A Regression Analysis of Economic Performance

by Russell Land

May 5, 2020

Abstract

This final report will analyze data from an economic study, rating the performances of select firms using six different independent variables. These variables include: "Economic Rate of Return," "Good Money," "Gross Operating Surplus," "Operating Gross Margin Rate," "Cash Flow/Turnover," "EBIT (Earnings before interest and tax) Margin," and "Total Debt Turnover." The response variable is "Economic Rate of Return." This response variable is a common measure of economic performance in the finance industry. All of the firms used in the study are all in the same industry and are all located in Romania. The data used in this study comes from a paper titled "Using Linear Regression In The Analysis of Financial-Economic Performances." [1] Our paper will run a regression analysis on the data and determine a model we can effectively use to predict future economic performance from these firms. We will first run a simple linear regression between the regressor "EBIT Margin" and the response variable. Next, we will run a multiple linear regression on all regressors and response variable to determine a best model. We will also check for adequacy of each model, possible multicollinearity issues, and check if the data will need a transformation.

Introduction

The variables in this study are common calculations found in financial summaries, balance sheets, and accounting documents. Let's discuss what each of them mean and how they may influence our regression model. If there is dependence among some variables then we may have to consider looking for multicollinearity. Also, by analyzing where these variables come from may help us predict what variables will be significant in our model, if any.

The first independent variable is Good Money (GM_i). This variable can defined as:

$$\mathrm{GM}_i = \frac{\text{Liquid assets}}{\text{Current debts}}.$$

The term liquid asset is an asset that has preset purchasing power; the asset is not a physical item and is not required to be traded for cash or credit. An example of a liquid asset is cash. Current debt is the term used for money that is owed to a lender at the present time. Together, these two concepts are used to compute Good Money.

The second independent variable is Gross Operating Surplus (GOS_i). This variable can be defined as:

$$\mathrm{GOS}_i = \mathrm{VA} + \mathrm{Se} - \mathrm{It} - \mathrm{Cp}.$$

where 'VA' is added value, 'Se' is operating subsidies, 'It' is the value of rates and taxes owed, and 'Cp' is personnel expenditures. To simplify, this measurement is the money left over after a

firm has produced a product and before they have sold the product. The product adds value to the firm and there could be subsides they receive. These two concepts are considered assets. The value of rates and taxes owed and personnel expenditures are both self-explanatory and can be thought of as a liability. By combining the assets and liabilities mentioned above we arrive at the remaining balance, the Gross Operating Surplus.

The third independent variable is Operating Gross Margin Rate (R_{mb}). This variable can be defined as:

$$R_{mb} = \frac{\text{Gross operating surplus (GOS)}}{\text{Turnover (TO)}}.$$

The numerator is the previous variable we discussed. The denominator Turnover is a measurement to find how fast a company makes their revenues. The faster a firm's revenue flow, the higher the turnover. Another thing to notice about this measurement is it is a function of the previous variable. We have two variables that are dependent of each other. This is something to keep in mind as we start the linear regression analysis.

The forth independent variable is Cashflow/Turnover (CFR_i). For the remaining of this report let's refer to this as just Cash Flow Rate. This variable is defined as:

$$CFR_i = \frac{\text{Operational cash flow (CFO)}}{\text{Turnover (TO)}}.$$

The denominator Turnover is the same definition discussed previously. The numerator Operational Cash Flow is self-explanatory. This refers to the amount of cash (asset or liability) flowing in or our of the firm during operations. The CFR will always be between 0 and 1 because the cash flow can be at most as much as they turnover in a given period of time.

The fifth independent variable is EBIT Margin (M_{EBIT}), or Earning Before Interest and Taxes Margin. The term EBIT by itself is defined as:

$$\text{EBIT} = \text{Gross profit} + \text{Interest expenses}.$$

When discussing the EBIT margin, this is defined as:

$$M_{EBIT} = \frac{\text{EBIT}}{\text{Total income (Vt)}} = \frac{\text{Gross profit} + \text{Interest expenses}}{\text{Total income (Vt)}}.$$

The EBIT, also referred to as Operating Earnings, is a measurement of a firms income before they pay income tax and pay interest for borrowing. This is useful in seeing a firms operations without the cost of capital and the taxes expenses, which can make a big impact. The total income of a firm is the amount of money received over a given period of time, usually one year. This variable is a good indicator of the current financial situation of a firm.

The sixth and last independent variable we will look at is Total Debt Turnover (R_{Dt}). This variable is defined as:

$$R_{Dt} = \frac{\text{Turnover (TO)}}{\text{Total debt (Dt)}}.$$

Again, we have defined Turnover. The total debt is the combined sum of all loans, cost of capital, and operating expenses a firm compiles over a given period of time, usually one year. Together, these two concepts form the Total Debt Turnover measurement.

The response variable from our data is Economic Rate of Return (R_{ec}). This variable is defined as:

$$R_{ec} = \frac{\text{EBIT}}{\text{Average balance of total assets } (\overline{\text{At}})} = \frac{\text{Gross profit} + \text{Interest expenses}}{\text{Average balance of total assets } (\overline{\text{At}})}.$$

The numerator EBIT has already been defined. The denominator Average balance of total assets has been defined by the paper; they say "the ($\overline{\text{At}}$), was determined as an average of the sums reported at the beginning and at the end of the financial period 2008." [cite]

