STA 206 001 FQ 2022 Final

Sirapat Watakajaturaphon

TOTAL POINTS

95 / 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) **5 / 5**

✓ **- 0 pts** Correct

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) **4 / 5**

✓ **- 1 pts** numerical error

QUESTION 10

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

✓ **- 0 pts** Correct

QUESTION 11

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

✓ **- 0 pts** Correct

QUESTION 12

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

✓ **- 0 pts** Correct

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) **4 / 5**

✓ **- 1 pts** calculation error

Statistics 206

Fall 2022

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

Print name: Sirapat Warakajaturaphon

Print ID (all digits): 920226951

Sign name: Sirapat W.

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.

   $\boxed{\text{False}}$

   Explanation:

   $$R_a^2 = 1 - \frac{n-1}{n-p} \cdot \frac{SSE}{SSTO}$$

   If p increases, R_a^2 can decrease.

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

   $\boxed{\text{True}}$

   Explanation:

   $$H 1_n = 1_n \Rightarrow \begin{pmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{n1} & \cdots & h_{nn} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

   $$\Rightarrow \begin{matrix} h_{11} + \cdots + h_{1n} = 1 \\ h_{n1} + \cdots + h_{nn} = 1 \end{matrix} \quad \Rightarrow \text{sum of all elements equals to } \underbrace{1 + \cdots + 1}_{n \text{ times}} = n. \quad \#$$

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

   $\boxed{\text{False}}$

   Explanation: WLOG: assume X_1 is uncorrelated with Y.

   suppose $\hat{\beta}_1 = 0$. So, $\hat{\beta}_1^*$ is also equal to 0.

   consider $\hat{\beta}_1^* = \sqrt{n-1}\, S_Y\, r_{XX}^{-1} [1, \;] \, r_{XY} = \sqrt{n-1}\, S_Y [C_1, \ldots, C_{p-1}] \begin{bmatrix} cor(X_1, Y) = 0 \\ \vdots \\ cor(X_{p-1}, Y) \end{bmatrix}$

   This means $C_2\, cor(X_2, Y) + \cdots + C_{p-1}\, cor(X_{p-1}, Y) = 0$

   which contradiction because all of those terms are positive. #

   (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.

   $\boxed{\text{True}}$

   Explanation:

   $$\left.\begin{aligned} cov(e, \hat{Y}) &= cov((I_n - H)Y, HY) \\ &= (I_n - H)\, \sigma^2\{Y\}\, H' \\ &= (I_n - H)(\sigma^2 I_n) H \\ &= \sigma^2 (I_n - H) H \\ &= \sigma^2 (H - H) \\ &= 0_{n \times n} \end{aligned}\right\}$$

   no correct model assumption needed. (what requires is $\sigma^2\{Y\} = \sigma^2 I_n$).

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.

Consider regressing Y onto X_1 and X_2. **Calculate (a) – (e) under the standardized regression model.** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

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* $ad - bc \neq 0$ $r_{XX} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \to r_{XX}^{-1} = \begin{bmatrix} 1.33 & -0.67 \\ -0.67 & 1.33 \end{bmatrix}$

(a) the fitted regression intercept

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

$\begin{bmatrix} \hat{\beta}_1^* \\ \hat{\beta}_2^* \end{bmatrix} = \sqrt{n-1}\, S_Y\, r_{XX}^{-1} r_{XY} = \sqrt{49}(1) \begin{bmatrix} 1.33 & -0.67 \\ -0.67 & 1.33 \end{bmatrix} \begin{bmatrix} 0.3 \\ 0.2 \end{bmatrix} = \begin{bmatrix} 1.855 \\ 0.455 \end{bmatrix}$

Thus, $\hat{Y}_i = 0.1 + 1.855 X_{i1}^* + 0.455 X_{i2}^*$.

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

$\hat{\beta}_1^* = 1.855$

$\hat{\beta}_2^* = 0.455$

$$1^2 = \frac{\Sigma(Y_i - \bar{Y})^2}{n-1} \Rightarrow \Sigma(Y_i - \bar{Y})^2 = 49$$

(c) the total sum of squares

$SSTO = \Sigma(Y_i - \bar{Y})^2 = 49$ because sample sd of $Y = 1$.

