STAR511 HW8

Questions 1 through 11 (Prestige): Data for n = 102 occupations was collected. The data is available from Canvas as **Prestige.csv**. The variables include:
prestige (Y): Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.
income (X1): Average income of incumbents, dollars, in 1971.
education (X2): Average education of occupational incumbents, years, in 1971.
women (X3): Percentage of incumbents who are women.

Note: Use code to import the data using the occupation name as the row.name. For example:

```
PrestigeData <- read.csv("Prestige.csv", row.names = 1)
```

1. Create pairwise scatterplots for all 4 variables.
2. Calculate pairwise (Pearson) correlations for all 4 variables.
3. Test against the null that the correlation between prestige (Y) and income (X) is zero. Show the output (including p-value) in your assignment.
4. Regress prestige (Y) against income (X). Show the "summary" output in your assignment.
5. Do we have evidence of an association between prestige and income? Briefly justify your response based on your Q3 and/or Q4.
6. Give an interpretation of the slope for income. In other words, explain what this slope is quantifying.
7. Using the model from Q4, create the plots of (A) residuals vs fitted values and (B) qqplot of residuals.
   Note: You can show just the two plots of interest (and save a little space) using code something like this:
   ```
   par(mfrow=c(1,2))
   plot(Model, which = c(1:2))
   ```
8. Regress prestige (Y) against income **and** education. Include the "summary" output in your assignment. This can be done with code like the following.
   ```
   PrModel2 <- lm(prestige ~ income + education, data = PrestigeData)
   summary(PrModel2)
   ```
9. Give an interpretation of the slope for income for the output in Q8 (note that it will not be identical to the interpretation in Q6).
10. Using the model from Q8, create the plots of (A) residuals vs fitted values and (B) qqplot of residuals.
    Note: You should find that these diagnostic plots look noticeably better than the corresponding plots for Q6.
11. Briefly interpret the R^2 value (labeled Multiple R-squared) shown in the output from the Q8.

Questions 12 through 17 (Steel): An engineer was interested in the association between Strength (Y) and coating Thickness (X) in Steel. An experiment was done where data was collected for n = 20 units. The data is available from Canvas as **Steel.csv**.

12. Create a plot of Strength vs Thick.
13. Regress Strength (Y) against Thick (X). Create plots of (A) residuals versus fitted values and (B) qqplot of residuals.
14. Considering your plots from the previous questions, do the regression assumptions appear to be met? Briefly discuss.
15. Perform a quadratic regression and include the "summary" output in your assignment. This can be done with code like the following.

    ```
    SteelModelQ <- lm(Strength ~ Thick + I(Thick^2), data = Steel)
    summary(SteelModelQ)
    ```

16. Report plots of residuals vs fitted values and a qqplot of residuals for the model in Q15. Write a couple sentences describing any noteworthy ways in which they differ from the plots in Q13.
17. Create a scatterplot with the fitted quadratic curve overlaid. This can be done with code like the following.

    Option 1 (Base R):

    ```
    plot(Strength ~ Thick, data = Steel)
    curve(b0 + b1*x + b2*x^2, add = TRUE)
    ```

    Important Note: b0, b1, b2 need to be replaced with estimates from Q15.

    Option 2 (tidyverse):

    ```
    library(tidyverse)
    qplot(x = Thick, y = Strength, data = Steel) +
    geom_smooth(method = "lm", formula = y ~ poly(x, 2), se =
    FALSE)
    ```