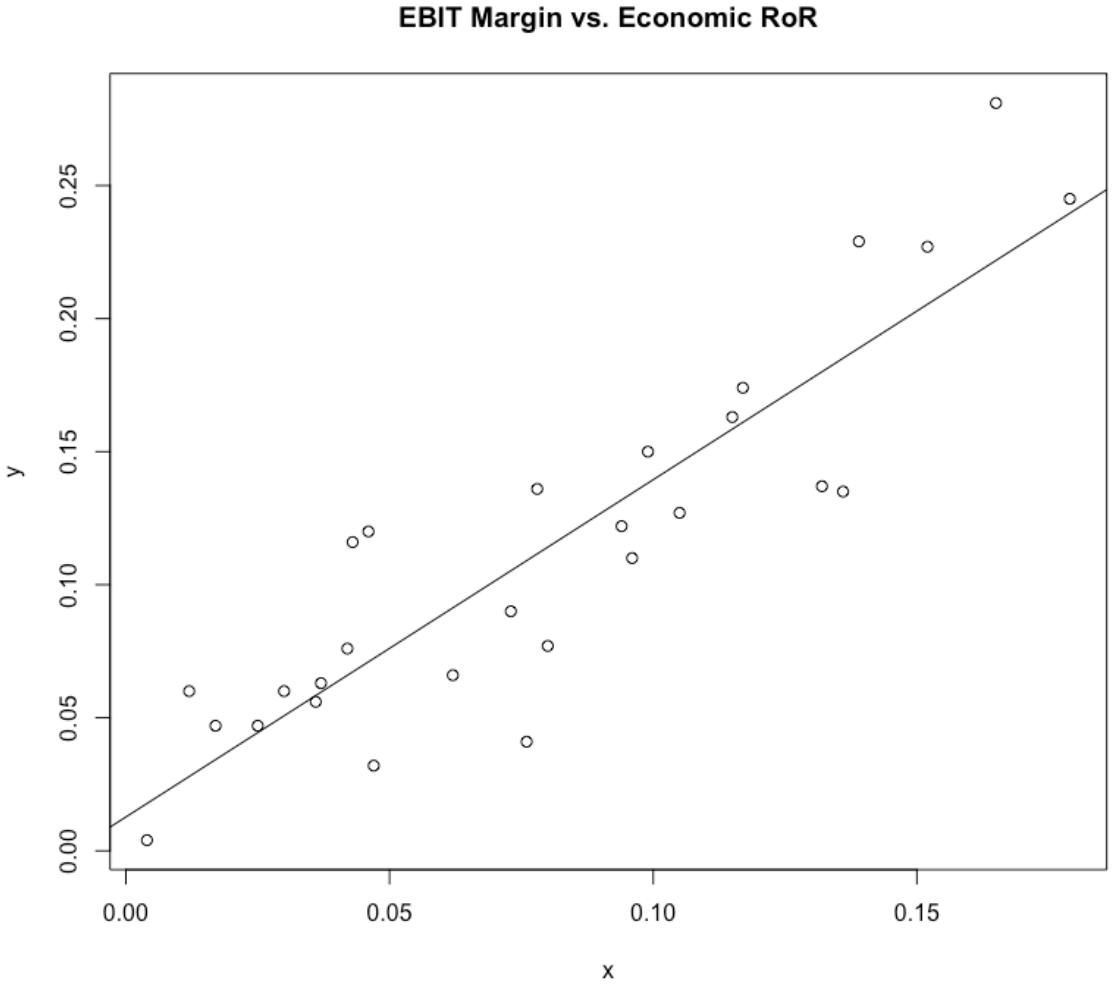
There is a trend we can notice from just analyzing the variables above. Performance of a firm seems to rely heavily on the turnover of a firm. The measurement turnover is in three of the variables above. The measurement EBIT also seems to be an important indicator of economic performance. This measurement is in one of the independent variables, and is the response variable as well. Before we run an analysis, I am going to predict that the variable M_{EBIT} will be a significant variable because itself and the response variable both contain the measurement EBIT, indicating a some kind of relationship.

Firm	Economic Rate of Return	Good Money	Gross Operating Surplus	Operating Gross Margin Rate	Cash Flow/ Turnover	EBIT Margin	Total Debt Turnover
1	0.056	0.442	10383120	0.087	0.057	0.036	3.513
2	0.076	0.022	9869544	0.070	0.027	0.042	2.565
3	0.227	1.111	26685671	0.249	0.136	0.152	5.201
4	0.041	0.003	12138277	0.110	0.024	0.076	1.400
5	0.063	0.060	12032959	0.096	0.056	0.037	1.704
6	0.077	0.004	10798008	0.083	0.012	0.080	0.905
7	0.032	0.007	10696810	0.112	0.052	0.047	1.601
8	0.281	2.452	23165418	0.212	0.194	0.165	9.502
9	0.047	0.006	5421437	0.058	0.027	0.025	1.807
10	0.174	0.043	13324530	0.132	0.063	0.117	1.937
11	0.127	0.062	14044440	0.154	0.039	0.105	1.355
12	0.110	0.006	13390931	0.149	0.058	0.096	1.344
13	0.060	0.040	1492526	0.024	0.007	0.012	3.051
14	0.245	0.025	20313159	0.227	0.046	0.179	1.358
15	0.047	0.044	3735283	0.053	0.023	0.017	3.479
16	0.122	0.087	13251207	0.153	0.083	0.094	3.600
17	0.004	0.095	3989975	0.046	0.057	0.004	2.723
18	0.135	0.177	14794181	0.222	0.034	0.136	0.703
19	0.163	0.165	11972850	0.144	0.087	0.115	2.385
20	0.060	0.044	3873616	0.082	0.023	0.030	1.400
21	0.066	0.009	9019110	0.108	0.059	0.062	1.338
22	0.116	0.164	3811847	0.056	0.003	0.043	1.878
23	0.090	0.004	14018015	0.207	0.110	0.073	1.116
24	0.150	0.074	11856849	0.159	0.061	0.099	1.243
25	0.136	0.261	8376274	0.194	0.064	0.078	1.706
26	0.137	0.025	15355836	0.279	0.092	0.132	0.795
27	0.229	0.058	19331352	0.287	0.180	0.139	2.343
28	0.120	0.346	4741264	0.073	0.039	0.046	3.295

Regression Analysis

Simple Linear Regression Analysis

For our simple linear regression analysis I am going to choose the variable M_{EBIT} to be the regressor and the have the response variable be R_{ec}. Both of these measurement are dependent on the measurement EBIT. Let's see how they relate. We can make a scatterplot to see a visual and make some naive assumptions about linearity.



The model we want to test is a simple linear regression model defined as:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Before we solve for the regression parameters we need to make a few assumptions. These are known as the the regression assumptions:

1. The random error is distributed normally with mean 0 and variance σ^2.
2. Each observation is independent of one another.
3. The variance is constant for any value of x.

