THIS IS A PRACTICE MIDTERM EXAM -- YOUR EXAM WILL BE DIFFERENT.

Online STAT 705

Sample Midterm Exam Name ______Answer Key________

Question 1

An experiment was conducted to explore the relationship between the temperature (in degrees) at which a chemical reaction occurs and the yield (in ppm) of the reaction.

Use the provided SAS output to answer the following questions. The SAS output contains information about two fitted models:

- a linear model $\left(Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i\right)$, and
- a quadratic model $\left(Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \varepsilon_i\right)$.

a. (5 points) Based on the results in the SAS output, does it appear that the linear regression model is appropriate for these data? Briefly explain.

> No. The residual plot shows a quadratic shape, when it should have no pattern.

b. (5 points) Use the estimated linear regression equation to estimate the yield when the temperature is 110 degrees. (round your answer to 3 decimal places)

$$\hat{Y} = 51.26492 + 0.20965 * 110$$
$$\hat{Y} = 74.326 \text{ ppm}$$

c. (5 points) Use the estimated quadratic regression equation to estimate the yield when the temperature is 110 degrees. (round your answer to 3 decimal places)

$$\hat{Y} = 77.72933 - 0.15359 * 110 + 0.00120 * 110^2$$
$$\hat{Y} = 75.354 \text{ ppm}$$

d. (5 points) Which of these two estimates is more reliable? Briefly explain.

> The estimate from the quadratic model is more reliable because the linear model violates an assumption.

Question 2

In a project to study age and growth characteristics of selected mussel species, researchers measured the age (in years) and weight (in pounds) of numerous mussels at three distinct locations (A, B and C).

a. (10 points) Write the complete model specification for an interaction model. Include all subscripts and define all the variables.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

where

$$Y = \text{weight of } i^{th} \text{ mussel}$$

$$x_{1i} = \text{age of } i^{th} \text{ mussel}$$

$$x_{2i} = \begin{cases} 1 \text{ if } i^{th} \text{ mussel is from location A} \\ 0 \text{ otherwise} \end{cases}$$

$$x_{3i} = \begin{cases} 1 \text{ if } i^{th} \text{ mussel is from location B} \\ 0 \text{ otherwise} \end{cases}$$

$$x_{4i} = x_{1i} \times x_{2i} \text{ (the interaction between age \& location for location A}$$

$$x_{5i} = x_{1i} \times x_{3i} \text{ (the interaction between age \& location for location B}$$

$$\varepsilon_i = \text{random error for } i^{th} \text{ mussel}$$

$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ and β_5 are population parameters

Assume $\varepsilon_i \sim \text{NIID}(0, \sigma^2)$

Use the SAS output to answer the following questions.

b. (5 points) Would an additive model be appropriate for these data? Explain.

No. The test for interaction has test statistic F = 24.15 and p-value < .0001.

c. (5 points) Write the estimated equation for location B. (simplify, and round to 3 decimal places)

$$\text{weight}_{\text{locB}} = (-0.861 - 0.214) + (0.343 + 0.306) \times \text{Age}$$

$$\text{weight}_{\text{locB}} = -1.075 + 0.649 \times \text{Age}$$

This question continues on the next page.

Question 2, continued

d. (5 points) In the parameter estimates table, look at the line for 'location A'. The test statistic is 0.46, with p-value 0.6464. Does this mean "location A" can be removed from the model? Briefly explain.

No. This would be removing a single indicator variable from the model. The categorical variable 'Location' has two indicator variables. If we wanted to test whether or not Location needed to be included in the model, we would have to look at the line for "location" (in the table below), which has p=0.5315, so we would not reject the null hypothesis that BOTH of location's indicator variables can be removed from the model. We do, however, need to keep the age*location interaction (p<.0001), and it is customary to keep all of the main variables that are involved with an interaction.

e. (5 points) Consider this table in the SAS output:

```
Source        DF  Type III SS   Mean Square   F Value   Pr > F
location       2   1.1045181     0.5522591     0.64    0.5315
age            1  379.9989681  379.9989681   438.34    <.0001
age*location   2   41.8750322   20.9375161    24.15    <.0001
```

In the first line, the test statistic is 0.64 and the p-value is 0.5315. Write the null and alternative hypotheses being testing, using the notation you defined in part (a).

$H_0 : \beta_2 = 0$ and $\beta_3 = 0$

$H_a : \beta_2 \neq 0$ or $\beta_3 \neq 0$

f. (5 points) According to the assumptions for a linear model, the errors are independent, normally distributed with mean 0 and common variance σ^2. What is the estimate for σ^2?

$\hat{\sigma}^2 = MSE = 0.8669$

Question 3.

