta}^*) = \sigma^2 \begin{pmatrix} \frac{1}{50} & 0 & 0 \\ 0 & \frac{4}{3} & -\frac{2}{3} \\ 0 & -\frac{2}{3} & \frac{4}{3} \end{pmatrix}$$

$$0.3 = \sum_{i=1}^{n} e_{i1}^2 = \sum_{i=1}^{n} (X_{i1} - \hat{X}_{i1}(X_2, X_3))^2 = \sum_{i=1}^{n} (X_{i1} - \bar{X} + \bar{X} - \hat{X}_{i1}$$

$$= \sum_{i=1}^{n} (X_{i1} - \bar{X})^2 + 2(X_{i1} - \bar{X})(\bar{X} - \hat{X}_{i1}(X_2, X_3)) + (\bar{X} - \hat{X}_{i1}(X_2, X_3))^2$$

3. (25 points) Consider a data set with n cases and four variables: Y, X_1, X_2, X_3. Let $\hat{\beta}_1$ denote the least-squares (LS) fitted regression coefficient of X_1 when regressing Y onto X_1, X_2, X_3.

Let X_{i1} denote the ith observation of X_1, $\bar{X}_1 := \frac{1}{n}\sum_{i=1}^{n} X_{i1}$; Let $e_{i1} = e_i(X_1|X_2, X_3)$ denote the ith residual by regressing X_1 onto X_2, X_3; and Let Y_i denote the ith observation of Y. The following summary statistics are given:

$$\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)^2 = 100, \quad \sum_{i=1}^{n} e_{i1}^2 = 0.3, \quad \sum_{i=1}^{n} Y_i \cdot e_{i1} = 1.2$$

$(X-\bar{X})^T(X-\bar{X})$ $\quad e^Te$ $\quad Y^Te$

(a) Calculate the variance inflation factor for $\hat{\beta}_1$. Comment on the degree of multicollinearity among X_1, X_2, X_3.

$$VIF_1 = \frac{1}{1 - R^2_{1|23}} = \frac{1}{1 - 0.003} = 1.00301 \approx 1$$

$$R^2_{1|23} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(e_i - \bar{e})}{\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)^2} = \frac{0.3}{100} = 0.003$$

this is little multicollinearity between X_1 and X_2, X_3

(b) Denote the residuals of regressing Y onto X_2, X_3 by $e(Y|X_2, X_3)$ and denote the residuals of regressing X_1 onto X_2, X_3 by $e(X_1|X_2, X_3)$. In class, you have learned that $\hat{\beta}_1$ *equals the LS fitted regression slope when regressing* $e(Y|X_2, X_3)$ *onto* $e(X_1|X_2, X_3)$. Using this fact, show that $\hat{\beta}_1$ equals the LS fitted regression slope when regressing Y onto $e(X_1|X_2, X_3)$.

We know $\hat{\beta}_1$ equals the LS fitted coefficient when

$$e(Y \mid X_{(1)}) \sim e(X_1 \mid X_{(1)})$$

However, since $VIF_1 = 1$, then we see that the coefficient that would result from regressing Y onto X_1 given X_2 and X_3 is equivalent to Y regressing onto X_1.

(c) Calculate $\hat{\beta}_1$. since $E[e] = 0$

$$\hat{\beta}_1 = \frac{\bar{Y}\,\cancel{E[e]^{0}} - \frac{1}{n}\sum_{i=1}^{n} e_i Y_i}{\cancel{E[e]^{2}{}^{0}} - \frac{1}{n}\sum_{i=1}^{n} e_i^2} = \frac{-\frac{1}{n}\sum_{i=1}^{n} e_i Y_i}{-\frac{1}{n}\sum e_i^2} = \frac{1.2}{0.3} = 4$$

4. (30 points) A city tax officer was interested in predicting residential home sales price by finished square footage and the quality of construction (high, medium or low). Data was collected on 522 home sales made in last year. A snapshot of the data is shown below.

case	sales-price (1000$)	square-footage (1000SQ)	quality
1	360.0	3.032	medium
2	340.0	2.058	medium
...	...	...	...
69	585.0	2.558	high
70	549.9	4.000	high
...	...	...	...
521	124.0	1.480	low
522	95.5	1.184	low

A model by regressing sales price onto square footage, construction quality and the interaction between square footage and construction quality (Model 1) is fitted to the data. Relevant R outputs are given below.

```
Call:
lm(formula = sales ~ Sq + quality + Sq:quality, data = data)


Coefficients:
                   Estimate  Std. Error t value Pr(>|t|)
(Intercept)         337.74      38.40    8.796  < 2e-16 ***
Sq                   61.51      11.25    5.469 7.06e-08 ***
qualitylow         -289.32      47.31   -6.115 1.91e-09 ***
qualitymedium      -333.83      41.67   -8.011 7.62e-15 ***
Sq:qualitylow        12.82      19.54    0.656    0.512
Sq:qualitymedium     54.75      13.14    4.167 3.62e-05 ***
---

Residual standard error: 62.57 on 516 degrees of freedom
Multiple R-squared: 0.7962,     Adjusted R-squared: 0.7942
F-statistic: 403.2 on 5 and 516 DF,  p-value: < 2.2e-16


Analysis of Variance Table

Response: sales
             Df Sum Sq Mean Sq   F value    Pr(>F)
Sq            1 6655486 6655486 1700.1635 < 2.2e-16 ***
quality       2 1157409  578705  147.8318 < 2.2e-16 ***
Sq:quality    2   78075   39037    9.9722 5.633e-05 ***
Residuals   516 2019942    3915
```

(a) Write down the fitted regression function corresponding to low construction quality (`'qualitylow'`) under **Model 1**. Note that, you should clearly define notations in the regression function (e.g., "x stands for...").

$$\text{Sales}_i = 337.74 - 289.32 + (12.82 + 61.51)\,\text{Sq}$$

$$= \boxed{48.42 + 74.33\,\text{Sq}}$$

(b) Is the interaction between square footage and construction quality significant? Explain your answer.

We see that the interaction between square footage and medium quality has a p-value of 3.62×10^{-5}. Therefore, we see that this suggests an definite nonzero coefficient.

(c) Calculate BIC for **Model 1**.

$p = 6$, $n = 522$

$$BIC_p = 522 \cdot \log \frac{SSE}{n} + \log(n)\, p$$

$$= 522 \cdot \log \frac{2019942}{522} + \log(522) \cdot 6$$

$$= \boxed{4349.74}$$

(d) Calculate BIC for the first-order model without interaction (referred to as **Model 2**). Which one, **Model 1** or **Model 2**, is preferred by BIC? Explain your answer.

$$SSE_p = 2019942 + 78075 = 2098017$$

$$p = 4$$

$$BIC = 522 \log\left(\frac{2098017}{522}\right) + \log(522) \cdot 4$$

$$= \boxed{4357.02}$$

Very similar, although BIC for model 2 is slightly smaller. So we can select model 2.

(e) Use **Model 1** as the full model, calculate C_p for **Model 2**. What is suggested by the C_p statistic? Explain your answer.

$$MSE_{full} = (62.57)^2 = 3915$$

$$p = 4$$

$$C_p = \frac{SSE_p}{MSE_{full}} - (n - 2p) = \frac{2098017}{3915} - (522 - 2 \cdot 4)$$

$$= \boxed{21.8913}$$

A desirable C_p would be one such that $C_p \approx p = 4$. Since $C_p \gg 4$, it's clear Model 2 isn't sufficient.

END OF EXAM.