From the scatterplot there seems to be evidence of an upward linear relationship between X and Y. Next, let's estimate the regression coefficients for our model above. To estimate β_0 we can use

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and we can estimate β_1 using

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

where

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$
$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

After performing these calculations we obtain $\hat{\beta}_0 = 0.01271747$ and $\hat{\beta}_1 = 1.267282$. So, the estimated regression equation is:

$$\hat{y} = 0.013 + 1.267\hat{x}$$

Next, let's test for significance of regression. There are a few ways to do this; let's start by constructing an ANOVA table. An ANOVA table uses our assumptions of linear regression to equate the analysis of variance with the comparison of population means. Each observation can be treated as its own population. For an analysis of variance test we are testing the following hypothesis:

$$\begin{array}{ll} H_0: \ \mu_{y_1} = \mu_{y_2} = \cdots = \mu_{y_k} & \qquad\qquad H_0: \ \sigma_B^2 = \sigma_W^2 = \sigma^2 \\ H_1: \ \text{at least one } \mu_{y_i} \text{ is different} & \iff \qquad H_1: \ \sigma_B^2 > \sigma_W^2 = \sigma^2 \end{array}$$

Now that we have established our hypothesis we can compute the ANOVA table in R. Below is the ANOVA table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	0.1039	1	0.1039	100.7644
Residual	0.0268	26	0.0010	
Total	0.1307	27		

Our test statistic for the ANOVA test is 100.7644. To compute a p-value, we can use a built-in R function to obtain 1.958e-10. This number is very small and we have a significant test at very small levels of significance. At the 1% level of significance we can conclude there is strong evidence for linear regression. There is another way to test for significance of linear regression under a simple linear regression model. We can calculate a confidence interval and check to see if 0 is contained in that interval. Let's find a 99% confidence interval to test for significance of regression. The estimated parameter $\hat{\beta}_1$ is normally distributed with mean β_1 and variance σ^2/S_{xx}. To construct this confidence interval we can use this information. Since the variance is unknown we can use the mean squared error to estimate it. The confidence interval for the estimated β_1 parameters is:

$$\hat{\beta}_1 - t_{\alpha/2,n-2} \cdot s_e\left(\hat{\beta}_1\right) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2} \cdot s_e\left(\hat{\beta}_1\right)$$

Using R, our 99% confidence interval for the regression parameter β_1 is:

$$0.9164791 \leq \beta_1 \leq 1.6180857$$

After using these two different methods, I think it is safe to assume there is significant evidence for linear regression. In other words, we are 99% confident that the average slope of the regression model will be in the interval $(0.9164791, 1.6180857)$. We could also perform a t-test to check for significance of regression, but this is already enough evidence.

Now that we have evidence of regression let's try to predict the average Economic Rate of Return given that the EBIT Margin is 10%.

$$E[y|x = 0.1] = 0.013 + 1.267(0.1) = 0.139$$

In fact, we can make many predictions. Let's make many predicts for the average Economic Rate of Return and future values of Economic Rate of Return. Below let's define the confidence intervals for both the average Economic Rate of Return and future values of Economic Rate of Return. For the average Economic Rate of Return:

$$\hat{y}_0 - t_{\alpha/2,n-2} \cdot s_e\left(\hat{y}_0\right) \leq \mu_{y|x_0} \leq \hat{y}_0 + t_{\alpha/2,n-2} \cdot s_e\left(\hat{y}_0\right)$$

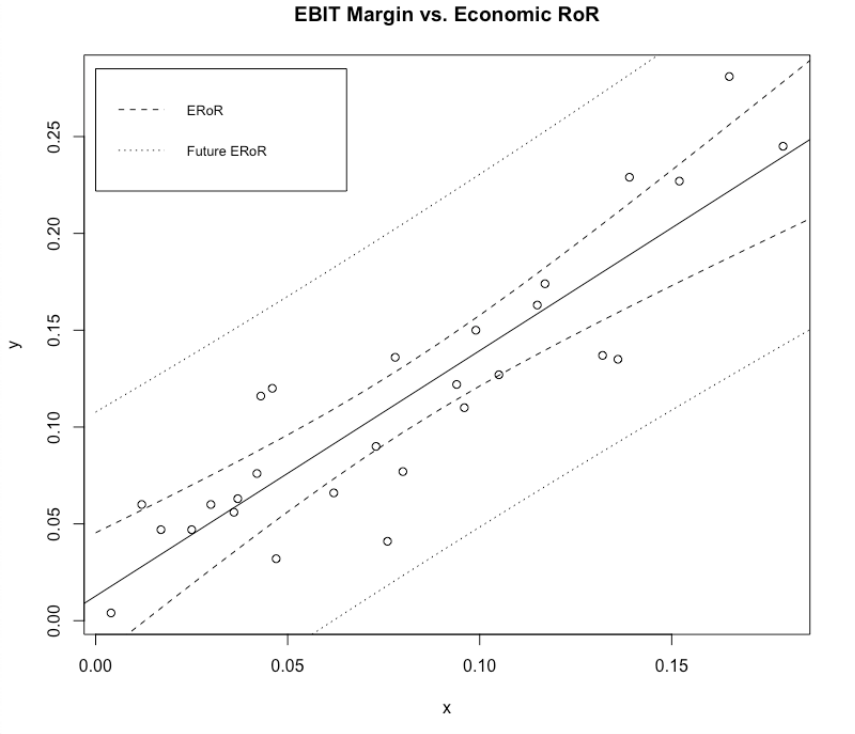
and the confidence interval for future values of Economic Rate of Return:

$$\hat{y}_0 - t_{\alpha/2,n-2} \cdot s_e\left(\Psi\right) \leq y_f \leq \hat{y}_0 + t_{\alpha/2,n-2} \cdot s_e\left(\Psi\right)$$

The respective standard errors are:

$$s_e\left(\hat{y}_0\right) = \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) MS_{Res}} \qquad s_e\left(\Psi\right) = \sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) MS_{Res}}$$

Now we can construct bounds for a 99% confidence interval for both of these quantities. The graph is below:



Beyond the confidence intervals, there are other ways to check for linearity. Another statistic is called the R-squared value. This measure is the ratio between the sum of squares regression and the total sum of squares. Another way to think of this measure is like:

$$R^2 = \frac{SS_R}{SS_T} = \frac{SS_T - SS_{Res}}{SS_T}$$

In other words, the more error there is in the model (SS_{Res}), the lower the R-squared value will be [3]. The R-squared value for this model is 0.795. This is a relatively high value, but there can be room for improvement. Since, this is predicting the economic performance of firms, this is acceptable. Once we introduce more regressors in the model, I will introduce the adjusted R-squared statistic.