Consider an experiment in which the researcher wants to determine a relationship between the seal strength of a bread wrapper stock (Y) and three predictor variables: sealing temperature (x_1), cooling bar temperature (x_2), and percent polyethylene in the stock (x_3). The researcher felt that it might be appropriate to include interaction terms and terms that are quadratic in the predictors, so he generates the model

Model 1: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{1i}^2 + \beta_5 x_{2i}^2 + \beta_6 x_{3i}^2 + \beta_7 x_{1i} x_{2i} + \beta_8 x_{1i} x_{3i} + \beta_9 x_{2i} x_{3i} + \varepsilon_i$

After fitting the model and reviewing the SAS output, the researcher believes that he can remove all three of the interaction terms (x1x2, x1x3, and x2x3). He removes these terms and fits the new model:

Model 2: $Y_i = \tau_0 + \tau_1 x_{1i} + \tau_2 x_{2i} + \tau_3 x_{3i} + \tau_4 x_{1i}^2 + \tau_5 x_{2i}^2 + \tau_6 x_{3i}^2 + \varepsilon_i$.

Now the researcher must choose which model he should use to complete his analysis.

Use the provided SAS code and output to conduct a nested model F test for these hypotheses:

$H_0 : \beta_7 = \beta_8 = \beta_9 = 0$

H_a : at least one of β_7, β_8, and β_9 is not 0

Please show your work in a logical fashion on the next page, and answer the specific questions below.

a. (3 points) What is the value of the test statistic? **0.876**

b. (3 points) What is the critical value for this test? (use α = 0.05) **3.71**

c. (3 points) Which hypothesis $(H_0$ or $H_a)$ do you decide is true? **H_0**

d. (3 points) Based on the available information, which model (Model 1 or Model 2) should the researcher use? **Model 2**

e. (3 points) What other information would you like to see about these two models before you choose a model? **Diagnostic plots, in order to assess model assumptions**

Use this page to show your work for Question 3.

The full model is Model 1. SSE = 12.0215, dfE = 10

The reduced model is Model 2. SSE = 15.18512, dfE = 13

The test statistic is

$$\begin{aligned} F &= \frac{\left(\text{SSE}_{\text{Reduced}} - \text{SSE}_{\text{Full}}\right) / \left(\text{dfE}_{\text{Reduced}} - \text{dfE}_{\text{Full}}\right)}{\text{SSE}_{\text{Full}} \,/\, \text{dfE}_{\text{Full}}} \\ &= \frac{\left(15.18512 - 12.02513\right) / \left(13 - 10\right)}{12.02512 \,/\, 10} \\ &= 0.876 \end{aligned}$$

The critical value is 3.71
(This is from the F table with 3 and 10 degrees of freedom, using α=0.05.)

The test statistic is NOT greater than the critical value, so we do NOT reject the null hypothesis. So we conclude that the three β's are 0, and that the reduced model is adequate.

Question 4.

Obesity is a common, serious and costly disease. In 2010, Mayor Bloomberg of New York City supported a law restricting the sale of sugary drinks (Coke, Pepsi, etc.) because he believes that consuming sugary drinks contributes to the rate of obesity, so reducing the consumption of sugary drinks will cause a reduction in the rate of obesity. To support his position, Mayor Bloomberg surveyed residents in various neighborhoods across the city. The analysis consists of describing the relationship between two variables:

X = percent of adults in the neighborhood who drink at least one sugary drink per day, and
Y = percent of adults in the neighborhood who are obese.

Use the SAS output to answer the following questions.

a. (5 points) How many neighborhoods are in the sample?

34. (This is because the degrees of freedom for Corrected Total is 33, and this is always one smaller than the sample size.)

b. (5 points) Since there is a relatively strong correlation between X and Y (the sample correlation is 0.76), Mayor Bloomberg argued that a reduction in X will cause a reduction in Y. Is this a valid argument? Briefly explain.

This argument is NOT valid. The data are from an observational study, and causation can be inferred only from data derived from an experimental study.

c. (5 points) Suppose another neighborhood in New York City has 5 percent of adults that drink at least one sugary drink per day. Why would it be in appropriate to use the results of this analysis to estimate the percent obese for this neighborhood?

It would NOT be appropriate because the smallest X is about 10 and using X = 5 would be extrapolation.

Question 5.

An assistant in the district sales office of a national cosmetics company is analyzing data on sales and expenditures in several of the district's territories. The data contain four variables:

Y = sales (in thousands of cases)
X1 = expenditures for point-of-sale displays in department stores
X2 = expenditures for local media advertising
X3 = expenditures for national media advertising

X1, X2 and X3 are in thousands of dollars.

A multiple linear regression model was fit to the data, and the results are shown in the SAS output. (Note: None of the variables have been transformed.)

a. (5 points) Interpret the slope on local media expenditures.

If expenditures for local media advertising increases by $1,000 and expenditures for point-of-sales displays and national media advertising remain unchanged, then we expect the sales to increase, on average, by 0.743 thousand cases (i.e. 743 cases).

b. (5 points) Does it appear that there are outliers in the data? Explain.

As shown in the studentized residual graph (middle graph on top row), one observation has studentized residual smaller than –2. This would be considered a low outlier.

c. (5 points) Does it appear that multicollinearity is present? Explain.

Yes. Two of the three variance inflation factors are greater than 10.