STA-590 Applied Statistics Qualifier January 2020

Assume 95% levels of confidence (α = .05) unless otherwise indicated. You should hand in your computer software output along with your analysis of each problem.

1. [30 points] (prob1.xls) The yield (Y) of a chemical process depends on the temperature (X_1) and pressure (X_2). The following regression model has been found to be appropriate for this process.

$$Y = \gamma_0 (X_1)^{\gamma 1}(X_2)^{\gamma 2} + e$$

Based upon this information you are asked to do the following:

A. (8 points) Using an appropriate method, obtain the starting values for γ_0, γ_1 and γ_2. State these starting values and show, mathematically and with software output, how you obtained them.

B. (7 points) Using the starting values obtained in part A, find the least squares estimates of the three parameters. Why are the starting values obtained in part A often important and/or necessary?

C. (7 points) Evaluate the validity of your model using the appropriate diagnostics. Provide the necessary output and your comments on the diagnostics you have selected. Is this a valid model?

D. (8 points) Assuming that large-sample inferences can be employed reasonably here and the diagnostics are acceptable (they may or may not be), obtain simultaneous approximate 95% confidence intervals for γ_1 and γ_2. Justify your selection of the appropriate simultaneous inference procedure (Bonferroni, Scheffe, Tukey).

2. [30 points] (prob2.xls) The following data set represents n=200 subjects from a larger study on the survival of patients following admission to an adult intensive care unit (ICU) at a local hospital. The hospital wishes to predict the probability of death for these patients until their discharge from the hospital. The dependent variable (Y) is patient status which has also been re-coded as the variable STA (0 = Patient Lived until discharge 1 = Patient Died before discharge). The remaining variables are coded as follows:

i. AGE: Patient's age in years
ii. CAN: Is cancer part of the present problem? (0 = No, 1 = Yes)
iii. INF: Infection (patient temperature above 101^0 F) probable at ICU admission? (0 = No, 1 = Yes)
iv. SYS: Systolic blood pressure at ICU admission (in mm Hg)
v. TYP: Type of admission (0 = Elective, 1 = Emergency Room Admission)

Based on this information, you are asked to do the following:

A. (7 points) Create an appropriate statistical model that can predict the probability of a patient dying before discharge. Provide your final statistical model (with final equation and numerical coefficients). Assume that the diagnostics for you chosen model are acceptable.

B. (8 points) After constructing a valid model, use the results to discuss the statistical influence each significant predictor has on the probability of dying. In other words, examine the coefficients within the context of your model. How does each significant predictor affect the dependent variable?

C. (7 points) A new patient is admitted with the following characteristics:

AGE: 65
CAN: No
INF: Yes
SYS: 145
TYP: Emergency Room Admission
STA: Unknown

Assume this is not extrapolation. Based on this patient's characteristics, and using your chosen model, estimate the probability and the odds that this patient will die before being discharged from this hospital.

D. (8 points) To increase ICU efficiency, the hospital is considering only one of two possible measurements (SYS [systolic blood pressure] or TYP [type of admission 0=Elective, 1=Emergency Room]). You are asked to recommend which of these two measurements alone would be more accurate in predicting the probability of a patient dying. You may examine each of these variables in the ability to predict the dependent variable (STA). Choose one of these variables for your recommendation and using the standard measures of association (Results tab in Minitab or Association Table in SAS), to justify your choice. After making your choice, discuss the overall influence this variable has on the dependent variable STA.

3. [40 points] (prob3.xls) A local consulting firm is interested in studying the effects of newspaper section placement (News, Business, Sports) and day of the work week (Monday, Tuesday, Wednesday, Thursday, and Friday) on the number of inquiries a company receives resulting from n=4 randomly selected purchased advertisements in the local newspaper. The data set is organized as follows:

   DAY: 1= Monday, 2=Tuesday, 3=Wednesday, 4=Thursday, 5=Friday

   SECTION: News, Business, Sports

   RESPONSE: Number of Inquiries made the following day to the company placing the advertisement.

Consider this a two-factor balanced and completely crossed Analysis of Variance study. The overall objective of this study is to investigate the influence both DAY and SECTION have on RESPONSE. We will employ a simultaneous comparison procedure with the primary objective being the pairwise comparison of factor level means.

A. (8 points) Create an appropriate and valid statistical model that can examine the pairwise comparison of factor means. Justify the validity of your model in terms of diagnostics using software output and your comments. If any modifications to the data are required, discuss your modifications then demonstrate the validity of your modified model. Any further analysis can only be done on a valid model.

B. (8 points) Discuss the results of your model. Specifically:

1. **Is there evidence of any statistically significant interaction between the factors? Discuss why or why not. What does the possibility of interaction between the factors actually indicate within the context of this analysis?**

C. (12 points) As previously stated, the research objective is the pairwise comparison of factor means. We wish to understand the influence of the simple main or main effects on the dependent variable. For this problem, assume we are analyzing both factors, but only report the SECTION factor analysis using an appropriate multiple comparison procedure. Use the underlining technique in your analysis of the SECTION factor means. Justify the multiple comparison procedure you have chosen. It is recommended that <u>hand calculations</u> along with some select software output be used instead of just interpreting the Minitab or SAS output. This is because this software uses slightly different assumptions in calculations than what the course textbook recommends.

D. (12 points) Based on the results of part C, answer the following questions:

1. **Based on the results of your analysis, is there a particular SECTION that is universally "better"? In other words, can we recommend a SECTION that is always best for advertisement? Make a recommendation and justification. Remember that the higher the response, the more successful the advertisement is judged.**

2. **If I wished to advertise on FRIDAY, what sections(s) should I advertise in? Make a recommendation and justification. Remember that the greater the response, the more successful the advertisement is judged.**
3. **If I wished to advertise on MONDAY, what sections(s) should be avoided? Make a recommendation and justification. Remember that the greater the response, the more successful the advertisement is judged.**

THE END