Multiple Linear Regression Analysis

For our multiple linear regression analysis, we will introduce five more regressors. The model will include all of the regressors from Table 1 with Economic Rate of Return as the response. We will use matrix formulation for the remainder of the regression analysis. The model is defined as:

$$\vec{y} = X\vec{\beta} + \vec{\varepsilon}$$

where

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{16} \\ 1 & x_{21} & x_{22} & \cdots & x_{26} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & X_{n2} & \cdots x_{n6} \end{bmatrix}, \qquad \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_6 \end{bmatrix}, \qquad \vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The same regression assumptions apply for multiple linear regression. With our model defined and the regression assumptions implied, we can solve for the estimated regression parameters using the least-squares estimate expression:

$$\vec{\beta} = \left(X^t X\right)^{-1} X^t Y$$

When building the model there was an issue because the inverse matrix was turning singular in R. Now the matrix was not actually singular, but was computationally singular, meaning it involved precision that R cannot handle. To avoid this issue I will scale the data, then unscale after I build the model. Using the coefficients γ_i and the scaled coefficients and β as the unscaled coefficients, we can establish the relationship:

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \gamma_0 + \gamma_1 \left(\frac{x_1 - \mu_1}{\sigma_1}\right) + \cdots + \gamma_p \left(\frac{x_p - \mu_p}{\sigma_p}\right)$$

Using all of this, we are finally able to build our model:

$$y = -3.390 - 1.227x_1 - 4.845x_2 - 4.616x_3 + 2.256x_4 + 1.635x_5 + 1.364x_6$$

I am speculating the reasons why this complication happened was because of highly correlated data. This issue was not a surprise as it was mentioned in the paper the data came from [1]. This concept is called multicollinearity. This is a problem that happens when the data is highly correlated and can be fixed by analyzing the the Variance Inflation Factors (VIF). Below are the VIFs for this model, calculated in R.

```
       GM       GOS      OGMR       CFR     EBITM       TDT
 8.114270  6.843655 11.109788  6.427460  8.295903 10.179212
```

The rule of thumb for recognizing multicollinearity is VIF values above 10. Immediately we can see the regressors Operating Gross Margin Rate and Total Debt Turnover are potential risk for the model, so we will rid these of the model. Our new model now includes four regressors. Now let's rebuild our model without these:

$$y = 2.537 + 3.234x_1 - 5.002x_2 + 3.124x_4 + 1.505x_5$$

Next, let's compute an ANOVA table to see if our model is significant. We can compute these all manually in R.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	0.1148	4	0.0.0287	41.3876
Residual	0.0159	23	0.0007	
Total	0.1307	27		

Our F statistic is 41.3876, which corresponds to a p-value of 3.4488e-10. At the 1% level of significance, we can conclude we have significant evidence of linear regression using the regressors listed above.

Let's now introduce another statistic that tests linearity, the adjusted R-squared. The adjusted R-squared is a very helpful measure when more regressors are introduced to the model. The original R-squared value increases as more regressors are introduced to the model, regardless of the importance of them. However, the adjusted R-squared addresses this issue by adjusting for this (dividing by degrees of freedom). The adjusted R-squared is defined as:

$$R^2_{\text{adj}} = 1 - \frac{SS_{Res}/n - p}{SS_T/n - 1}$$

In this model we have both a high adjusted R-squared value and a high R-squared value. These can be interpreted in the following way. We have an adjusted R-squared value of 0.8568, so around 86% of the data is accounted for by the model.

Now that we have established evidence of regression, let's test each coefficient using a t-test. Let's test whether the coefficient for Cash Flow contributes to the response. Let's test the following hypotheses:

$$H_0 : \ \beta_4 = 0$$
$$H_a : \ \beta_4 \neq 0$$

The standard error can be calculated from the covariance matrix and the mean squared error. The standard error for the i^{th} regression coefficient is defined as $\sqrt{C_{ii} \cdot MS_{Res}}$. The test statistic can be defined as:

$$t = \frac{\hat{\beta}_4 - \beta_4}{s_e\left(\hat{\beta}_4\right)} = \frac{3.124 - 0}{0.1791} = 1.744$$

We can repeat this process for each coefficient. We can let software do the rest. We will test at the 5% level of significance.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.536287e-02	1.131812e-02	2.240909	3.498089e-02
XGM	3.233781e-02	1.327691e-02	2.435642	2.302719e-02
XGOS	-5.002444e-09	2.177756e-09	-2.297064	3.105353e-02
XCFR	3.124164e-01	1.790962e-01	1.744405	9.443960e-02
XEBITM	1.505211e+00	2.294826e-01	6.559149	1.081190e-06

All coefficients are significant at the 5% level except for β_3, the one we tested earlier. We can conclude that β_0, β_1, β_2, and β_5 contribute to the response variable at the 95% confidence level.

Next, let's reduce our model one more time using the extra-sum-of-squares method. We will test the regressors x_1, x_2, and x_5. To do this method we will create two subsets for the regression coefficients, $\vec{\beta}_1 = (\beta_0\ \beta_3\ \beta_4\ \beta_6)^t$ and $\vec{\beta}_2 = (\beta_1\ \beta_2\ \beta_5)^t$. We want to test the following hypotheses:

$$H_0 :\ \beta_1 = \beta_2 = \beta_5$$
$$H_a :\ \text{at least one } \beta_j \neq 0$$

Now we can define the test statistic and compute it in R.

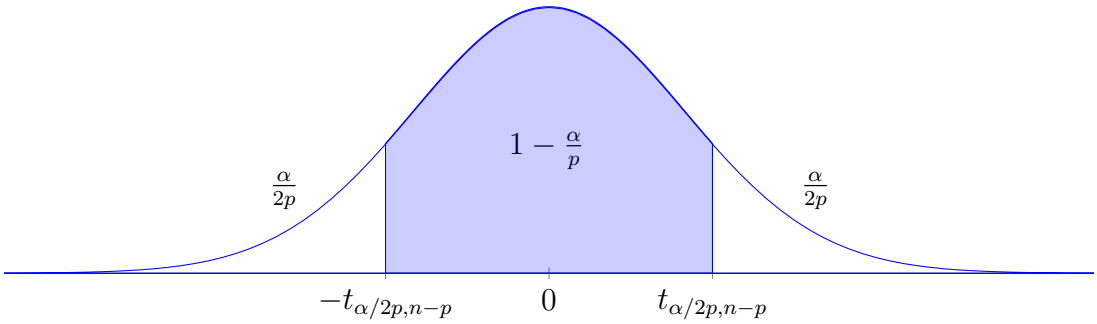
$$F_0 = \frac{MS_R\left(\vec{\beta}_2|\vec{\beta}_1\right)}{MS_{Res}} = 0.0072/0.0006 = 11.0328$$

With the test statistic computed above, we can compute a p-value of 1.4716e-04. At the 1% level of significance, we can conclude that the regressors x_1, x_2, and x_5 contribute to the response variable given that the other regressors are already in the model.

After these couple of tests, I think we have reached an optimal model with the given regressors. Next, let's compute simultaneous confidence intervals using the Bonferoni Method at the 5% confidence level using the optimal reduced model. For each β_j the confidence interval can be defined as:

$$\hat{\beta}_j - cv \cdot s_e\left(\hat{\beta}_j\right) \leq \beta_j \leq \hat{\beta}_j + cv \cdot s_e\left(\hat{\beta}_j\right)$$

where cv is the critical value for the simultaneous confidence intervals. Finding this critical value is a little different than normal. Since it is a simultaneous critical value, we must consider all of the regressors at once. Consider the graph below.



