**Lab 4 Activity**

Data Set: Duncan’s Data on the Prestige of U.S. Occupations Data

Source: Duncan, O. D. (1961) A socioeconomic index for all occupations. In Reiss, A. J., Jr. (Ed.) Occupations and Social Status. Free Press [Table VI-1].

Variables:

***Id***: Name of profession

***type***: Type of occupation: prof, professional and managerial; wc, white-collar; bc, bluecollar.

***income***: Percent of males in occupation earning $3500 or more in 1950.

***education***: Percent of males in occupation in 1950 who were high-school graduates.

***prestige***: Percent of raters in NORC study rating occupation as excellent or good in prestige.

-https://socialsciences.mcmaster.ca/jfox/Books/Applied-Regression-2E/datasets/Duncan.pdf

**Directions**

1. Create a new R project in a new directory.
2. Copy the Duncan.csv data into your project folder
3. Start a new R Notebook
4. Install and/or load any libraries that you need to complete the activity
5. Import the data into R
   1. Get descriptive statistics to familiarize yourself with the data

**Visualize and Model the Data**

1. We are interested in the effects of ***education*** and ***income*** on ***prestige***
   1. Create a scatterplot matrix of the three variables to help yourself visualize the data
   2. Write a few sentences about what you see. Are there relations between the variables? Are they linear? Are there outliers? Etc…
2. Estimate a linear model that regresses ***prestige*** on ***education*** and ***income***.
   1. Write a few sentences interpreting your output

**Checking Assumptions: Linearity**

An OLS regression model assumes that the form of the relationship between the independent variables and the outcome is linear. That is, that the model is linear in the parameters. To check this assumption, we can:

1. Plot the residuals against each predictor and fitted values using car::residualPlots()
   1. Briefly interpret the plot for each predictor vs. residuals
   2. Briefly interpret the plot for the fitted values vs predictor
   3. Interpret the statistical tests for curvature for each predictor vs residuals plot and the Tukey test for curvature in the fitted values vs residuals plot
2. Create component + residual plots (also called partial-residual plots) for each predictor.
   1. Briefly interpret each plot

**Homoscedasticity**

An OLS regression model assumes that the variance of the errors is constant across the linear combination of the predictors. If the variance of the residuals is related to any of the predictors or to the predicted values (i.e., y-hats, the fitted values), then this assumption may be violated. Violation of this assumption compromises the accuracy of our significance tests because the standard errors might be incorrect. To check this assumption, we can:

1. Reexamine the predictor vs residuals plots you created previously for homoscedasticity
   1. Briefly interpret each plot re: homoscedasticity
2. Conduct a non-constant error variance test using car::ncvTest()
   1. Briefly interpret the results

**Normality of Errors**

An OLS regression model assumes that the errors are normally distributed. Violation of this assumption compromises the accuracy of our significance tests. To check this assumption, we can:

1. Make a histogram of the standardized residuals. Use ***broom::augment()*** to get the standardized residuals.
   1. Briefly interpret the plot
2. Make a QQ plot of residuals (plot empirical quantiles of the studentized residuals, against the quantiles of a normal distribution).
   1. Briefly interpret the plot
3. Conduct a formal test of the normality of the residuals.
   1. Is the test significant or nonsignificant? What does that mean regarding normality of errors?



