Home Work # 3. AMS 380

Name:________________________________SBU ID:________________________________

Dear all, the homework is due on Tuessday, Sep 21, 2021, at 11:59 PM. Please submit your homework to the Blackboard in a pdf or word (.doc) document. Rmarkdown is highly recommended.

Please include (1) R code; (2) Output from R; (3) Answers to all the questions asked.

1. Is there a simple linear relationship between income and happiness? The dataset 'income.data.csv' tabulates these two variables from a random sample of 498 people. Please write up the entire R code necessary to answer the following questions. You may refer to the following website for the R procedures:

   http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/
   http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r
   http://www.r-tutor.com/elementary-statistics/simple-linear-regression/residual-plot
   https://rpubs.com/iabrady/residual-analysis
   http://www.sthda.com/english/wiki/normality-test-in-r

   (a) Find the least squares regression line.
   (b) Plot the points and the regression line in the same figure.
   (c) Test at $\alpha = 0.05$ whether there is a significant linear relationship between these two variables. What assumptions are necessary? Please test these assumptions.
   (d) Compute the sample correlation coefficient between the two variables and test whether the corresponding population correlation is zero or not at $\alpha = 0.05$.
   (e) Report the coefficient of determination – does this statistic indicate a good linear model fit? (Note: Recall that for simple linear regression, the coefficient of determination is simply the squared sample Pearson correlation coefficient.)
   (f) Generate ANOVA table of the regression and test at $\alpha = 0.05$ whether the regression effect is significant.

2. You are a public health researcher interested in social factors that influence heart disease. You survey 498 towns and gather data on the percentage of people in each town who smoke, the percentage of people in each town who bike to work, and the percentage of people in each town who have heart disease. The dataset 'heart.data.csv' tabulates these variables. Is there a simple linear relationship between the percentage of people in each town who bike to work and the percentage of people in each town who have heart disease?
   Please write up the entire R code necessary to answer the following questions.

   (a) Find the least squares regression line.
   (b) Plot the points and the regression line in the same figure.
   (c) Test at $\alpha = 0.10$ whether there is a significant linear relationship between these two variables. What assumptions are necessary? Please test these assumptions.
   (d) Compute the sample Pearson correlation coefficient between the two variables and test whether the corresponding population Pearson correlation is zero or not at $\alpha = 0.10$.

(e) Report the coefficient of determination – does this statistic indicate a good linear model fit?

(f) Predict the percentage of people in the town who have heart disease if the percentage of people who bike to work is 65% in that town.

(g) Generate 90% confidence interval of the coefficients β_0 and β_1