STAT 408 Homework 3

Due by 11:55 pm, Sunday, 10/16/2022

50 points

Please provide detailed calculation and explanation in your solution. Points will be deducted for skimpily written answers. This homework will also require coding in R. On the coding part, the homework solutions should also include detailed description, R code, and output. Write your answers, scan them, and combine to a single pdf file. Name this file as yourname_hw3 and upload to Sakai.

1. (10 points) In this question, we will use the prostate dataset. Import this dataset and answer following questions.

a. Compute a 95% CI for the parameter associated with age. Use the manual method.

b. Compute a 90% CI for the parameter associated with age. Use the manual method.

c. Based on these two CIs, what can we expect the p-value of this parameter in t-test? Compare your conclusion with the p-value output by summary function.

d. Conduct a permutation t-test for predictor age in this model.

e. Remove all the predictors not significant at the 5% level. Use anova function to conduct an F test to test this model against the original full model. Which model is preferred? Give your reason.

2. (10 points) In this question, we will use the cheddar dataset. Import this dataset and answer following questions.

a. Fit a regression model with taste as the response and the three chemical contents as predictors. Identify the predictors that are statistically significant at the 5% level.

b. Acetic and H2S are measured on a log scale. Fit a linear model where all three predictors are measured on their original scale. Identify the predictors that are statistically significant at the 5% level for this model. Hint: exponential function is exp().

c. Can we use an F-test to compare these two models? Which model provides a better fit to the data? Explain your reasoning for these two questions.

d. If H2S is increased 0.01 for the model used in (a), what change in the taste would be expected?

3. (10 points) In this question, we will use the teengamb dataset. Import this dataset and answer following questions.

a. Fit a model with gamble as the response and the other variables as predictors. Which variables are statistically significant at the 5% level?

b. Ccheck the meaning of each variable. Does the variable significance in (a) make sense? Give your reasoning.

c. Fit a model with just income as a predictor and use an F-test to compare it to the full model.

4. (10 points) In this question, we will use the sat dataset. It was collected to study the relationship between expenditures on public education and test results. It contains the following variables

Expend: Current expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of dollars)

Ratio: Average pupil/teacher ratio in public elementary and secondary schools, Fall 1994

Salary: Estimated average annual salary of teachers in public elementary and secondary schools, 1994-95 (in thousands of dollars)

Takers: Percentage of all eligible students taking the SAT, 1994-95

Verbal: Average verbal SAT score, 1994-95

Math: Average math SAT score, 1994-95

Total: Average total score on the SAT, 1994-95

Using the sat dataset, fit a linear model with the total SAT score as the response and expend, salary, ratio, and takers as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant. Some questions may be subjective. Show the most valid judgment and give your seasons.

a. Plot residual vs. fitted response to check the constant variance assumption for the errors.
b. Use Q-Q plot to check the normality assumption. What is the shape of the error distribution?
c. Use studentized residuals to check the outliers. Set the cutoff of being "large" as the 5% critical value in t distribution. Note that we need to consider the two sides.
d. Using defbeta function to plot the change of parameter estimation and check influential points. Check influential points in terms of each parameter except the intercept.

5. (10 points) The divusa dataset records the US divorce and social-economic variables in 77 years.

Year: the year from 1920-1996

Divorce: divorce per 1000 women aged 15 or more

Unemployed: unemployment rate

Femlab: percent female participation in labor force aged 16+

Marriage: marriages per 1000 unmarried women aged 16+

Birth: births per 1000 women aged 15-44

Military: military personnel per 1000 population

Fit a model with divorce as the response and the other variables, **except year,** as predictors. Check for error correlation using three methods: plot residuals vs. years; plot residual (t+1) vs. residual (t); fit a linear model of residual (t+1) ~ residual (t). Give your insight and conclusion.