Analysis Report for:

Private Firm Asset Correlation

By:

Chengdao Yang

[redacted]

ADMS 4370: Data Analysis Systems

[redacted]

Before fitting the data through a model, we have done some preliminary data analysis in order to assess the data at hand. At first we will use graphical measures to see if there is any relationship between R-squared and Asset value. We initially use scatter plot for the correlation.

Referring to the Scatter Plot in Figure 1, it can be seen that Asset value and R-squared are positively linearly correlated. The slopes for Rating Categories 3 and 4 are very close to each other but are much lower than the other categories. The slopes for Rating Categories 1, 2, 5, 6, 7 and 8 are also positively linearly correlated but have a steeper slope compared to the other categories. The slopes of these Rating Categories are very similar to each other.

Figure 1: Scatter Plot of R-squared by Rating Category

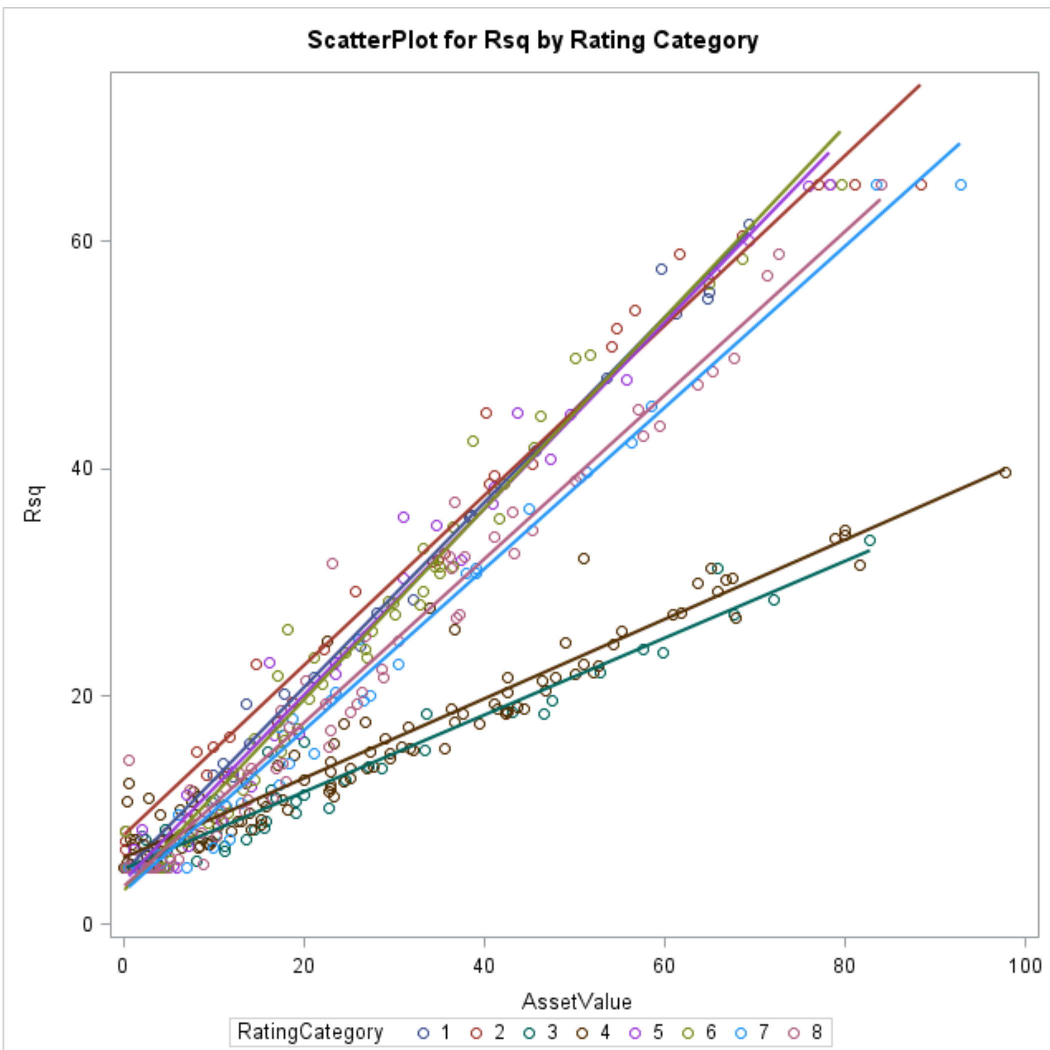
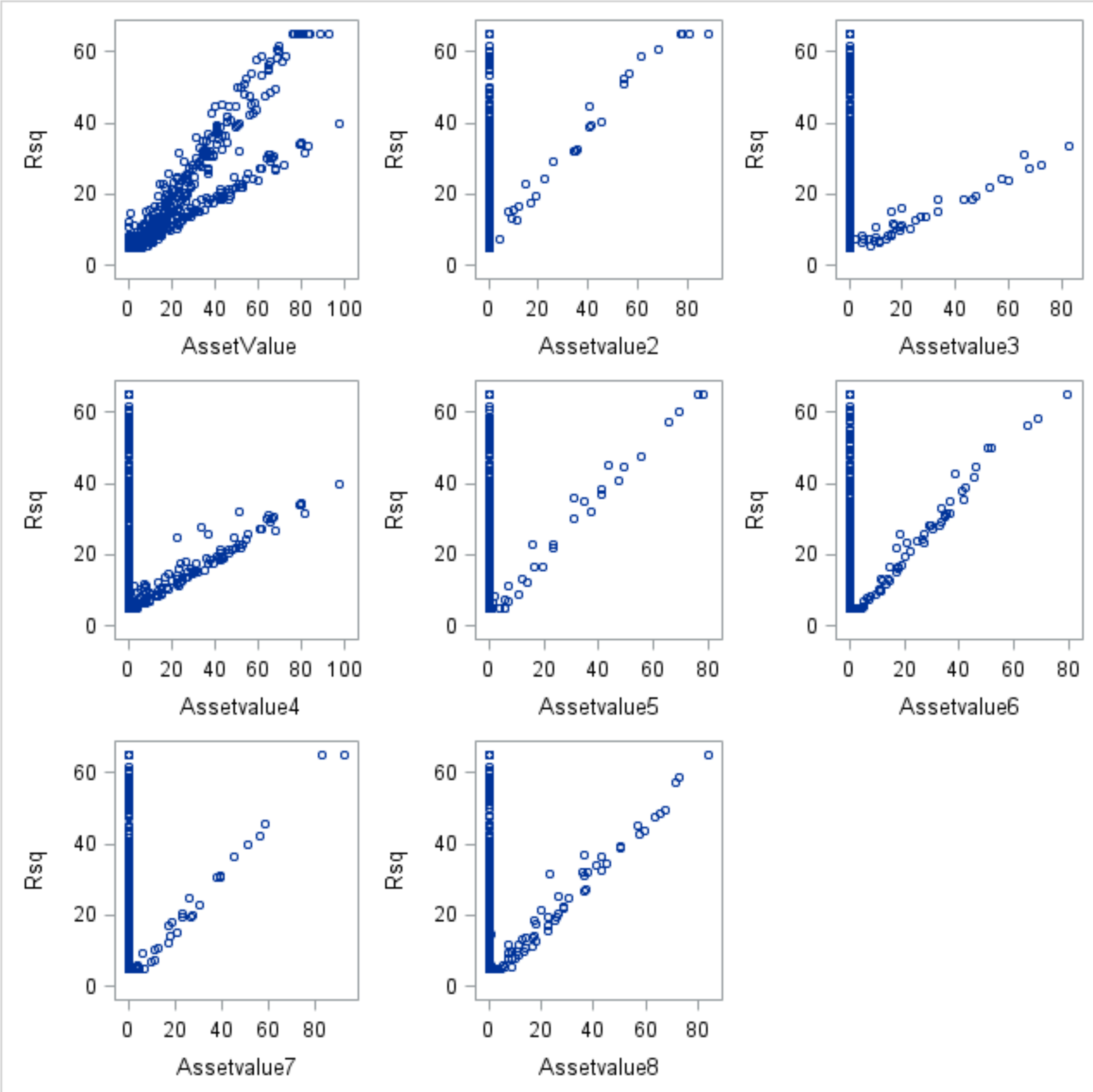


Figure 2: Scatter plot on each Rating Category



The figure above shows the same correlation but on separate scatter plots for each categories. It shows the correlation between each asset values through rating 1-8, taking rating category 1 as the baseline. From the graph it can be seen that all the R-squared has a linear relationship with their corresponding asset values, further illustrating a relationship between the two.

Next, we look at the boxplots of the numerical data to depict rating categories through their quartiles.