The critical value can be computed in R as 2.700. Next, we can construct our confidence interval by finding the standard error for each coefficient. Below are the confidence interval.

```
                 SCIlower     SCIupper
(Intercept) -5.319910e-03 5.825209e-02
XGM          8.475578e-03 7.599939e-02
XGOS        -8.817808e-09 2.171471e-09
XEBITM       8.212542e-01 2.105101e+00
```

There are a few observations we can make from this simultaneous confidence interval. Two of the confidence intervals contain 0, indicating they are not significant. We have already establish however that our model is significant. One thing to be careful about is the fact that these values are already very close to 0, so these confidence intervals may not be very helpful.

To conclude the multiple linear regression analysis, I think we have found a good representative model for the Economic Rate of Return. In Finance and Economics it is sometimes peoples jobs to predict and forecast the performance of a firm. Linear regression can play a significant role in making predictions and future forecasts for firms. A decision made from a regression analysis may affect a whole firm and could give an advantage to greedy investors. Firms such as hedge funds, who move money and make massive investment, need to make adequate forecasts to make smart decision. At the end of the day, most firms are actually moving around other peoples money. This include hedge funds, investment firms, insurance companies, etc. This leads us into our next topic: model adequacy. It is important to make accurate decision, but it is just as important to know if your model is adequate. An inadequate model can lead to unexpected and devastating returns.

Check Model Adequacy

The next topic is model adequacy. This idea will explore the initial assumptions we made at the beginning as whether they were correct assumptions or not. Statisticians may never know if their assumptions are correct, but there are many ways to sufficiently check. The main ways that we will check for our regression assumptions will be through:

1. Checking whether the current model is in-fact linear.
2. Checking if there is constant variance with respect to the random error.
3. Checking if the random error is normally distributed.

If these assumptions turn out to not be satisfied there may be a few actions taken. A transformation may be applied to change the model and perhaps fix the issue. We may also need to consider a different model, or in the worst case consider different sampling methods or re-sampling. We will only apply a transformation if necessary. Let's again state our model so we are clear what we are analyzing:

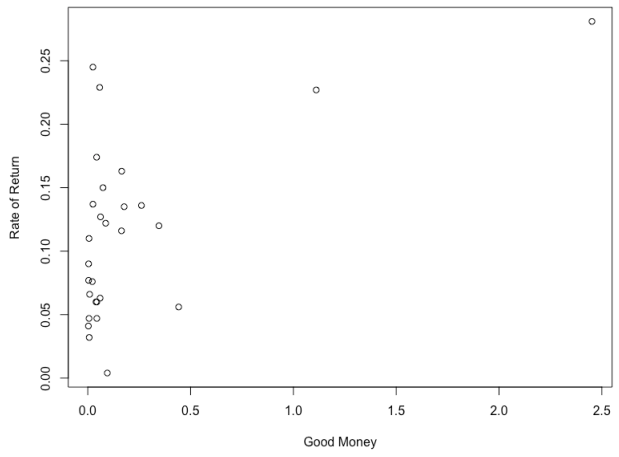
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_5 x_5 + \varepsilon$$

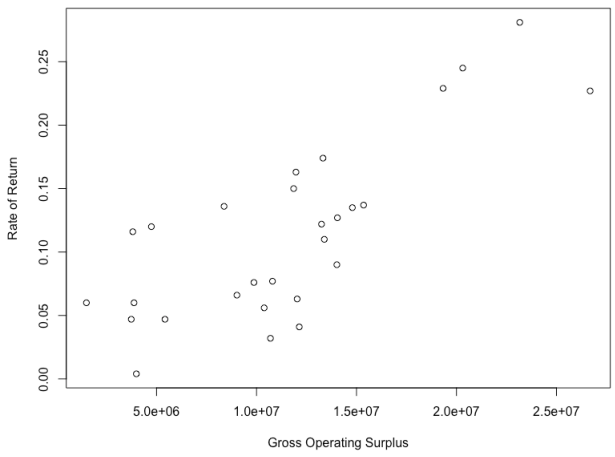
Our first method for checking model assumptions is to check for linearity. We have not calculated a complete regression analysis on the new model. We can do this in R.

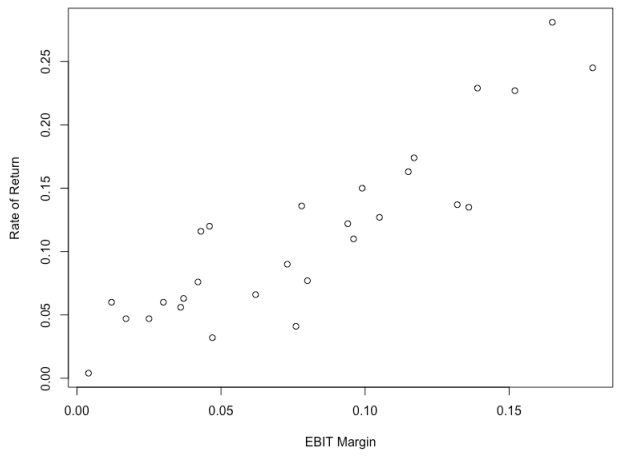
```
Call:
lm(formula = Y ~ X)

Residuals:
```

```
      Min        1Q    Median        3Q       Max
-0.056457 -0.014874  0.001402  0.019397  0.060944

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.647e-02  1.177e-02   2.248  0.03401 *
XGM          4.224e-02  1.250e-02   3.378  0.00249 **
XGOS        -3.323e-09  2.035e-09  -1.633  0.11550
XEBITM       1.463e+00  2.377e-01   6.155 2.33e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02743 on 24 degrees of freedom
Multiple R-squared:  0.8619,Adjusted R-squared:  0.8446
F-statistic: 49.92 on 3 and 24 DF,  p-value: 1.815e-10
```

