ISyE 2028 – Basic Statistical Methods - Fall 2015
Bonus Project: "Big" Data Analytics
Final Report

By: **Alan G. Johnson**

Title: **Analysis of the Relationship between University Ranking and Endowment**

Introduction:

Students deciding which school to attend often use university rankings as an indication of what school they should attend. A university endowments are financial assets that a school, university system, or foundation invest in order to gain an income to use to hire professors, upgrade facilities, fund scholarships, or lower student tuition. A university's endowment could be a factor that impacts many important factors that attract students to universities; therefore, I investigated the relationship between a university's endowment and its ranking.

Analysis:

I used the most up to date data that was available to me at the time of this analysis. The data for the ranking of the universities was the 2016 rankings of national universities by *U.S. News and World Report.* The data for the university endowments was collected online from the *National Association of College and University Business Officers* which publishes university endowment information for all U.S. and Canadian institutions each year. University endowments are often shared and reported as part of a university system endowment; therefore, I found the endowments of the institutions that comprise the University of California System on the website for the University of California Office of the President.

The data sources for the information and the question of interest are ideal for linear regression. Before conduction the investigation, I thought that there would be a linear model for the relationship between university ranking and endowment.

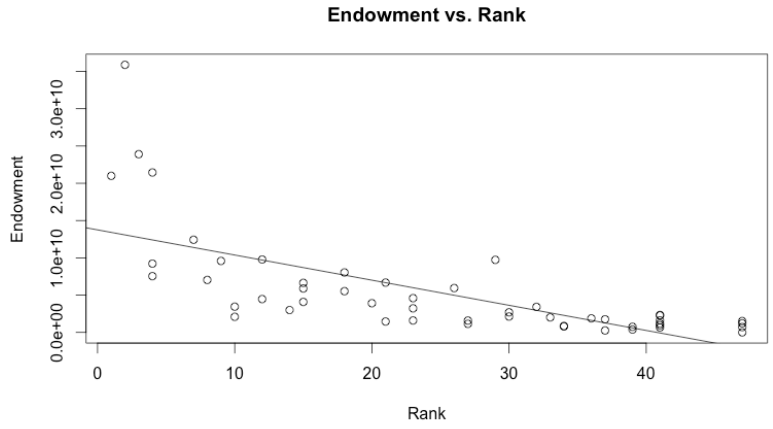
The Simple Linear Regression Model was calculated using R:

$$Y_{Endowment} = 1.375\times10^{10} - 3.377\times10^{8} \times \mathrm{X}_{Rank}$$

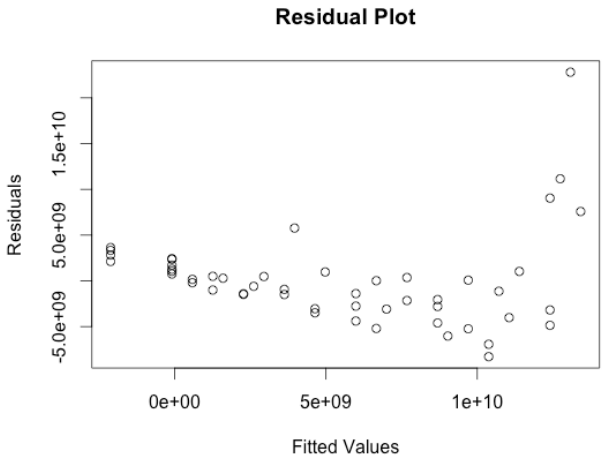
The R summary of the Simple Linear Regression Model is:

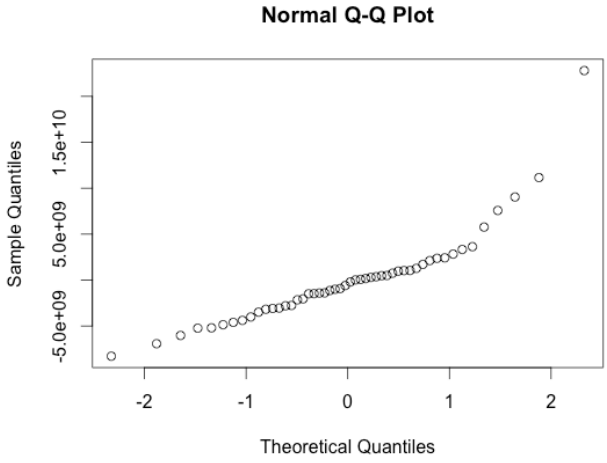
```
Residuals:
Min          1Q      Median          3Q         Max
-8.284e+09 -2.985e+09 -3.991e+08  1.217e+09  2.280e+10

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.375e+10  1.459e+09   9.425 1.70e-12 ***
rank        -3.377e+08  5.139e+07  -6.571 3.34e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.081e+09 on 48 degrees of freedom
Multiple R-squared:  0.4736,    Adjusted R-squared:  0.4626
F-statistic: 43.18 on 1 and 48 DF,  p-value: 3.341e-08
```

Before creating a simple regression model, I used R to create a scatterplot of the data:



I also used R to create a residual plot and a normal probability plot of the residuals:





The data in the scatterplot of the data indicates that a linear model could be a good fit for the data except for a few of the points with higher endowments and lower rankings. These points are influential in the calculation of the simple linear regression model; however, there is no statistical reason to remove these points from consideration.

In order for the linear regression model to be a good fit for the relationship, the residual plot should show a random distribution; however, the residual plot follows a roughly "U" shaped distribution that is not random, which indicates a nonlinear distribution. The Normal Probability Plot (Normal Q-Q Plot) has long tails, which indicate more variance than expected in a normal distribution. The analysis of these graphs indicate that the linear model may not be the best fit for the data.

The adequacy of a regression model can also be measured by the coefficient of determination: it is the square of the correlation coefficient between Y and X. The coefficient of determination for the regression model is 0.4736. The coefficient of determination predicts that around 47 percent of the variation in the endowment can be predicted by rank.

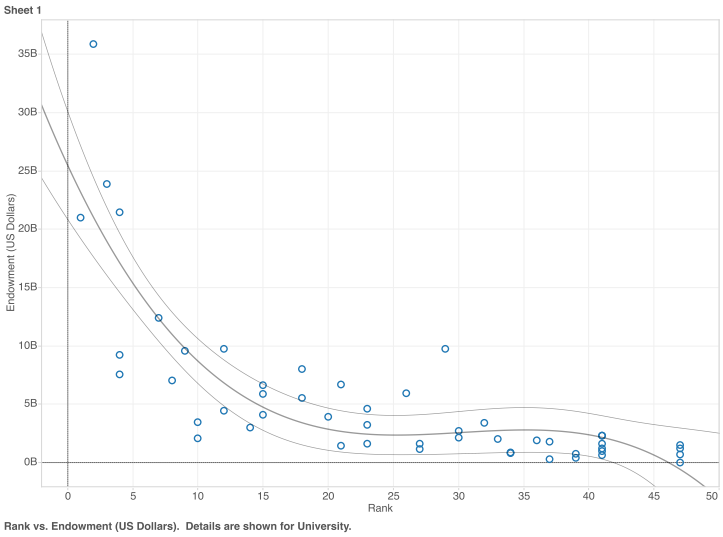
Conclusion:

The scatterplots of the data and the residuals, the normal probability plot, and the coefficient of determination indicate that the linear regression model do not fit the data as well as I predicted.

The best regression model might be better determined with a larger sample size. The data is available for the endowments of over eight-hundred universities in the U.S. and Canada, but not all of the universities in the U.S. and Canada are ranked, which would limit the sample size of the regression model. Other tests could also be used to determine the significance of the linear relationship observed, and if the linear relationship was supported, the regression equation could be used to predict the endowment of a university given its rank.

Using the statistical visualization software Tableau, nonlinear regression models can be easily created, and I found that a polynomial regression model fit the data better than the linear model. In each case the F statistic can be used to determine whether the relationship is statistically significant. While, I did not find the linear model that I hoped to, I suspect that another model would fit this data. On the following pages I included an example of a nonlinear regression model created by Tableau and the linear regression model created by Tableau.

This is an example of a nonlinear regression model obtained in Tableau:



This plot has confidence bands (the upper and lower regression lines) which form the 95% prediction limits.

