# Exam 3

STA 3024 Spring 2023

Class #: 16898 (Zheng)

Name: _____

UFID: _____

**Instructions:**

1. This examination contains 8 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.<br><br>Signature: _____ |

1. (5 points) To use $X$ to predict $Y$, we collect $n = 100$ observations and fit a simple linear regression model $Y = \alpha + \beta X + \varepsilon$, $\varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. We have computed the summary statistics: $\bar{X} = 10$, $\bar{Y} = 20$, $S_X = 1$, $S_Y = 2$, $r = -0.8$. Which one of the following is the LSR equation? Circle your choice and write the letter in the blank below.

   A. $\hat{Y} = 24 - 0.4X$
   B. $\hat{Y} = 36 - 1.6X$
   C. $\hat{Y} = 40 - 2X$
   D. $\hat{Y} = -0.8X$
   E. $\hat{Y} = -1.6X$
   F. None of the above

   1. _____

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

   (a) ___ If we use the Least Squares Regression method to fit an SLR model, the sum of the residuals must be zero, i.e. $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$.

   (b) ___ Having an outlier in a regression model is not always bad. But if it is an influential outlier, we might need to consider getting rid of it.

   (c) ___ In an SLR model, the coefficient of determination $R^2$ is 0.39. It tells us that the linear association between the response and the predictor is strong.

   (d) ___ When computing residuals, we need to be cautious because we might commit extrapolation for some observations.

   (e) ___ Both PI and CI for response are the narrowest when $x = \bar{x}$.

   (f) ___ Normal probability plot (NPP) is a graphical technique for assessing whether the random error variance is constant.

   (g) ___ In multiple regression, the t-test for an individual predictor tells us if the predictor provides significant information about the response after taking into account all other predictor variables.

   (h) ___ $R^2_{\text{adj}}$ can decrease if the newly-included predictor variable is bad, but $R^2$ might still increase.

   (i) ___ Multicollinearity is conducive to our prediction because the correlation between different predictors can give us more information about the response.

   (j) ___ After fitting a quadratic regression model, we do NOT interpret the estimate of the coefficient of the linear term because that can lead to extrapolation.

3. A study was conducted to measure the effect of the number of layers of panels in cloth fabric on the velocity needed for half of a ballistic discharge to penetrate the fabric, separately for 3 types of bullets (Rounded, sharp, fsp). The response variable in the study was V50, which is the velocity at which approximately half the projectiles within a specified narrow striking velocity range penetrate the panel, in meters/second. In this exam, we use V50Csq=$(V50/100)^2$ as the response variable, i.e., $Y$ =V50Csq. Some of the predictor variables are

$X_1$ = the number of layers (quantitative)
$X_2$ = sharp (dummy variable coded as 1 = sharp, 0 = otherwise)
$X_3$ = fsp (dummy variable coded as 1 = fsp, 0 = otherwise)

In total there are $n = 25$ observations.

We first fit a full model without interaction, called **Model 1**, as follows:

$$Y = \alpha' + \beta_1' X_1 + \beta_2' X_2 + \beta_3' X_3 + \varepsilon.$$

Part of the output is shown in Table 1.

We then fit a full model with interaction, called **Model 2**, as follows:

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

Part of the output is shown in Table 2.

Table 1: Model 1 Coefficients

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Constant | 1.588 | 0.724 | 2.194 | 0.040 |
| layers | 0.951 | 0.023 | ?? | ?? |
| sharp | 3.948 | 0.744 | 5.310 | 0.000 |
| fsp | 3.368 | 0.723 | 4.659 | 0.000 |

Table 2: Model 2 Coefficients

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Constant | 3.643 | 0.836 | 4.357 | 0.000 |
| layers | 0.855 | 0.034 | 25.349 | 0.000 |
| sharp | 0.769 | 1.182 | 0.650 | 0.523 |
| fsp | 0.499 | 1.136 | 0.439 | 0.666 |
| layers∗sharp | 0.150 | 0.048 | 3.138 | 0.005 |
| layers∗fsp | 0.137 | 0.047 | 2.933 | 0.009 |

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

(a) (4 points) Write down the baseline model for **Model 1** and **Model 2** respectively.

(b) (8 points) Briefly interpret the following parameters:

- $\beta_2'$, the coefficient of sharp in **Model 1**

- $\beta_5$, the coefficient of the interaction of layers and fsp in **Model 2**

(c) (5 points) Write down the null and alternative hypotheses for testing whether the slope for sharp is significantly different than fsp in **Model 2**.

(d) (8 points) Using the computer output, predict V50Csq when type=sharp and the number of layers=20 for **Model 1** and **Model 2** respectively.

(e) (5 points) Interpret the number 0.951 in Table 1.

(f) (5 points) In Table 2, the p-values for sharp and fsp are pretty large. In light of this observation, should we eliminate them and thus fit a simpler model? Explain.

(g) (4 points) In Table 1, the t-statistic value and the p-value for layers are missing. Please find them (Round the p-value to three decimals places).

(h) (3 points) If we conduct the ANOVA test on **Model 2**, what will the degrees of freedom for Regression, Error, and Total be, repectively?

(i) (5 points) In **Model 2**, there are interaction terms $X_1X_2$ and $X_1X_3$. Should we include the interaction $X_2X_3$ in the model as well? Explain.

4. Infidelity data, known as Fair's Affairs, is based on a cross-sectional survey conducted by Psychology Today in 1969 and is described in Greene (2003) and Fair (1978). This data contains several variables collected on 601 respondents which hold information such as

   - $Y$: whether they have affairs during the past years (1=yes, 0=no)
   - $X_1$: age
   - $X_2$: years married
   - $X_3$: how religious they are (on a 5-point scale from 1=anti to 5=very)
   - $X_4$: a self-rating on happiness toward their marriage (from 1=very unhappy to 5=very happy)

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

$$P(Y=1) = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4}}{1+e^{\alpha+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4}}.$$

   (a) (5 points) Which model equation(s) below is(are) equivalent to the model above? [**There may be more than one correct choice**]

   A. $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$
   B. $\text{logit}(P(Y=1)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$
   C. $\text{logit}(P(Y=1)) = e^{\alpha+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4}$
   D. $P(Y=1) = \text{logit}(\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)$
   E. $P(Y=1) = \frac{1}{1+e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4)}}$

   (a) _____

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

Table 3: Logistic regression model coefficients

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| Constant | 1.931 | 0.610 | 3.164 | 0.002 |
| age | −0.035 | 0.017 | −2.032 | 0.042 |
| yearsmarried | 0.101 | 0.029 | 3.445 | 0.001 |
| religiousness | −0.329 | 0.089 | −3.678 | 0.000 |
| rating | −0.461 | 0.089 | −5.193 | 0.000 |

(c) (4 points) Interpret the number 0.101 in Table 3.

(d) (6 points) Predict the probability of having affairs for a 32-year-old, very religious respondent who has been married for 8 years and has been very happy toward the marriage.

5. (5 points) (a) (2 points) Write down the interpretation of $R^2$ for the regression model:

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \quad \varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

(b) (3 points) Consider an SLR model. Draw and use a plot to illustrate the interpretation for $R^2$.