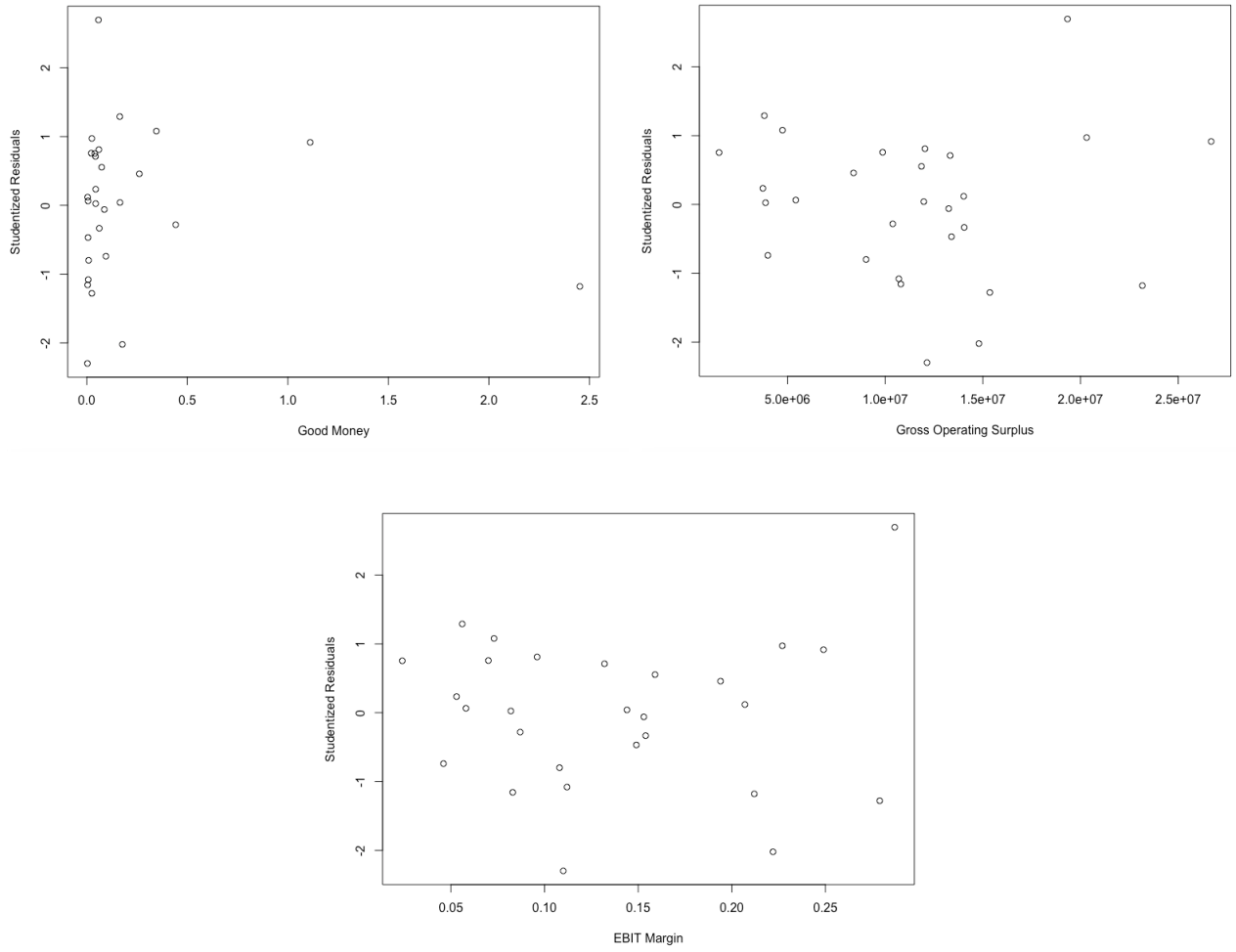
There is an adjusted R-squared value of 0.8446 and an R-squared value of 0.8619. In my opinion, there is a clear linear relationship present. To further check this assumption, let's make plot of each regressor against the response variable.



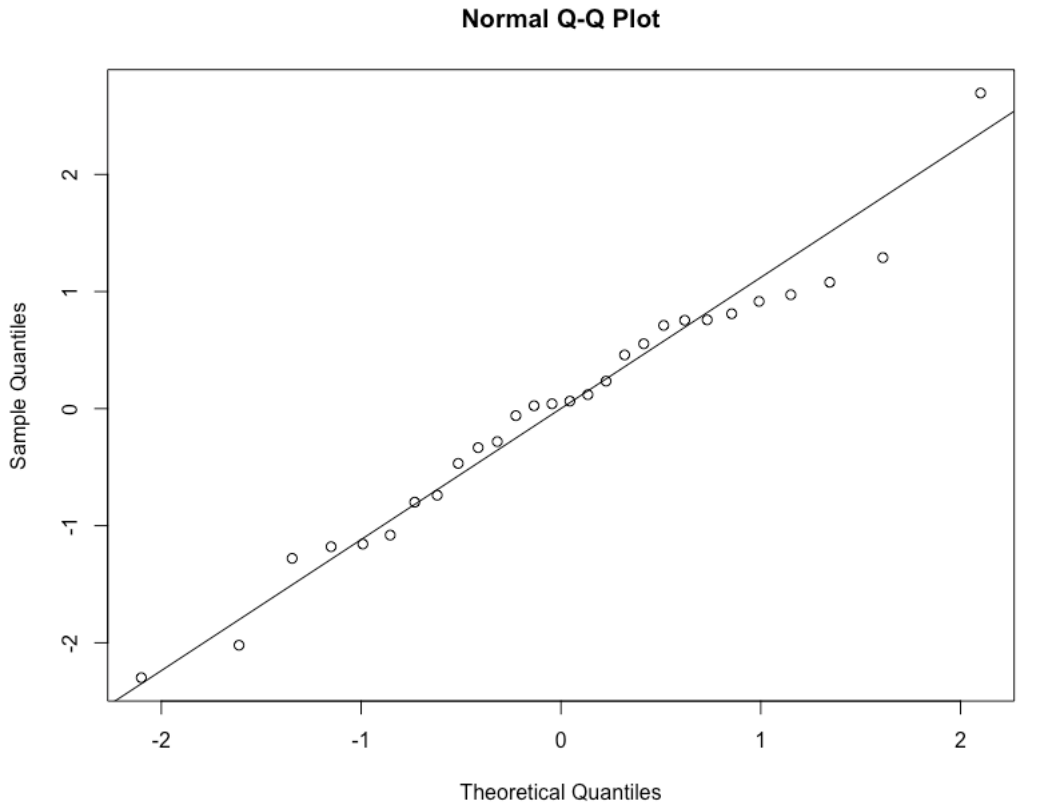




There were very interesting results from these scatterplots. Even though our regression results from earlier pointed toward a strong linear relationship, we may need to reconsider our model. Let's move onto the assumption of constant variance. This can be checked by analyzing the residual plot against the fitted values. Also, also with the residuals plotted against the regressors as well. These plots will be below in that respective order.