Figure 3: Distribution of Asset Value by Rating Category

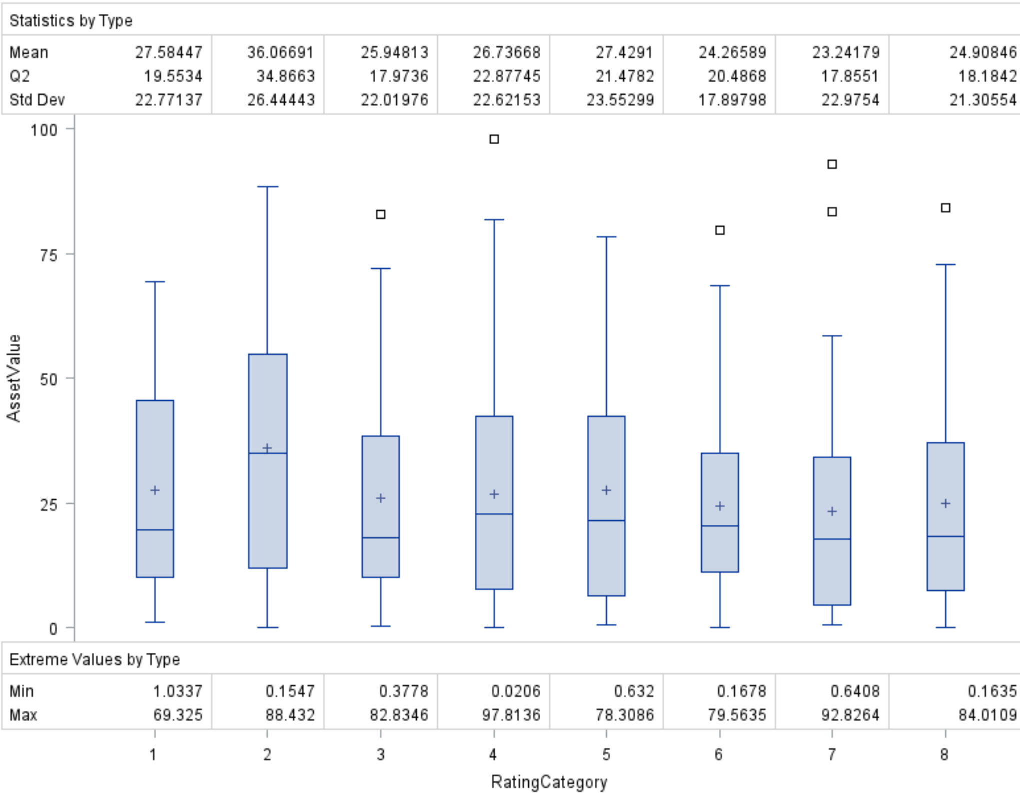


Figure 4: Distribution of R-Squared by Rating Category

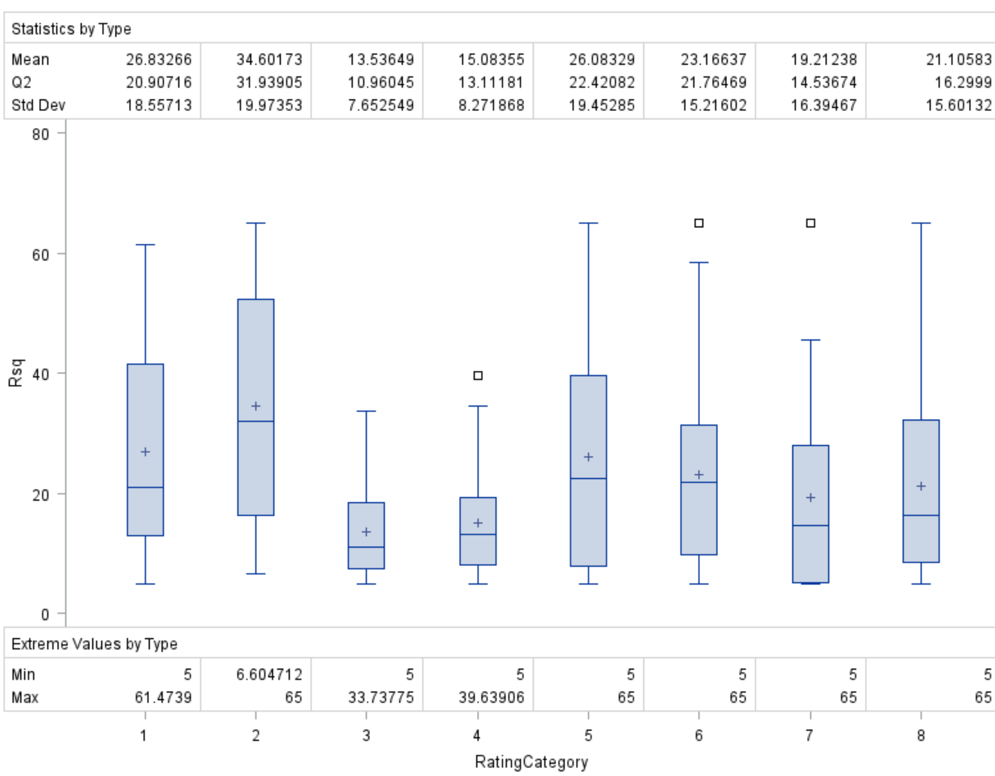


Figure 3 and 4 provide

box plots for Rating Category with Asset value and R-Squared. In figure 3, the outliers can also be seen for Rating Categories 3, 4, 6, 7 and 8. In figure 4, the outliers are seen in rating categories 4, 6, and 7. Rating Category 2 has the highest Asset value and R-Squared. Rating Categories 1 and 5 have similar summary statistics for Asset value with their means and standard deviations being almost equal. Rating Categories 3, 4, 6, 7 and 8 have similar summary statistics for Asset value. Rating Categories 3 and 4 have relatively low R-Squared values and they have similar summary statistics. For example, category 3 and 4 have very similar interquartile range. The same is true for category 7 and 8 as well. The spread of the distribution of Rating Categories is higher for R-Squared than for the box plots with Asset value.

Now that we have an idea of the spread of the data, we are ready to look at some non-graphic measures to further analyses the data. We use ANOVA procedure to test if R-squared is different for each category. From the figure below, since p-value is less than 0.05, we can reject our null hypothesis that R-squared is the same for each category. Therefore, now we know for sure that R-squared is different.

Figure 5: The ANOVA Procedure

Dependent Variable: Rsq

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	13697.64085	1956.80584	9.68	<.0001
Error	398	80480.72887	202.21289		
Corrected Total	405	94178.36972			

Source	DF	Anova SS	Mean Square	F Value	Pr > F
RatingCategory	7	13697.64085	1956.80584	9.68	<.0001

We are now ready to do a univariate test on y axis (R-squared) to test whether the distribution is normal or not.

Figure 6: The Univariate Procedure

Moments			
N	406	Sum Weights	406
Mean	20.4542094	Sum Observations	8304.40902
Std Deviation	15.2492355	Variance	232.539184
Skewness	1.22907189	Kurtosis	0.86076329
Uncorrected SS	264038.491	Corrected SS	94178.3697
Coeff Variation	74.553043	Std Error Mean	0.75680685

Basic Statistical Measures			
Location		Variability	
Mean	20.45421	Std Deviation	15.24924
Median	15.46419	Variance	232.53918
Mode	5.00000	Range	60.00000
		Interquartile Range	20.17597

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	65.00000
99%	65.00000
95%	55.51957
90%	42.79939
75% Q3	28.43176
50% Median	15.46419
25% Q1	8.25580
10%	5.21953
5%	5.00000
1%	5.00000
0% Min	5.00000

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	t	27.02699	Pr > |t|	<.0001
Sign	M	203	Pr >= |M|	<.0001
Signed Rank	S	41310.5	Pr >= |S|	<.0001

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.861008	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.155425	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.708029	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	16.72527	Pr > A-Sq	<0.0050

The tables above give us the 5 number summary for the distribution. The location test shows that the mean of the distribution is different from zero as p-value is less than 0.05. The test of normality shows that the distribution is not normal. Both the Kolmogorov-Smirnov and the Anderson-Darling fails the normality test. Figure 7 and 8 below shows a graphical representative of the distribution which, clearly, is not normal.