Trend Lines Model

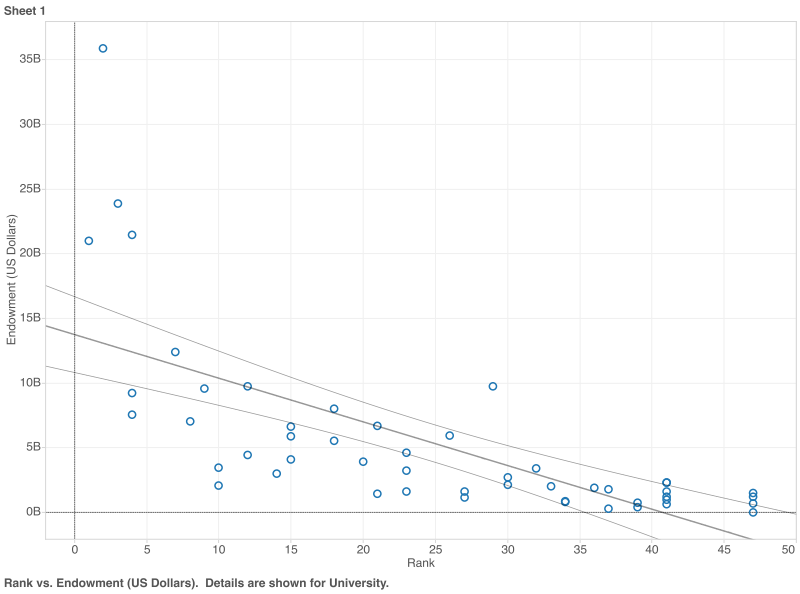
A polynomial trend model of degree 3 is computed for Endowment (US Dollars) given Rank. The model may be significant at p <= 0.05.

Model formula:	(Rank^3 + Rank^2 + Rank + intercept)
Number of modeled observations:	50
Number of filtered observations:	0
Model degrees of freedom:	4
Residual degrees of freedom (DF):	46
SSE (sum squared error):	6.92424e+20
MSE (mean squared error):	1.50527e+19
R-Squared:	0.705837
Standard error:	3.87978e+09
p-value (significance):	< 0.0001

Individual trend lines:

Panes		Line		Coefficients				
Row	**Column**	**p-value**	**DF**	**Term**	**Value**	**StdErr**	**t-value**	**p-value**
Endowment (US Dollars)	Rank	< 0.0001	46	Rank^3	-904043	257072	-3.5167	0.0009949
				Rank^2	8.1956e+07	1.90905e+07	4.29302	< 0.0001
				Rank	-2.40991e+09	4.0452e+08	-5.95745	< 0.0001
				intercept	2.55053e+10	2.30528e+09	11.0639	< 0.0001

This is the linear regression model from Tableau:



This plot has confidence bands (the upper and lower regression lines) which form the 95% prediction limits.

Trend Lines Model

A linear trend model is computed for Endowment (US Dollars) given Rank. The model may be significant at p <= 0.05.

Model formula:	(Rank + intercept)
Number of modeled observations:	50
Number of filtered observations:	0
Model degrees of freedom:	2
Residual degrees of freedom (DF):	48
SSE (sum squared error):	1.2391e+21
MSE (mean squared error):	2.58145e+19
R-Squared:	0.473593
Standard error:	5.0808e+09
p-value (significance):	< 0.0001

Individual trend lines:

Panes		Line		Coefficients				
Row	**Column**	**p-value**	**DF**	**Term**	**Value**	**StdErr**	**t-value**	**p-value**
Endowment (US Dollars)	Rank	< 0.0001	48	Rank	-3.37686e+08	5.13866e+07	-6.57147	< 0.0001
				intercept	1.37546e+10	1.45942e+09	9.42472	< 0.0001

Sources:

US News & World Report Rankings Data:

http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings

NACUBO University Endowment Data:

http://www.nacubo.org/Documents/EndowmentFiles/2014_Endowment_Market_Values_Revised 2.27.15.pdf

University of California Endowment Data:

http://www.ucop.edu/investment-office/_files/report/UC_Annual_Endowment_Report_FY2013-2014.pdf