Class: STAT_____ Name _____ ID _____

Due: 11:59 PM Friday, Oct 16 2020. Submit your work to Dropbox.

Instructions

Please do the Mid-term test questions all by yourself. You can use the textbook, lecture note, previous homework, but you are not allowed to communicate with anyone else. If you do not understand the questions, please feel free to ask the instructor. Do not share any part of your answer with others. If plagiarism is detected or suspected, the instructor will have to report it to the department based on the Academic Integrity policy. Academic Dishonesty is strictly proscribed and if found may result in student discipline up to and including dismissal from the university. The grade for this test will be counted as zero. All individuals involved in plagiarism will get the same penalty.

Include the questions in your answer sheet. Arrange your answer in the same order as the questions. Mark the question number clearly before the answer. Answer the questions one by one and separate questions by two lines line spacing.

Interpret the outputs and results in plain language. No interpretation, no credits.

For R users, please submit both the .RMD file and the .pdf file that generated by the .RMD file. For JMP users, please include screenshots of steps, outputs, graphs, and anything that are necessary to interpret your answer in the .pdf file.

Here is a letter written by Dr. Andrew Heiss that I want to share with you. "… Your grade on an exam says nothing about your value as a person. Exam grades are imperfect measures of what you have actually learned. You have worked hard so far this semester, and you have learned a lot, regardless of what your score might say. Grad School is hard, and just the fact that you are here in one of top programs in the country is so impressive! PLUS we're in the middle of a global pandemic, there's an election in less than one month….Just try your best…"

The data set Midtest_data shows summary statistics about the 10 national league baseball teams for the 1965 through 1968 seasons. Answer the following questions.

(10) 1. Using forward selection method to fit the best multiple linear regression model (AICc) for the response variable percentage of *wining*. Do not take the variable *year* into your predictor variable.

(10) 2. Discuss the significance of your fitted model. Your interpretation should include the test statistic and the p-value.

(10) 3. Use t-tests to assess the contribution of each predictor to the model. Discuss your findings.

(10) 4. Give the 95% confidence interval (CI) on the **mean percentage of winning** for the runs=707, ba=0.254, dp=152, walk=467, sa=916. Your answer must include the formula for calculating the CI. (You do not have to use all the values if your best model does not contain the corresponding variables.)

(10) 5. Give the 95% prediction interval (PI) on **a new observation of the percentage of winning** for the runs=707, ba=0.254, dp=152, walk=467, sa=916. Your answer must include the formula for calculating the PI. (You do not have to use all the values if your best model does not contain the corresponding variables.)

(10) 6. Compare the results from part 4 and part 5. Which interval has a longer interval length? Explain the reason.

(10) 7. Let lag=15, calculate the Sample ACF, the corresponding Z-Statistic, and Ljung-Box Statistic of the percentage of winning. The output should be similar to the Table 2.3 on page 71 in the textbook. It will be fine if your outputs include all information contained in Table 2.3. The **values** of ACF, Z-statistic, and Ljung-Box statistic are needed. Is there an indication of non-stationary behavior in the residuals?

(10) 8. Let the lag=15, calculate the variogram of the percentage of winning. What can you tell from the variogram?

(10) 9. Plot the 4 in 1 residual plots (QQ plot, Fitted value vs Residual, Histogram of Residual, and Observation order vs Residual) and interpret the graphs. The graphs should be similar to Figure 3.1 on page 138 in the textbook.

(10) 10. Discuss the model adequate by analyzing the residuals. Your output should be similar to the table 3.7 on page 141 in the textbook. Based on your outputs, answer the following questions: Are there any outliers? High leverage observations? High influential observations? Use criterions given in the textbook.