Figure 7. Histogram for R-squared

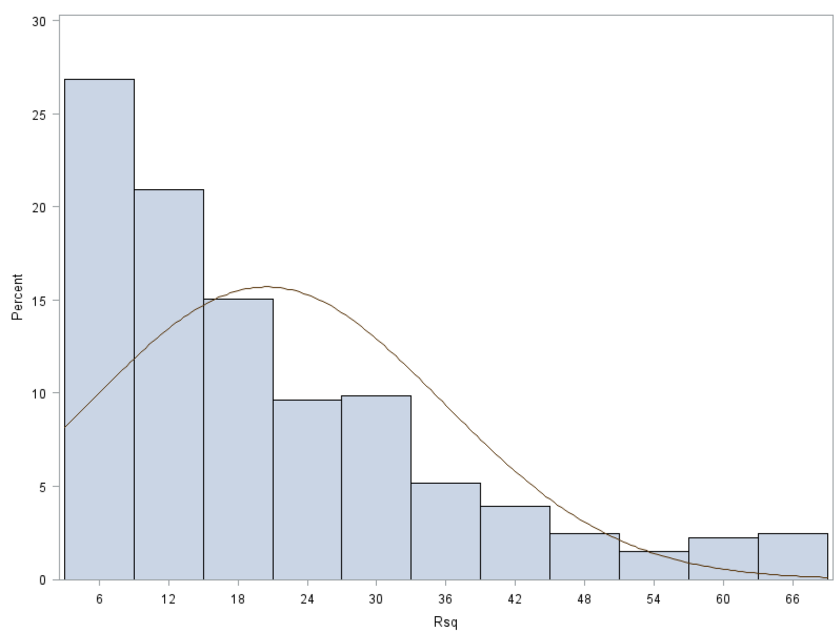
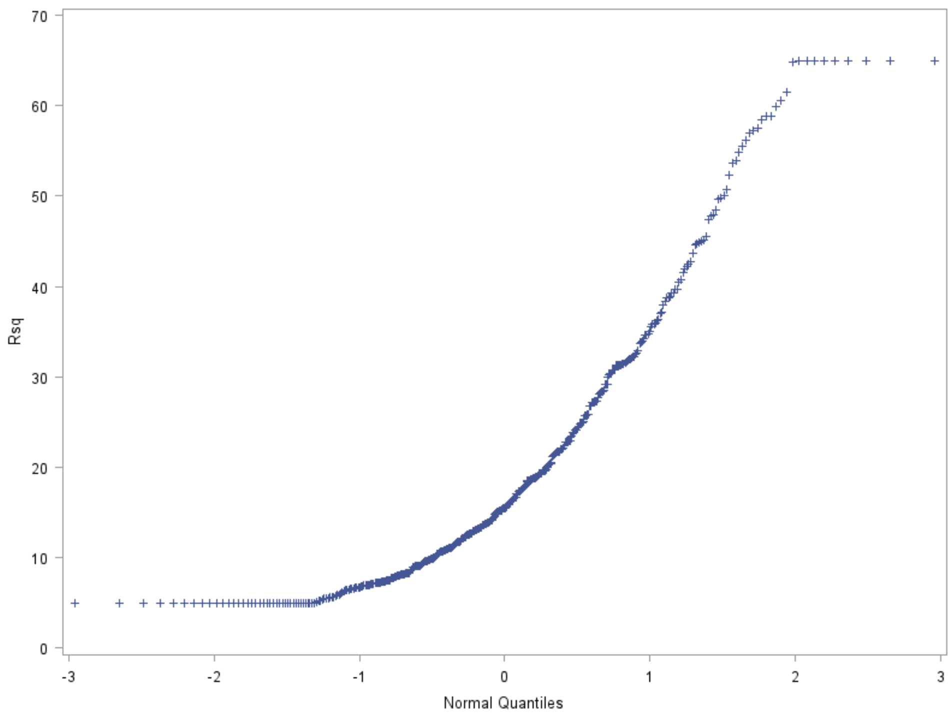


Figure 8: QQ plot for R-squared distribution



Finally, for our data analysis section, we look at correlation between all the variables, asset value, rating category and R-squared using Pearson and Spearmen test.

Figure 9: Pearson and Spearman Correlation

Pearson Correlation Coefficients, N = 406 Prob > |r| under H0: Rho=0			
	AssetValue	RatingCategory	Rsq
AssetValue	1.00000	-0.08642 0.0820	0.88026 <.0001
RatingCategory	-0.08642 0.0820	1.00000	-0.03953 0.4270
Rsq	0.88026 <.0001	-0.03953 0.4270	1.00000

Spearman Correlation Coefficients, N = 406 Prob > |r| under H0: Rho=0			
	AssetValue	RatingCategory	Rsq
AssetValue	1.00000	-0.07402 0.1365	0.91814 <.0001
RatingCategory	-0.07402 0.1365	1.00000	-0.02907 0.5592
Rsq	0.91814 <.0001	-0.02907 0.5592	1.00000

The p-value is significantly lower than 0.05 for R-squared in both the Pearson and Spearman Correlation analysis. R-Squared is highly correlated with Asset value with correlations of 88% and 91.8% for the Pearson and Spearman correlation, respectively. However, looking at the p-value R-Squared vs. Rating Category, they are higher than 0.05, making them statistically insignificant. R-Squared and Rating Categories do not have a significant correlation.

Having done a preliminary data analysis we can now move on to fit a linear model. At first, we fit a model taking R-squared Y (dependent) variable. For our x (independent) variables, we take asset values, rating category and an interaction term between asset value and rating category. For the purpose of this particular model, we treat rating category as a continuous variable.

Model:

R-squared = $\beta_0 + \beta_1$*AssetValues + β_2*RatingCategory + β_3*(AssetValues*RatingCategory)

Figure 10: Regression parameter estimates

Parameter Estimates										
						Heteroscedasticity Consistent				
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	5.36560	1.43068	3.75	0.0002	1.14887	4.67	<.0001	.	0
AssetValue	1	0.52434	0.03916	13.39	<.0001	0.06305	8.32	<.0001	0.16956	5.89768
RatingCategory	1	-0.21154	0.26784	-0.79	0.4301	0.18123	-1.17	0.2438	0.40960	2.44142
AR	1	0.01794	0.00766	2.34	0.0196	0.00986	1.82	0.0695	0.14564	6.86637

Fitting the parameters into the model we get,

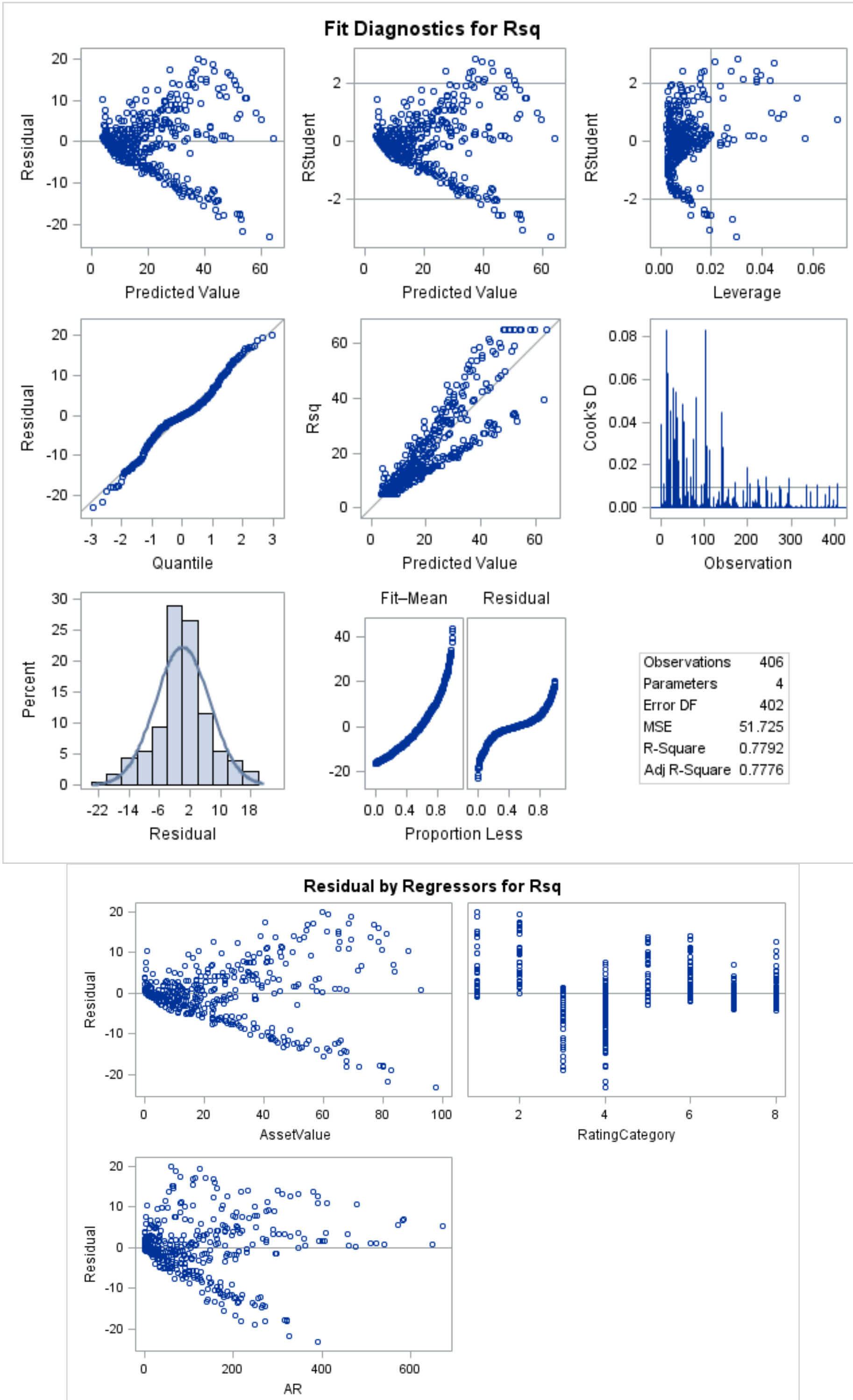
R-squared = 5.366 + 0.524AssetValues – 0.211RatingCategory + 0.018(AssetValues*RatingCategory)

Looking at the test of significance, we see that asset-value has a p-value of less than 0.05 therefore this variable is significant. However, the other two variables are not significant in this test with a p-value greater than 0.05.