(d) the regression sum of squares

$$SSR = SSTO(R^2)$$
$$= 49R^2$$

thus, $SSE = SSTO - SSR$
$$= 49 - 49R^2$$
$$= 49(1-R^2)$$

$\Rightarrow$ so, $MSE = \frac{SSE}{n-p} = \frac{SSE}{50-3} = \frac{SSE}{47}$

(e) the standard errors of the fitted regression slopes

$$s\{\hat{\beta}_1^*\} = \sqrt{MSE\, r_{XX}^{-1}[1,1]} = \sqrt{\frac{49(1-R^2)}{47} \times 1.33}$$

$$s\{\hat{\beta}_2^*\} = \sqrt{MSE\, r_{XX}^{-1}[2,2]} = \sqrt{\frac{49(1-R^2)}{47} \times 1.33}$$

$$\begin{bmatrix} var(\hat{\beta}_1^*) & cov(\hat{\beta}_1^*, \hat{\beta}_2^*) \\ cov(\hat{\beta}_2^*, \hat{\beta}_1^*) & var(\hat{\beta}_2^*) \end{bmatrix} = \sigma^2 r_{XX}^{-1}$$

$Y = \beta_0 + \beta_1 X_1$

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:

$$\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2 = 100, \quad \underbrace{\sum_{i=1}^n e_{i1}^2}_{\langle e_1, e_1 \rangle} = 0.3, \quad \underbrace{\sum_{i=1}^n Y_i \cdot e_{i1}}_{\langle Y, e_1 \rangle} = 1.2$$

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

$$R_1^2 = 1 - \frac{\sum e_{i1}^2}{\sum (X_{i1} - \bar{X})^2} = 1 - \frac{0.3}{100} = 0.997$$

Thus, $VIF_1 = \frac{1}{1 - R_1^2} = \frac{1}{0.003} = 333.33 \Rightarrow$ high multicollinearity.

(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 $\hat{\beta}_1 = \frac{\sum (e_{i1} - \overset{0}{\cancel{\bar{e}_1}})(e_{iY} - \overset{0}{\cancel{\bar{e}_Y}})}{\sum (e_{i1} - \bar{e}_1)^2} = \frac{\sum e_{i1} e_{iY}}{\sum (e_{i1} - \bar{e}_1)^2}$

see $\sum e_{i1} e_{iY} = \langle e_1, e_Y \rangle = \langle e_1, Y - \hat{Y}(X_2, X_3) \rangle$
$= \langle e_1, Y \rangle$.

thus, $\hat{\beta}_1 = \frac{\sum e_{i1} e_{iY}}{\sum (e_{i1} - \bar{e}_1)^2} = \frac{\sum e_{i1} Y_i}{\sum (e_{i1} - \bar{e}_1)^2} = \frac{\sum (e_{i1} - \bar{e}_1)(Y_i - \bar{Y})}{\sum (e_{i1} - \bar{e}_1)^2}$. #

the LS slope formula when regressing Y onto e_1.

(c) Calculate $\hat{\beta}_1$.

By (b), $\hat{\beta}_1 = \frac{\sum (e_{i1} - \overset{0}{\cancel{\bar{e}_1}})(Y_i - \bar{Y})}{\sum (e_{i1} - \bar{e}_1)^2} = \frac{\sum e_{i1}(Y_i - \bar{Y})}{\sum e_{i1}^2} = \frac{\sum e_{i1} Y_i}{\sum e_{i1}^2}$

$= \frac{1.2}{0.3} = 4.$ #

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
```

$n - p$

$= 522 - 6$

$= 516$

(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...").

Let $qualitylow = \begin{cases} 1 & low \\ 0 & otw \end{cases}$; $qualitymedium = \begin{cases} 1 & medium \\ 0 & otw \end{cases}$.

$$\hat{sales} = 337.74 + 61.51\,Sq - 289.32 + 12.82\,Sq$$

$$= 48.42 + 74.33\,Sq$$

is the fitted func. for quality low. #

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

Based on the ANOVA table,

	Pr(> F)
sq : quality	5.633×10^{-5} - pvalue very small

Thus, the interaction is significant. #

It's the test between model with sq and quality and the model with sq, quality, and the interaction.

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

$$BIC = n\log\left(\frac{SSE_F}{n}\right) + (\log n)\,p$$

$$= 522\log\left(\frac{2019942}{522}\right) + [\log(522)]\,6 \quad \text{(see ANOVA table)}$$

$$= 4349.75.$$

#

(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(model2) = SSE(model w/o interation) = 78075 + 2019942 (ANOVA)

= 2098017

BIC for model 2 = $522 \log\left(\frac{2098017}{522}\right) + (\log(522)) \times 4$

= 4357.02.

since BIC for model 1 is smaller, model 1 is preferred. #

(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.

$C_p = \frac{SSE_p}{\hat{\sigma}^2} - (n - 2p)$

$= \frac{2098017}{3915} - (522 - 2 \times 4)$

$= 535.89 - 522 + 8$

$= 21.89$

SSE_p = SSE(model2) see (d)

$\hat{\sigma}^2$ = MSE of full model

$= 62.57^2$

$= 3915$.

since $C_p >> p = 4$, Model 2 has a substantial model bias. #

END OF EXAM.