Lab Assignment 6: More Regression

Review: Investigating the Relationships between variables
Open the dataset Exercise.JMP in JMP-IN Extra Data. Let’s describe the distributions of the continuous variables in this data set.

1. Look at the distributions of all the continuous variables. Describe each of the distribution in terms of
   - Shape (Is it Normal?)
   - Gaps
   - Outliers
   - Box plots (Identify $Q_1$, $Q_3$, IQR, Upper and Lower Fence)

2. Next look at the correlations between all of the continuous variables. Which variables have the strongest positive correlation? Strongest negative correlation? Do these associations make sense? Which variables had the weakest positive and negative correlations?

Graph your data before fitting models
Some statistical packages don’t show graphs of the regression, while others require you to make an extra effort to see the graph. The following data illustrates what you can miss if you don’t look at the graph.

1. Open Anscombe.JMP from the subdirectory JMP IN Data. Click Analyze $\rightarrow$ Fit Y by X four times, to fit $Y_1$ by $X_1$, $Y_2$ by $X_2$, $Y_3$ by $X_3$, and $Y_4$ by $X_4$. For each pair, fit the regression line and answer the following questions:
   - $R^2$?
   - RMS Error?
   - Slope?
   - Is the Residual Plot OK?
   - Is there a problem with this regression? If so, what is it?

2. Look back at the scatter plot of $X_2$ and $Y_2$. The fit of the regression line is NOT adequate. However, a simple of transformation of $Y_2$ can solve all your problems. First, select the red arrow next to Bivariate Fit and then choose Fit Polynomial $\rightarrow$ 2, quadratic.
   - What is the new regression line?
   - $R^2$?
   - RMS Error?
   - Slope?
   - Is the Residual Plot OK?
Outliers and reversing the roles of X and Y

Open the file Movies.JMP from the subdirectory JMP IN Data. Choose Analyze → Fit Y by X, with Foreign $ as Y and Domestic $ as X. Fit a line to the data and examine the slope of the regression line. We will make the assumption/guess that the units are in millions of dollars.

- If a movie grosses 200 million in domestic dollars, how much would you predict it to gross in foreign dollars?
- On average, how far off do you expect your prediction to be?
- Write an explanation of this regression equation in terms anyone could understand.

What movie is the outlying observation at (600,1200)? To find out simply click on that data point. The movie title will appear. Let’s exclude this observation and rerun the regression. First remove the current regression line by clicking on Linear Fit → Remove Fit, then right click on the observation and choose Row Exclude, then refit the regression line, Fit Line.

- What is the new slope and intercept?
- Now for a movie that grosses 200 million in domestic dollars, how much would you predict it to gross in foreign dollars?
- On average, how far off do you expect your prediction to be?

Go back to the dataset, go to the red sign by the Titanic row, right click and go to Clear Row States, this puts Titanic back into the dataset. Now, rerun the analysis, using Foreign $ as X and Domestic $ as Y.

- Is $R^2$ the same?
- What about the slope of the regression line?
- Why does the slope change and the correlation does not?
- Which regression model made more sense? Why?