To further assess this model, we look at the diagnostic of the residuals.

In figure 11, from the first diagram of residuals verses predicted value, we see that the error variance is not constant. The residuals seem to value with the level of predicted value. Rstudent vs. predicted value also shows similar patterns. This shows that this model is a bad fit for the data. Not all residuals are explained by the model. The leverage vs. Rstudent graph displays a lot of outliers which may be influential in the model. Rstudent vs. quantile falls roughly in a normal distribution. R-squared vs. predicted value shows clusters of three lines along with quite a few outliers. The cook's D observation shows a lot of outliers as well. Percentage vs. residual is a uni-modal, roughly symmetric diagram. Lastly, the fit-mean vs. proportion less shows that the spread in residual is not greater than the spread in centered fit. Overall, even though the model explains some variability in the data, the model is not a good fit because it does not provide a constant variability in residuals.

Figure 11: Diagnostics of Residuals



In order to find a better fitted model, we drop the interaction term asset values and rating category because of multicollinearity. We hope to decrease the variability in the residual by omitting the insignificant variables from the model one by one.

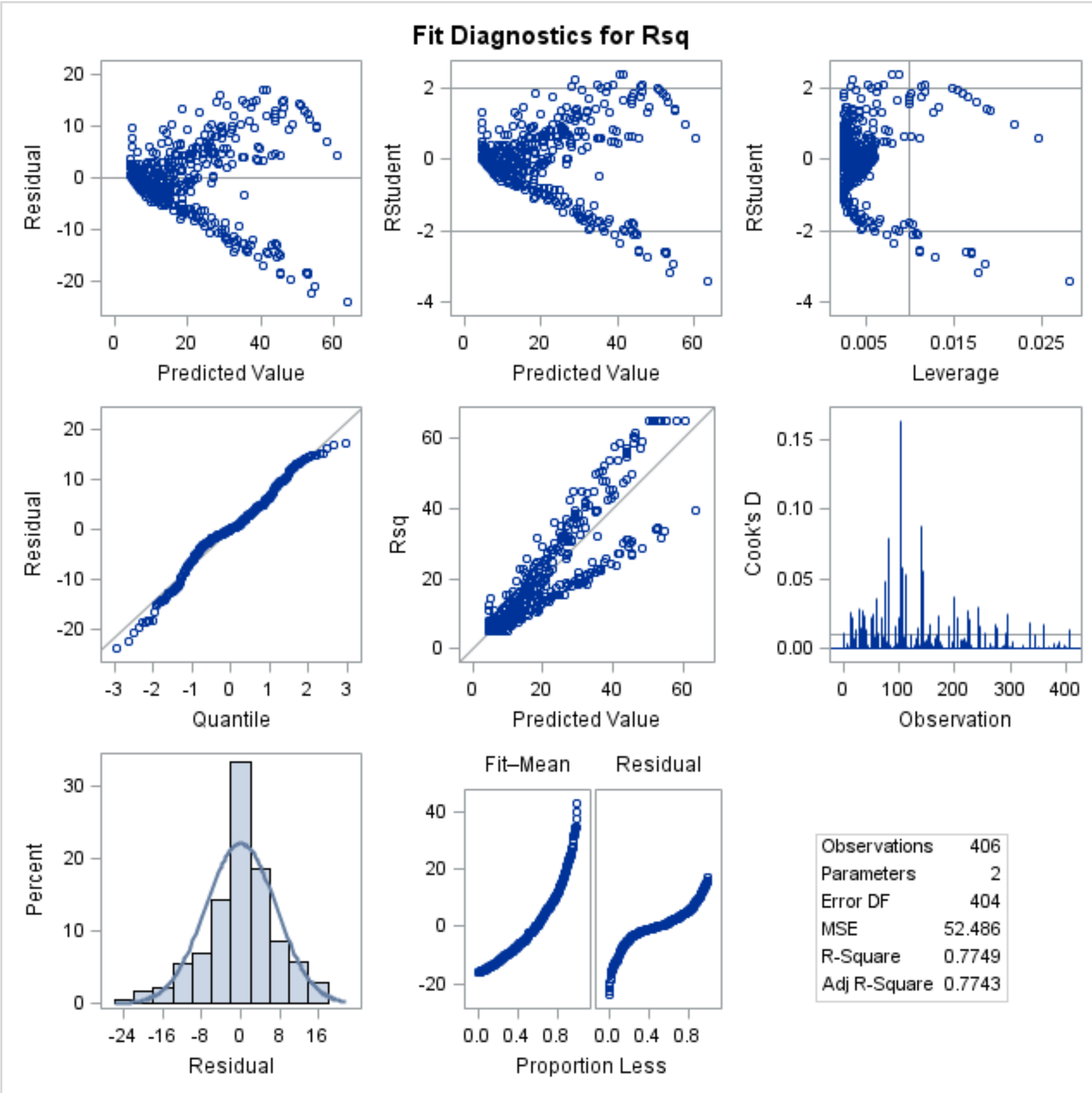
Figure 11: Parameter Estimates and Residual diagnostics

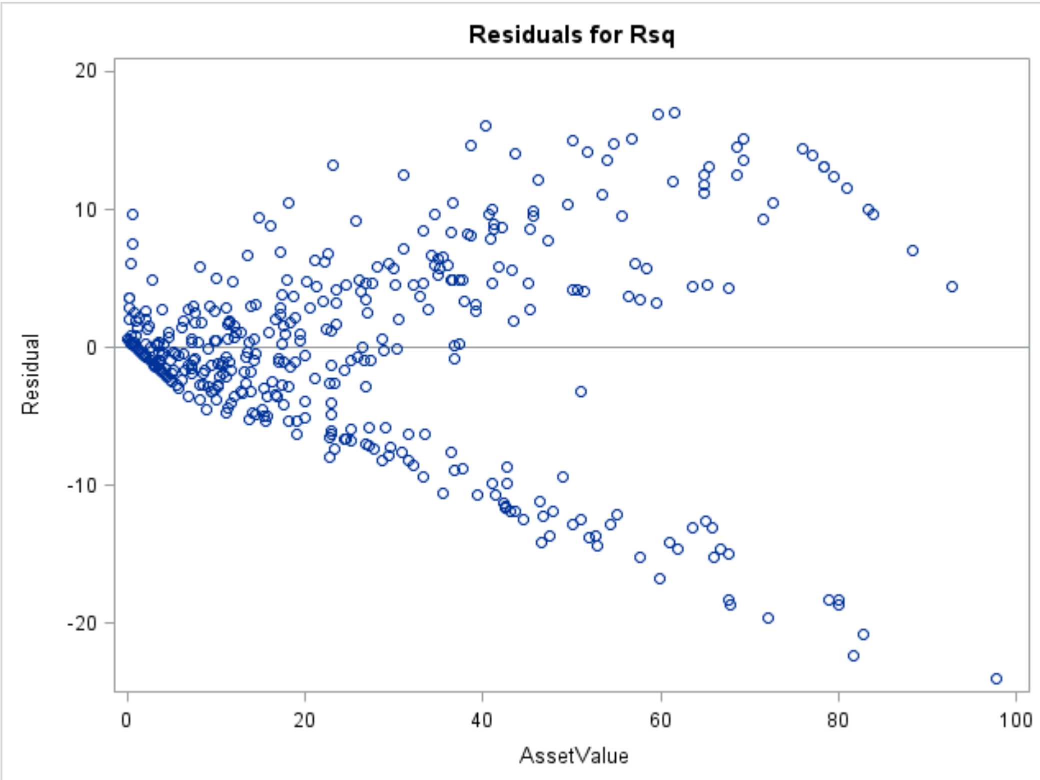
Parameter Estimates										
						Heteroscedasticity Consistent				
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	3.05366	1.04152	2.93	0.0036	0.83261	3.67	0.0003	.	0
AssetValue	1	0.60787	0.01628	37.35	<.0001	0.02444	24.87	<.0001	0.99253	1.00752
RatingCategory	1	0.26931	0.17302	1.56	0.1204	0.15433	1.75	0.0817	0.99253	1.00752