From these residual plot and the scatterplot above it is beginning to become obvious of some assumption violations. The first residual plot is not ideal as there are a few outliers. The last residual plot is good enough, but again the second plot is not ideal either. Let's continue with the last regression assumption of normality before we make a decision on how to proceed. Below is a normal probability plot of the studentized residuals.



The assumption of normality is satisfied. Consider the other assumption violations for regression we will try a transformation to try to fix these assumptions. The main violation that I have found is the violation of linearity. If we can fix this, I believe our model can become adequate. After trying many different transformations[2], consider the following model below:

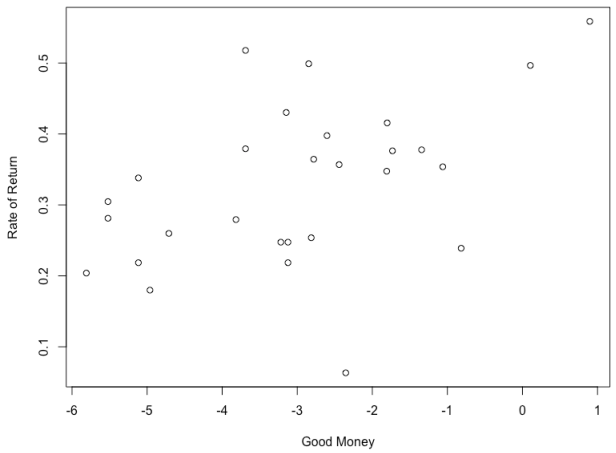
$$\sin^{-1}(\sqrt{y}) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \beta_5 x_5 + \epsilon$$

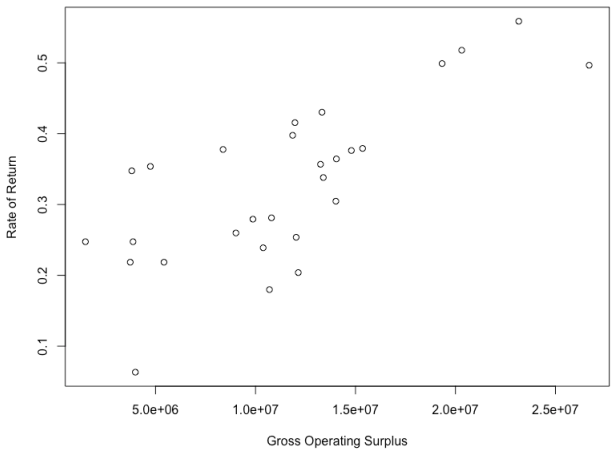
With these transformations defined let's rerun these assumptions tests and let's also run another regression analysis below. Below will be the regression results, residuals, residual plots, scatterplots, and finally the normal probability plot.

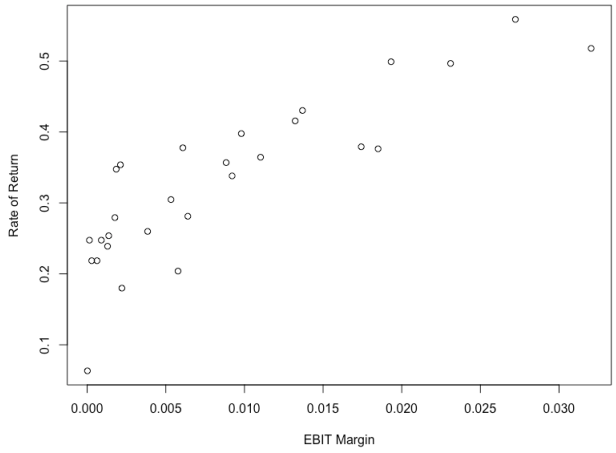
```
Call:
lm(formula = nY ~ nX)

Residuals:
      Min        1Q    Median        3Q       Max
-0.133448 -0.017118  0.006608  0.030542  0.067791

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.229854   0.026772   8.586 6.33e-09 ***
nX1         0.017276   0.005481   3.152  0.00418 **
nX2         1.886722   0.192277   9.813 4.69e-10 ***
---
```

