Practice Exam 3

STA 3024 Spring 2023

Class #: 16898 (Zheng)

Name: _____

UFID: _____

Instructions:

1. This examination contains 9 pages, including this page.
2. You have **50 minutes** to complete the exam.
3. The total score is 105. The extra 5 points serve as a buffer, so the highest score you can get is 100.
4. Write your answers clearly and legibly on the exam. Answers without sufficient work shown will not receive full credit.
5. You may use a scientific calculator. Do not share a calculator with anyone.
6. This is a closed-book exam. You may not use any resources including lecture notes, books, or other students.
7. Please sign the below Honor Code statement.

In recognition of the UF Student Honor Code, I certify that I will neither give nor receive unauthorized aid on this examination. Signature: _____

1. (5 points) To use X to predict Y, we collect some observations and fit a simple linear regression model $Y = \alpha + \beta X + \varepsilon$, $\varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Which one of the following statements regarding SLR is wrong? Circle your choice and write the letter in the blank below.

   A. $b = r\frac{S_Y}{S_X}$ is the estimate of β using the Least Squares method.

   B. $a = \bar{Y} - r\bar{X}$ is the estimate of α using the Least Squares method.

   C. $\sqrt{\text{MSE}}$ is the estimate of σ using the Least Squares method.

   D. $R^2 = r^2$.

   E. $R^2 = \text{SSR}/\text{SST}$.

   1. _____

2. (30 points) Are the following statements true or false? You do not need to give reasons.

   (a) ___ In SLR, we do AVONA test first. If the p-value is small, we proceed to do the t-test for the slope.

   (b) ___ The correlation coefficient r measures strength and direction of the linear association between x and y.

   (c) ___ In SLR, the slope estimate must be zero if the correlation between x and y is zero.

   (d) ___ PI for response is always wider than CI for response at the same value of x.

   (e) ___ An observation with large leverage helps us to make a better prediction.

   (f) ___ When a funnel shape is displayed in a residual plot (residuals versus x), we should consider adding a quadratic term in the model to get a better fit.

   (g) ___ In multiple regression, the t-test for an individual predictor tells us if the predictor provides significant information about the response WITHOUT taking into account other predictor variables.

   (h) ___ Oversaturated models do a perfect job of interpolation.

   (i) ___ As long as R^2 goes up after a new predictor is added to the model, we keep that predictor in our model.

   (j) ___ In model selection, the "best subsets regression" method usually requires fitting more models than the "Backward elimination" method.

3. Consider the "egyptcttn" dataset from Hamza, A.A. and Z.N. Sokkar (1981). "Brightness of Egyptian Cotton Fibers." Textile Research Journal, Vol. 51, pp. 587-590. Following are the description of variables: Variety: 5 different cotton varieties; LnGrade: log-transformed Grade score; Luminance: Luminance scores (response variable).

   The response variable is Luminance. LnGrade is a quantitative predictor variable. And five cotton varieties (Giza69, Giza67, Giza70, Giza68, Menoufi) are coded into four dummy variables with being certain variety as 1. Therefore, we have

$$\begin{aligned} Y &= \text{Luminance} \\ X_1 &= \text{LnGrade} \\ X_2 &= \text{Giza67} \\ X_3 &= \text{Giza70} \\ X_4 &= \text{Giza68} \\ X_5 &= \text{Menoufi} \end{aligned}$$

   For each cotton variety, there are 4 observations.

   Below are the three models we focus on:

   Model 1:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

   Model 2:

$$Y = \alpha' + \beta_1' X_1 + \beta_2' X_2 + \beta_3' X_3 + \beta_4' X_4 + \beta_5' X_5 + \beta_6' X_1 X_2 + \beta_7' X_1 X_3 + \beta_8' X_1 X_4 + \beta_9' X_1 X_5 + \varepsilon$$

   Model 3:

$$Y = \alpha'' + \beta_1'' X_1 + \beta_2'' X_2 + \beta_3'' X_3 + \beta_4'' X_4 + \beta_5'' X_5 + \beta_6'' X_1 X_2 + \beta_7'' X_1 X_4 + \varepsilon$$

   The computer output for coefficient estimates and related quantities are shown in Tables 1, 2, and 3.