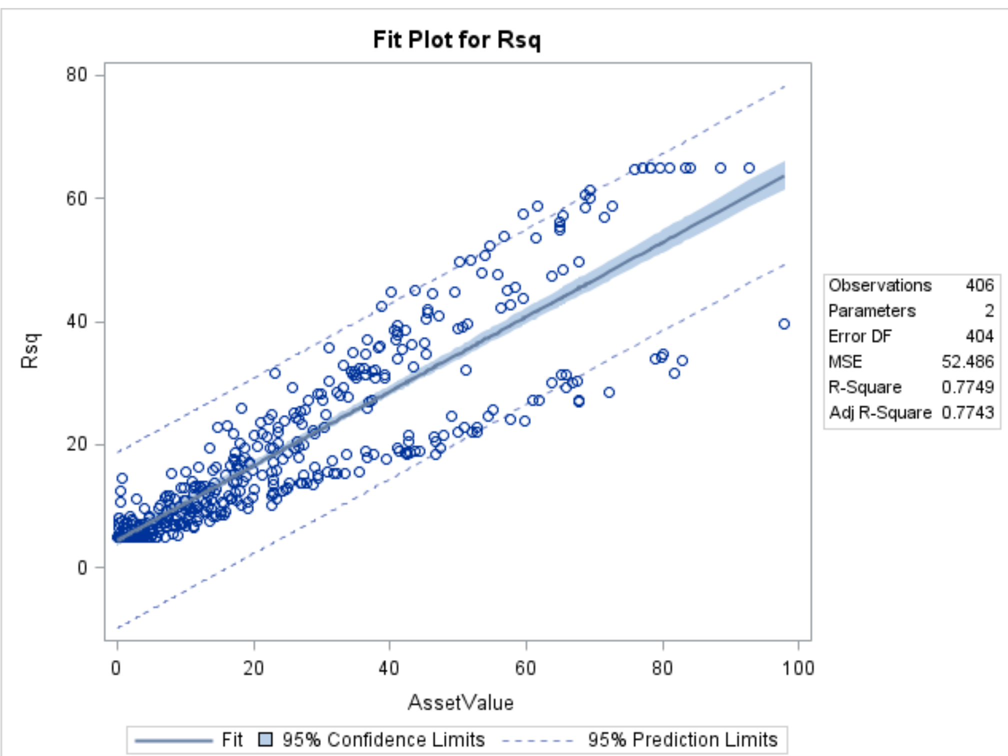
Since the variable rating category is still insignificant in the model as p-value is greater than 0.05, we also drop it. The final model is as follows:

Figure 12: Parameter Estimates and Residual diagnostics

Parameter Estimates								
						Heteroscedasticity Consistent		
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|
Intercept	1	4.42102	0.56050	7.89	<.0001	0.44438	9.95	<.0001
AssetValue	1	0.60568	0.01624	37.29	<.0001	0.02494	24.28	<.0001


Fit Diagnostics for Rsq
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
Rsq
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 406
Parameters 2
Error DF 404
MSE 52.486
R-Square 0.7749
Adj R-Square 0.7743


Residuals for Rsq
Residual
AssetValue
20
10
0
-10
-20
0
20
40
60
80
100


Fit Plot for Rsq
Rsq
AssetValue
Observations 406
Parameters 2
Error DF 404
MSE 52.486
R-Square 0.7749
Adj R-Square 0.7743
Fit
95% Confidence Limits
95% Prediction Limits

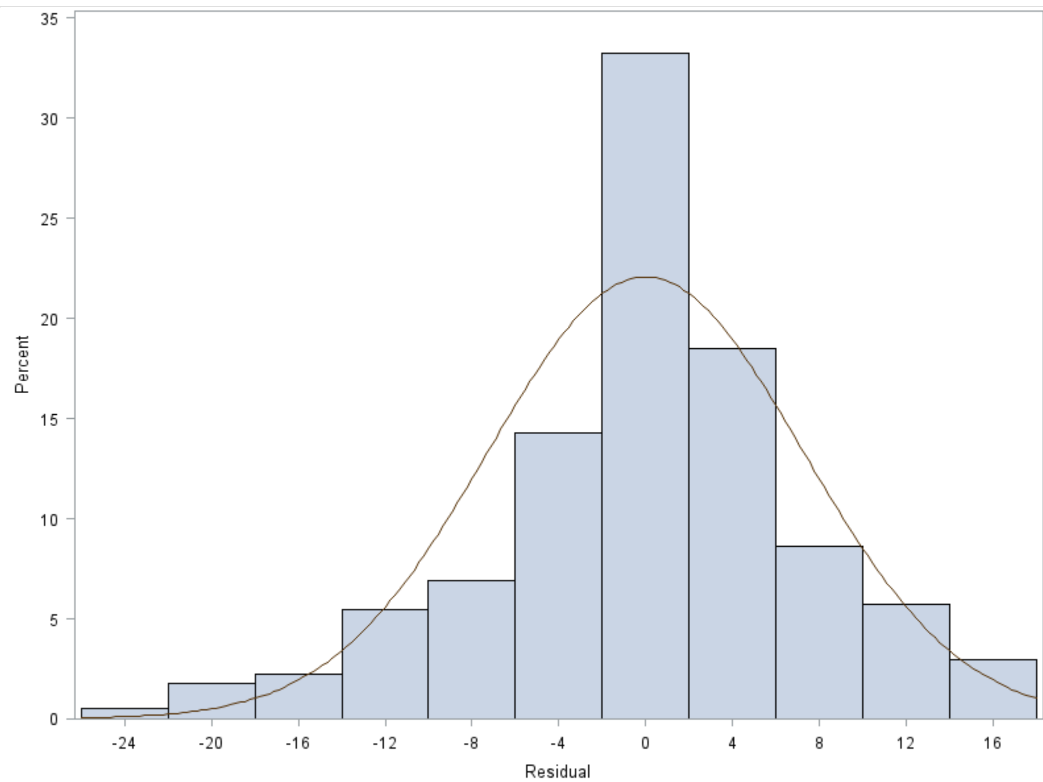
R-squared = 4.421 + 0.606AssetValues

Looking at the diagnostics residuals, the variations seem to have improved a little but the model is still not a good enough fit for the data. The error values are still roughly inconsistent.

Next we do a normality test to further assess the model. According to the data and graph below, this model also fails the normality test.

Figure 13: Test for normality

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.977508	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.088769	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.635822	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.397963	Pr > A-Sq	<0.0050



Then, we assume the residuals are independent, we carryout a correlation test to show the independence and then regress asset values over the residuals.

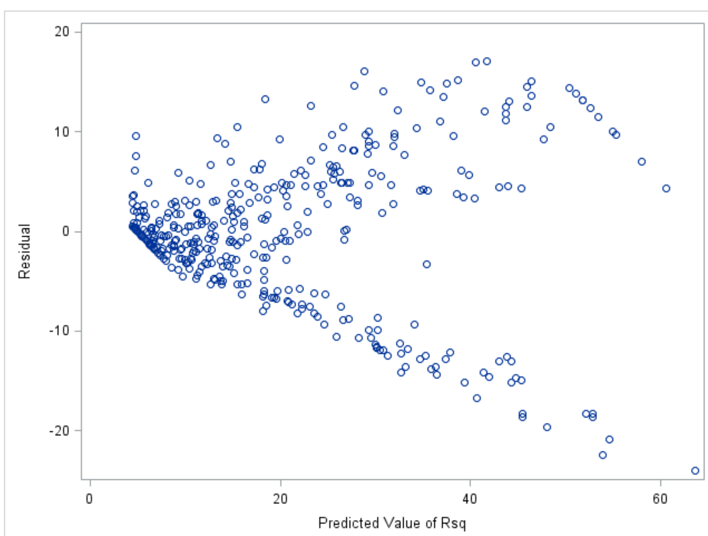
Figure 14. Correlation test and parameters of estimates

Pearson Correlation Coefficients, N = 406 Prob > |r| under H0: Rho=0		
	AssetValue	residual
AssetValue	1.00000	0.00000
		1.0000
residual	0.00000	1.00000
Residual	1.0000	

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	26.47132	1.10125	24.04	<.0001
residual	Residual	1	3.86945E-15	0.15238	0.00	1.0000

