MET CS 555 Assignment 4 – 20 points

SUBMISSION REQUIREMENTS: **Please submit a single document (word or PDF) for submission.  Your submission should contain a summary of your results (and answers to questions asked on the homework) as well as your R code used to generate your results (please append to the end of your submission). Please use R for the calculations whenever possible. You will lose points if you are not utilizing R.**

**The data “A04.csv” is from a Canadian 1970 census which collected information about specific occupations. Data was collected to develop a regression model to predict prestige for all occupations.** **Use R to calculate the quantities and generate the visual summaries requested below.**

**The first row is the full name and the second row is the name of variables in the data file.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Occupational Title** | **Education Level (years)** | **Income ($)** | **Percent of Workforce that are Women** | **Prestige Score** |
| Title | **EL** | **Inc** | **Perc** | **Score** |

1. To get a sense of the data, generate a scatterplot to examine the association between prestige score and years of education. Briefly describe the form, direction, and strength of the association between the variables. Calculate the correlation coefficient. **(3 points )**
2. Perform a simple linear regression with prestige score and years of education, and briefly summarize your conclusions (no need to do the 5-step procedure here). Generate a residual plot. Assess whether the model assumptions are met. Are there any outliers or influence points? If so, identify them by ID and comment on the effect of each on the regression. **(5 points)**
3. Calculate the least squares regression equation that predicts prestige score from education, income, and percentage of women. **Formally test** (using the 5-step procedure) whether the set of these predictors is associated with a prestige score at the α = 0.05 level (Hint: You should be performing the global test). **(5 points)**
4. If the overall model was significant, summarize the information about the contribution of each variable separately at the same significance level as used for the overall model *(no need to do a formal 5-step procedure for each one, just comment on the results of the tests)*. Provide interpretations for any estimates (of the slopes) that are significant. Calculate 95% confidence intervals for any significant estimates. **(4 points)**
5. Generate a residual plot showing the fitted values from the regression against the residuals. Is the fit of the model reasonable? Are there any outliers or influence points?  **(3 points)**