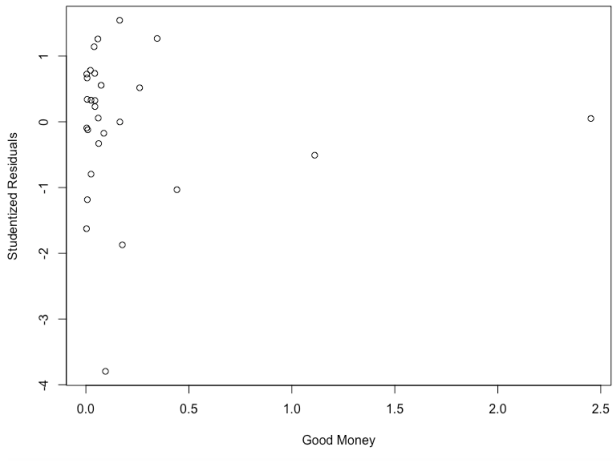
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04723 on 25 degrees of freedom
Multiple R-squared:  0.8398,Adjusted R-squared:  0.827
F-statistic: 65.54 on 2 and 25 DF,  p-value: 1.142e-10
```

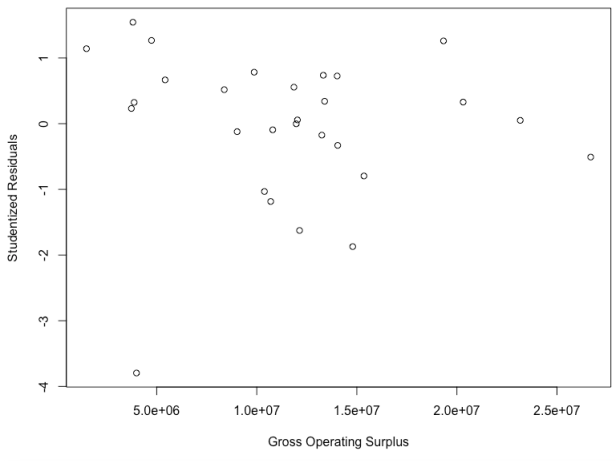
	Residuals	Standardized	Studentized
1	-0.0447621509	-1.031108424	-1.032468405
2	0.0361384204	0.789548778	0.783425869
3	-0.0218475732	-0.516118109	-0.508406223
4	-0.0689927860	-1.574801933	-1.625726099
5	0.0026521641	0.058182850	0.057011178
6	-0.0042248116	-0.095431650	-0.093520575
7	-0.0529565651	-1.175148070	-1.184587678
8	0.0020534819	0.051197688	0.050165915
9	0.0298909170	0.673678123	0.666141215
10	0.0340472255	0.743561629	0.736730710
11	-0.0155419497	-0.336796855	-0.330743368
12	0.0154689202	0.346910008	0.340722086
13	0.0505809082	1.134566590	1.141417616
14	0.0139568321	0.334935855	0.328907557
15	0.0105636843	0.235692382	0.231187425
16	-0.0082132733	-0.177593218	-0.174114971
17	-0.1334481901	-3.062278990	-3.795559818
18	-0.0802963680	-1.784158908	-1.871301895
19	-0.0001064937	-0.002331898	-0.002284784
20	0.0149733506	0.329731617	0.323774487
21	-0.0056345835	-0.123901400	-0.121435374
22	0.0677913449	1.503029661	1.544078338
23	0.0324956878	0.732550851	0.725579932
24	0.0260405983	0.563289491	0.555444785
25	0.0238849308	0.525350251	0.517601033
26	-0.0360185631	-0.802021591	-0.796126189
27	0.0560716858	1.245647283	1.260214415
28	0.0554331565	1.252074702	1.267150850

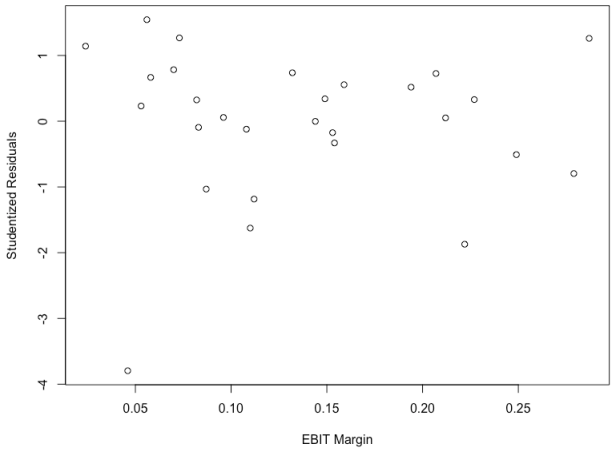

Rate of Return
Good Money
0.1
0.2
0.3
0.4
0.5
-6
-5
-4
-3
-2
-1
0
1

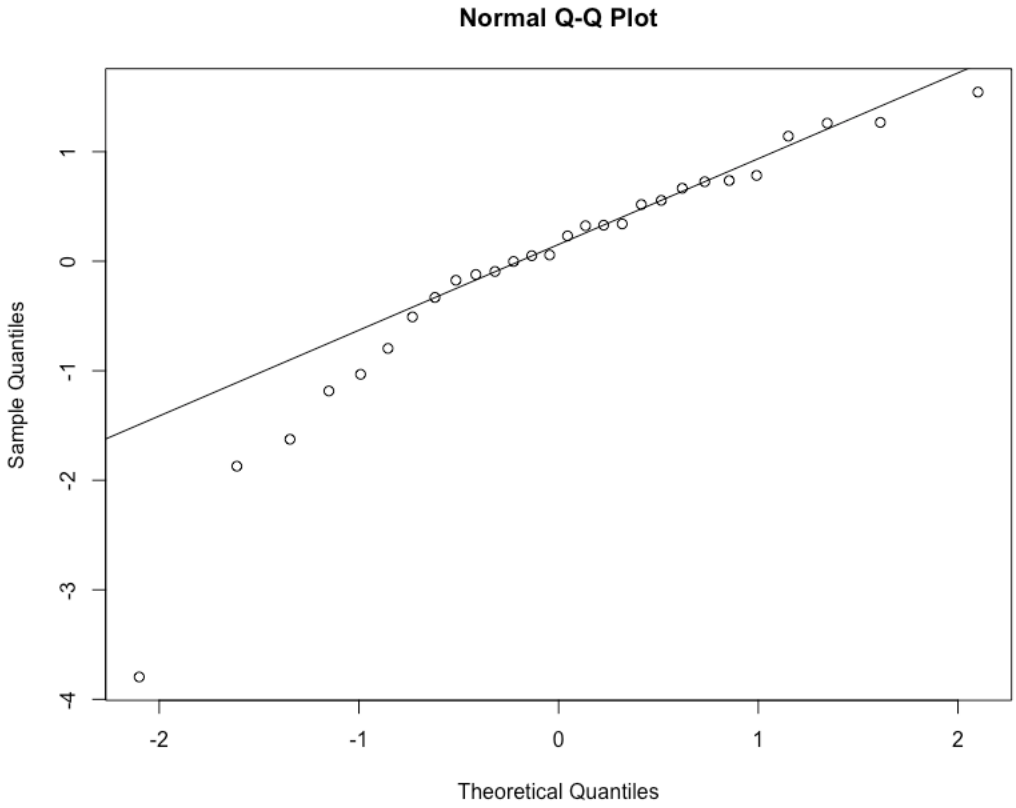

Rate of Return
Gross Operating Surplus
0.1
0.2
0.3
0.4
0.5
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07


Rate of Return
EBIT Margin
0.1
0.2
0.3
0.4
0.5
0.000
0.005
0.010
0.015
0.020
0.025
0.030


Studentized Residuals
Good Money
-4
-3
-2
-1
0
1
0.0
0.5
1.0
1.5
2.0
2.5


Studentized Residuals
Gross Operating Surplus
-4
-3
-2
-1
0
1
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07





The transformation looks like it helped our assumption of linearity. The scatterplots seemed to have improved some. However, the other assumptions mentioned earlier do not seem to have improved. To conclude, I think we should stick with the transformed model. The studentized residuals only produce one outlier, while the rest are below 2. This is good sign.

Conclusion

The regression analysis above revealed some interesting results, and unexpected results at that. I did not expect the issue arising regarding the computational singularity with R. This was an issue

that was finally fixed with a lot of different attempts. Regardless, I believe that our initial simple linear regression analysis with the one regressor EBIT Margin was a superior model to the optimal model found in the multiple linear regression analysis. All of the other regression just seemed to add 'noise' to the data. Sometimes a best model is not always the more complicated model. This project just goes to show that making economic prediction can be very difficult and unreliable at times.

References

[1] Lucian Buse, Mirela Ganea, and Daniel Circiumaru. Using linear regression in the analysis of financial-economic performances. Technical report, University of Craiova, Craiova, Romania, 2009.

[2] Montgomery Douglas C., Peck Elizabeth A., and Vining G. Geoffrey. *Introduction to Linear Regression Analysis*. John Wiley & Sons Inc., United States of America, fifth edition, 2012.

[3] R. Lyman Ott and Michael Longnecker. *An Introduction to Statistical Methods and Data Analysis*. Cancage Learning, Belmont, CA, sixth edition, 2010.