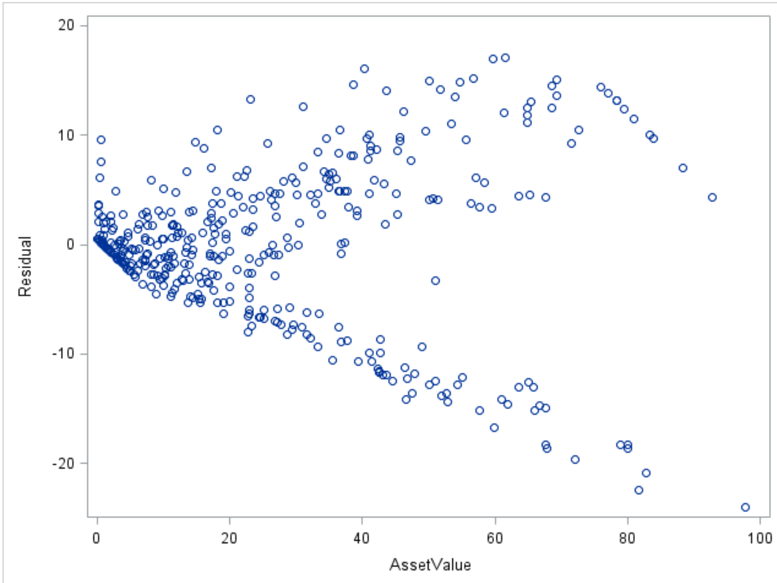
To show constant variance in residual we use scatterplot

Figure 15: Residual vs. predicted value R-squared scatter plot



To show residual linearity, we plot residual vs. asset value:

Figure 16: Residual vs. asset value scatter plot

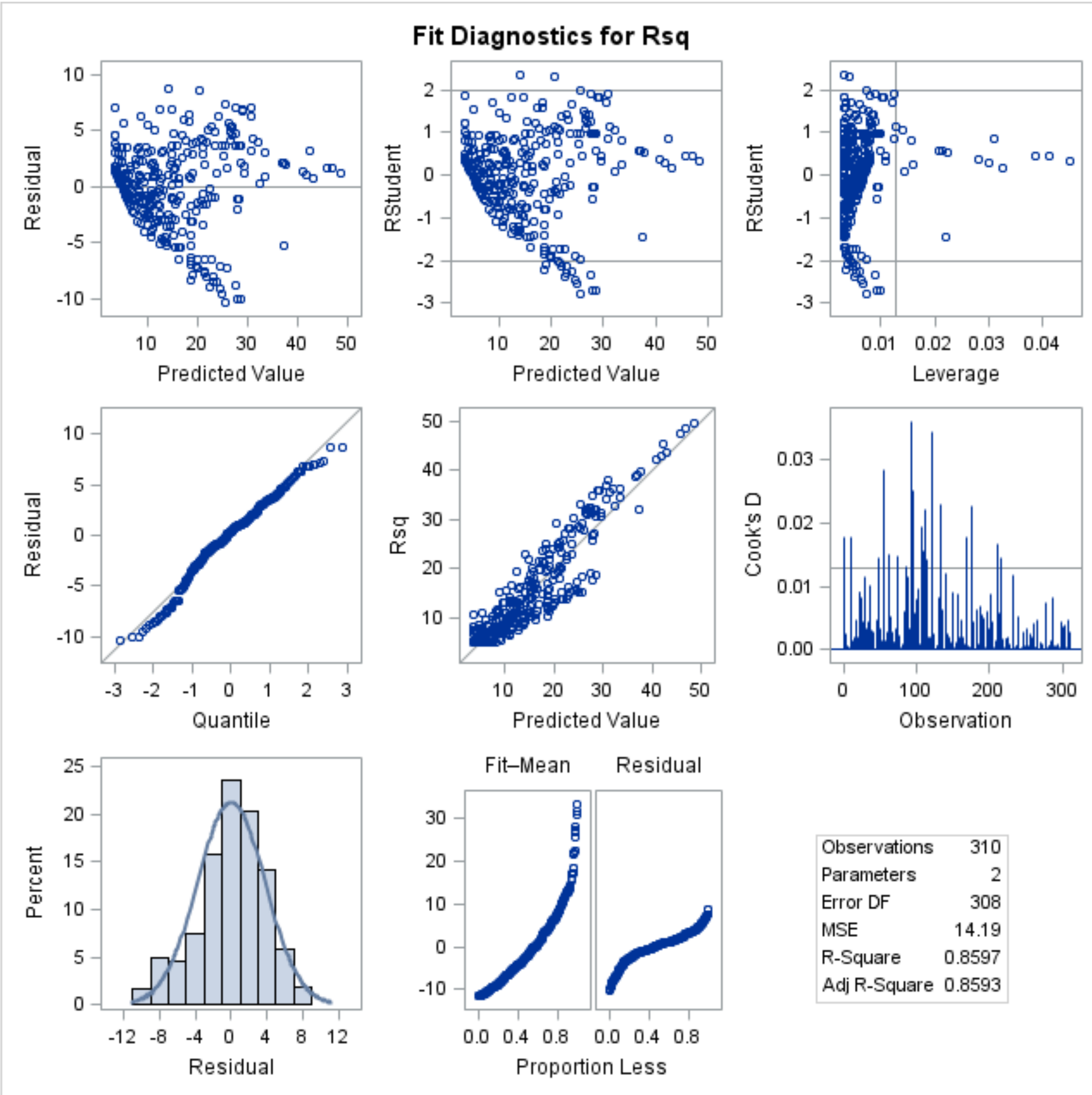


To see if the outliers are influential, we regress the model again after deleting the outliers using Cooks'D.

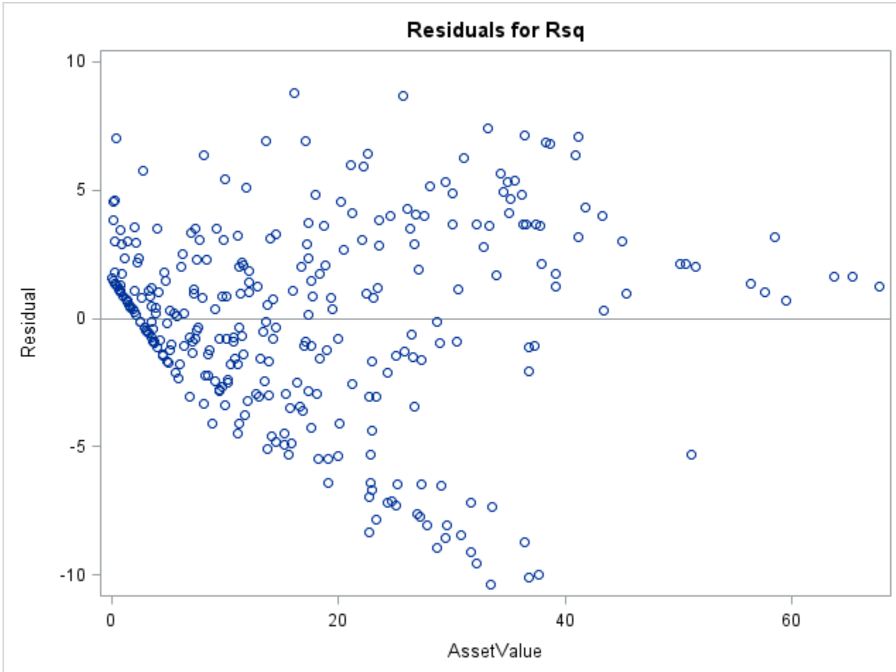
Figure 17. Parameters of Estimates and diagnostic residuals

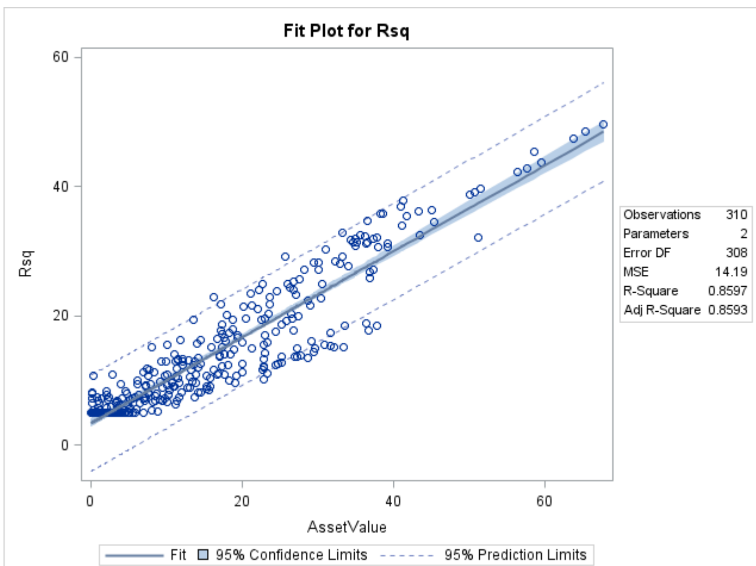
Parameter Estimates								
						Heteroscedasticity Consistent		
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|
Intercept	1	3.43184	0.34227	10.03	<.0001	0.22597	15.19	<.0001
AssetValue	1	0.66553	0.01532	43.45	<.0001	0.01442	46.15	<.0001

R-squared = 3.431 + 0.665AssetValue



Fit Diagnostics for Rsq
Residual
Predicted Value
RStudent
Leverage
Quantile
Rsq
Cook's D
Observation
Percent
Fit–Mean
Residual
Proportion Less
Observations 310
Parameters 2
Error DF 308
MSE 14.19
R-Square 0.8597
Adj R-Square 0.8593





Performing a normality test on the data after deleting the outliers, we see that the distribution is now normal.