Table 1: Model 1 Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
Constant	83.112	0.908	91.487	0.000
X1	??	0.270	7.402	0.000
X2	-0.228	0.293	-0.778	0.450
X3	-1.718	0.293	-5.872	0.000
X4	-1.670	0.293	-5.709	0.000
X5	1.542	0.293	5.273	0.000

Table 2: Model 2 Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
Constant	78.803	1.398	56.386	0.000
X1	3.314	0.424	7.810	0.000
X2	7.223	1.976	3.655	0.004
X3	2.052	1.976	1.038	0.324
X4	5.115	1.976	2.588	0.027
X5	5.080	1.976	2.570	0.028
X1*X2	-2.274	0.600	-3.790	0.004
X1*X3	-1.151	0.600	-1.918	0.084
X1*X4	-2.071	0.600	-3.452	0.006
X1*X5	-1.080	0.600	-1.800	0.102

Table 3: Model 3 Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
Constant	81.239	0.900	90.243	0.000
X1	2.570	0.270	9.506	0.000
X2	4.787	1.786	2.680	0.020
X3	-1.718	0.227	-7.570	0.000
X4	2.679	1.786	1.500	0.159
X5	1.542	0.227	6.799	0.000
X1*X2	-1.531	0.541	-2.831	0.015
X1*X4	-1.327	0.541	-2.455	0.030

*For part (a)-(c), you do NOT need to use the computer output, i.e., the three tables.

(a) (5 points) In **Model 2**, what are the slopes for the five cotton varieties?

- Giza69:
- Giza67:
- Giza70:
- Giza68:
- Menoufi:

(b) (4 points) Briefly interpret β_5, the coefficient of Menoufi in **Model 1**.

(c) (4 points) Write down the null and alternative hypotheses for testing whether the slope for Giza67 is significantly different than Giza68 in **Model 2**.

(d) (10 points) Using the computer output, predict the luminance score for the baseline group with LnGrade=3.5 for **Model 1** and **Model 2** respectively.

(e) (5 points) Interpret the number -1.080 in Table 2.

(f) (6 points) After looking at the **Model 2** output in table 2, we decide to fit a new model, **Model 3**. From **Model 2** to **Model 3**, we get rid of X_1X_3 and X_1X_5 but still keep X_3 in the model. Give the justification for such operation.

(g) (5 points) For **Model 3**, the coefficient of determination R^2 is 96.99% and R^2_{adj} is 95.23%. Point out which one has a clear interpretation and interpret it.

(h) (10 points) Below is an incomplete ANOVA table for **Model 3**. Complete it and write down the numbers in the blanks below the table. You do not need to show your calculation process. [Hint: You may find the quantities given in the previous part useful]

Source	DF	SS	MS	F
Regression	(i)	(iv)	(vi)	(viii)
Error	(ii)	(v)	(vii)	
Total	(iii)	41.01		

(i)_____ (ii)_____ (iii)_____ (iv)_____ (v)_____
(vi)_____ (vii)_____ (viii)_____

4. Consider the "cotton_spindle" dataset from T. Leung (2003). The dataset aims to capture the relationship between yarn count and output for New England Cotton spindles of types mule spinning and ring spinning. However, in this exam, we investigate if the yarn count (X_1) and output (X_2) can help us guess which spinning type is used in the production process. Suppose $Y = 1$ refers to ring spinning and $Y = 0$ refers to mule spinning.

   Suppose we want to use logistic regression to predict Y. The following is the model equation:

$$P(Y = 1) = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2}}{1 + e^{\alpha+\beta_1 X_1+\beta_2 X_2}}.$$

   (a) (4 points) Using the given notation above, write down the (theoretical) probability of using mule spinning ($Y = 0$) given $X_1 = 20$ and $X_2 = 121.9$.

   (b) (3 points) Table 4 shows part of the computer output for the logistic model. Based on the output, give the fitted equation.

Table 4: Logistic regression model coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
Constant	-5.191	1.920	-2.703	0.007
yarncount	0.024	0.018	1.279	0.201
output	0.041	0.014	2.955	0.003

   (c) (4 points) Interpret the number 0.201 in Table 4.

(d) (5 points) Predict the probability of using mule spinning ($Y = 0$) when $X_1 = 20$ and $X_2 = 121.9$.

(e) (5 points) When the yarn count is 20, how much output do we need to make a prediction $\hat{p} = 0.5$?

5. (5 points) In SLR, we learned the quantitative relationship SST=SSE+SSR, where the three quantities are given by

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

*You only need to choose one of the two following problems to answer. Explicitly point out which question you are answering.

(a) Prove the quantitative relationship [Hint: You can use the fact that $\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) = 0$].

(b) Draw and use a plot to illustrate where these sums of squares come from and what variabilities they stand for, respectively.