Figure 18: Test for Normality

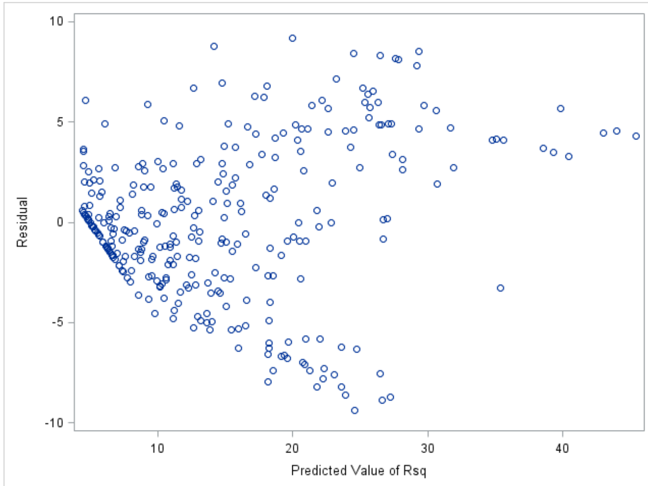
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.991398	Pr < W	0.0678
Kolmogorov-Smirnov	D	0.045734	Pr > D	0.1137
Cramer-von Mises	W-Sq	0.107508	Pr > W-Sq	0.0916
Anderson-Darling	A-Sq	0.666403	Pr > A-Sq	0.0847

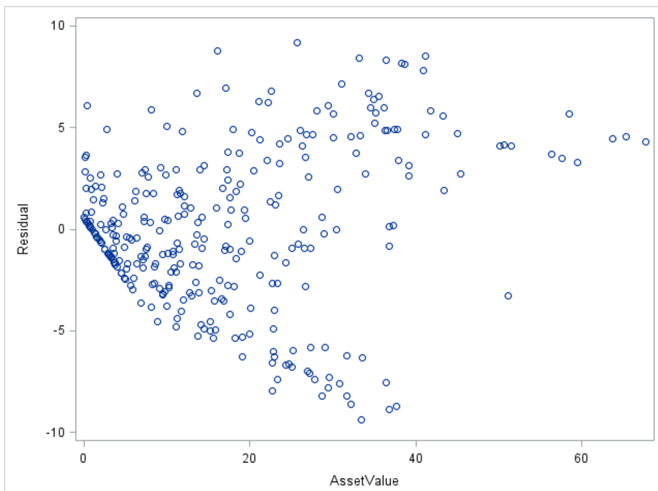
Carrying out the white test again for homoscedasticity, we get the following results:

Figure 19: White test

Pearson Correlation Coefficients, N = 310 Prob > |r| under H0: Rho=0		
	AssetValue	residual
AssetValue	1.00000	0.21731
		0.0001
residual	0.21731	1.00000
Residual	0.0001	

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	17.39938	0.77699	22.39	<.0001
residual	Residual	1	0.78911	0.20197	3.91	0.0001





There seems to be a constant variance now. Therefore without the outliers, the model is a good fit for the data. However, we prefer not to delete data from the sample. Instead we use a robust regression model to mitigate for the effects of the outliers. We base the robust regression on the model before deleting the outliers. Therefore our model is:

R-squared = $\beta_0 + \beta_1$*AssetValues

Figure 20: Robust regression parameters

Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Intercept	1	2.6057	0.3839	1.8534	3.3581	46.08	<.0001
AssetValue	1	0.7605	0.0111	0.7387	0.7823	4673.79	<.0001
Scale	1	4.3889					

R-squared = 2.606 + 0.761AssetValues

Compared to our model previously, R-squared = 4.421 + 0.606AssetValues, the robust model has a much lower intercept and a higher coefficient for asset values. The parameters of estimates do change drastically for the intercept but not so much for the coefficient of asset values. Furthermore, if we compare the parameters with that of the deleted outlier sample (R-squared = 3.431 + 0.665AssetValue), we see same result that the robust model has a lower intercept and a higher coefficient for asset values still compared to the deleted outlier sample.

In order to show association among asset value and R-squared, we now hold rating category as categorical values instead of as continuous variables as we did before. For this we create dummy variables for each category and use Rating Category 1 as baseline. Next we find the interaction terms for asset value and each category dummy variable. We put all that into one model and regress it using stepwise selection. The model we put in initially is as follows. Let Rating Category 1 be A, Rating Category 2 be B and so on. The interaction between Rating Category 1 and asset values is Assetvalues1, interaction between Rating Category 2 and asset values is Assetvalues2 and so forth. Therefore, we have the following model:

R-squared = β_0 + β_1Assetvalue + β_2B + β_3C + β_4D + β_5E + β_6F + β_7G + β_8H + β_9Assetvalue2 + β_{10}Assetvalue3 + β_{11}Assetvalue4+ β_{12}Assetvalue5 + β_{13}Assetvalue6 + β_{14}Assetvalue7 + β_{15}Assetvalue8.

The following model summary is generated by using the stepwise selection.

Figure 21: Stepwise selection regression

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	AssetValue		1	0.7458	0.7458	5028.21	983.01	<.0001
2	Assetvalue4		2	0.1380	0.8839	2118.74	396.95	<.0001
3	Assetvalue3		3	0.0855	0.9694	317.357	929.09	<.0001
4	Assetvalue8		4	0.0054	0.9747	206.479	70.25	<.0001
5	Assetvalue7		5	0.0056	0.9803	89.5645	94.94	<.0001
6	D		6	0.0022	0.9826	44.9711	41.79	<.0001
7	B		7	0.0006	0.9831	34.8305	11.23	0.0009
8	Assetvalue2		8	0.0005	0.9836	26.2783	10.02	0.0017
9	C		9	0.0004	0.9840	20.2155	7.82	0.0055

For assets under the rating category "7" and "8", SAS procedure has deleted the dummy variables but kept the interaction terms. Although the dummy variables in rating category 7 and 8 are not statistically significant, we think we should keep them due to the significance of their interaction terms. So we decide to drop one variable at a time to find the best model. The partial F-tests shows that assets under the rating category "5" and "6" and the interaction terms between them and the asset value are not statistically significant.

Our steps are as follows:

1. First run the regression model with all the dummy variables and the interaction terms.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.43542	0.78470	5.65	<.0001
AssetValue	1	0.81195	0.02211	36.72	<.0001
B	1	3.26570	1.12042	2.91	0.0038
C	1	0.33516	0.99993	0.34	0.7377
D	1	1.31934	0.86321	1.53	0.1272
E	1	-0.78659	1.04455	-0.75	0.4519
F	1	-1.62788	0.96694	-1.68	0.0931
G	1	-1.68814	0.98849	-1.71	0.0885
H	1	-1.23699	0.91556	-1.35	0.1775
Assetvalue2	1	-0.06610	0.02851	-2.32	0.0209
Assetvalue3	1	-0.47413	0.02871	-16.52	<.0001
Assetvalue4	1	-0.46304	0.02439	-18.99	<.0001
Assetvalue5	1	0.00596	0.02928	0.20	0.8389
Assetvalue6	1	0.02704	0.02902	0.93	0.3521
Assetvalue7	1	-0.10352	0.02884	-3.59	0.0004
Assetvalue8	1	-0.09302	0.02641	-3.52	0.0005

2. Test the significance of slopes for overall asset value terms

   H0: The slopes for Asset Value, Assetvalue2, Assetvalue3, Assetvalue4, Assetvalue5, Assetvalue6, Assetvalue7, Assetvalue8 ALL EQUAL 0

   Ha: at least one of the slope not equals 0

Test testnosig Results for Dependent Variable Rsq				
Source	DF	Mean Square	F Value	Pr > F
Numerator	8	9751.12701	1538.58	<.0001
Denominator	390	6.33773		

   Decision: P-value<0, reject H0

   Conclusion: The slopes for overall asset value terms are statistically significant.

3. Test the significance of all the interaction terms.

   H0: The slopes for Assetvalue2, Assetvalue3, Assetvalue4, Assetvalue5, Assetvalue6, Assetvalue7, Assetvalue8 ALL EQUAL 0

   Ha: at least one of the slope not equals 0

Test testnosig Results for Dependent Variable Rsq				
Source	DF	Mean Square	F Value	Pr > F
Numerator	7	1222.73148	192.93	<.0001
Denominator	390	6.33773		

   P-value<0 indicates that overall, the slopes of all the interaction terms are statistically significant.

4. Test if the asset value effect is the same for all interaction terms.

Test testsame Results for Dependent Variable Rsq				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	475.73473	75.06	<.0001

Test testsame Results for Dependent Variable Rsq				
Source	DF	Mean Square	F Value	Pr > F
Denominator	390	6.33773		

P-value<0 indicates that overall, the asset value effects are not the same for all interaction terms.

5. Test if assets under category 5 and its interaction term with asset value is significant.

Test testrc5 Results for Dependent Variable Rsq				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	2.91206	0.46	0.6320
Denominator	390	6.33773		

P-value>0 indicates that assets under category 5 and its interaction term with asset value are not statistically significant. Thus, we can drop them.

6. Test the model again without category 5 and the interaction term.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.99693	0.51714	7.73	<.0001
AssetValue	1	0.81539	0.01448	56.33	<.0001
B	1	3.70420	0.95145	3.89	0.0001
C	1	0.77366	0.80652	0.96	0.3380
D	1	1.75784	0.62964	2.79	0.0055
F	1	-1.18938	0.76536	-1.55	0.1210
G	1	-1.24964	0.79233	-1.58	0.1156
H	1	-0.79849	0.69951	-1.14	0.2544
Assetvalue2	1	-0.06954	0.02307	-3.01	0.0027

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Assetvalue3	1	-0.47757	0.02332	-20.48	<.0001
Assetvalue4	1	-0.46648	0.01775	-26.28	<.0001
Assetvalue6	1	0.02360	0.02370	1.00	0.3201
Assetvalue7	1	-0.10697	0.02349	-4.55	<.0001
Assetvalue8	1	-0.09646	0.02043	-4.72	<.0001

Category 6 and its interaction term with the asset value raise some cautions so we will test the significance of these two variables.

P-value>0 indicates that overall, the asset under the Category 6 and its interaction term with the asset value are not statistically significant. Thus, we can drop these two variables.

Test testrc6 Results for Dependent Variable Rsq				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	8.07615	1.28	0.2798
Denominator	392	6.32025		

7. This is the model without category 6 and its interaction model.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.43406	0.37780	9.09	<.0001
AssetValue	1	0.82532	0.01144	72.17	<.0001
B	1	4.26706	0.88400	4.83	<.0001
C	1	1.33652	0.72548	1.84	0.0662
D	1	2.32070	0.52147	4.45	<.0001
G	1	-0.68677	0.70964	-0.97	0.3338
H	1	-0.23562	0.60409	-0.39	0.6967
Assetvalue2	1	-0.07947	0.02131	-3.73	0.0002
Assetvalue3	1	-0.48750	0.02157	-22.60	<.0001
Assetvalue4	1	-0.47641	0.01538	-30.98	<.0001
Assetvalue7	1	-0.11690	0.02176	-5.37	<.0001
Assetvalue8	1	-0.10639	0.01841	-5.78	<.0001

B, G and H’s p-values >0.05 indicates that they may not be statistically significant so we are going to test their significance first.

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable Rsq				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	2	257.79095	40.73	<.0001
Denominator	394	6.32916		

The REG Procedure
Model: MODEL1

Test 2 Results for Dependent Variable Rsq				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	2	288.57894	45.60	<.0001
Denominator	394	6.32916		

The REG Procedure
Model: MODEL1

Test 3 Results for Dependent Variable Rsq				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	2	3512.53576	554.98	<.0001
Denominator	394	6.32916		

For all three tests, the p-value< 0.05.It means assets under category 2,7 and 8 and their interaction terms are significant. Therefore, we have to keep them.

In order to simplify the model further and taking into consideration of the association showed in the box plots, we decide to test the similarity between group 7 and 8, 3 and 4.

The REG Procedure
Model: MODEL1

Test 4 Results for Dependent Variable Rsq				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	1	1.26758	0.20	0.6547
Denominator	394	6.32916		

The REG Procedure
Model: MODEL1

Test 5 Results for Dependent Variable Rsq				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	2.20941	0.35	0.5550
Denominator	394	6.32916		

For both tests, P-value>0.05,failed to reject H0.It means the asset value effect and the category effects are the same for both categories. So we can group them into one group.

The REG Procedure
Model: MODEL1

Test 6 Results for Dependent Variable Rsq				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	1.76729	0.28	0.5975
Denominator	394	6.32916		

The REG Procedure
Model: MODEL1

Test 7 Results for Dependent Variable Rsq				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	11.95515	1.89	0.1701
Denominator	394	6.32916		

For both tests, P-value>0.05,failed to reject H0.It means the asset value effect and the category effects are the same for both categories. So group 3 and 4 can be put in one group.

8. Now test the model again.

This is the model without deleting the outliers:

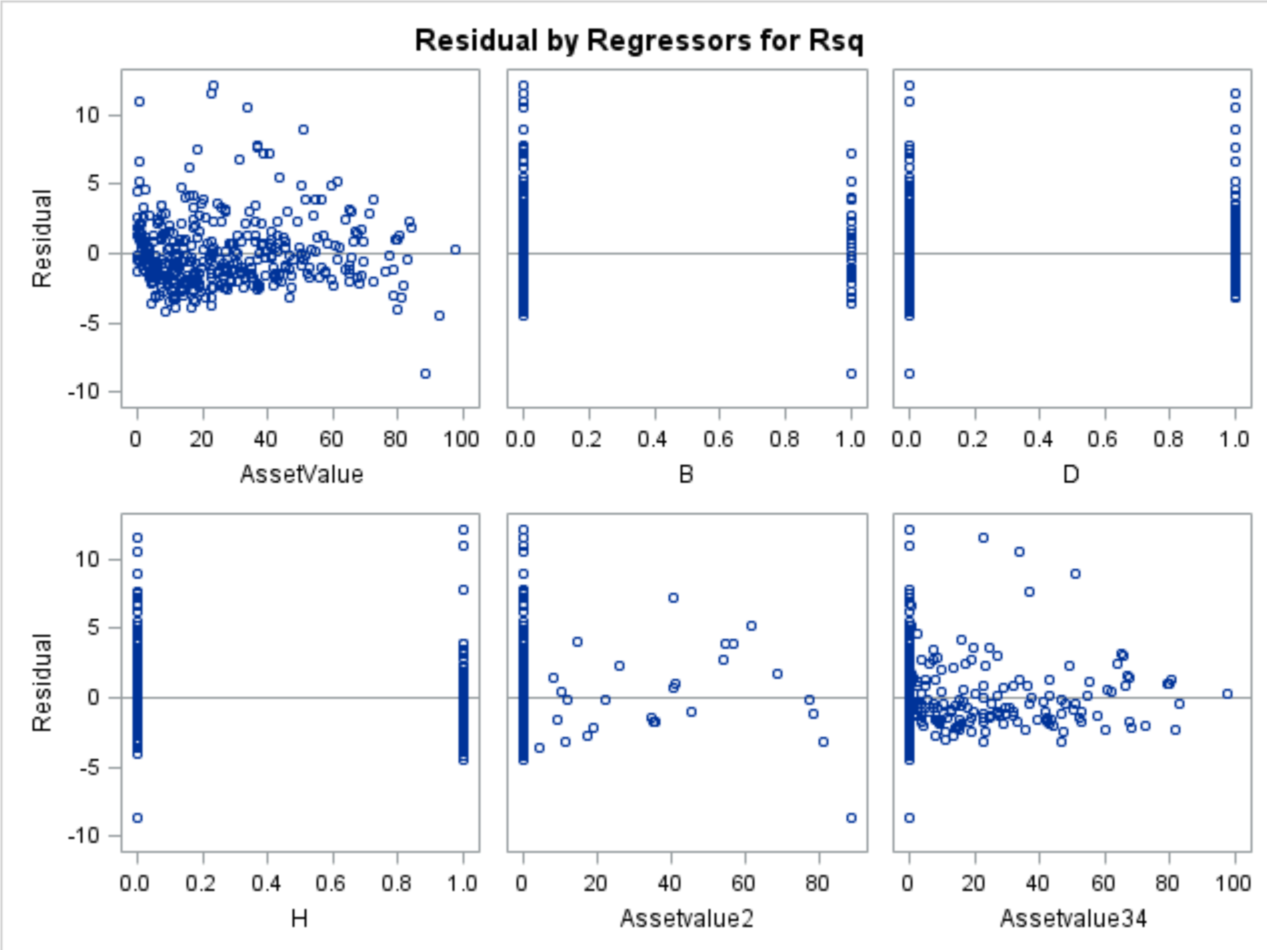
Parameter Estimates										
						Heteroscedasticity Consistent				
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	3.43406	0.38062	9.02	<.0001	0.33497	10.25	<.0001	.	0
AssetValue	1	0.82532	0.01152	71.64	<.0001	0.01050	78.61	<.0001	0.24331	4.10993
B	1	4.26706	0.89060	4.79	<.0001	0.93305	4.57	<.0001	0.30077	3.32483

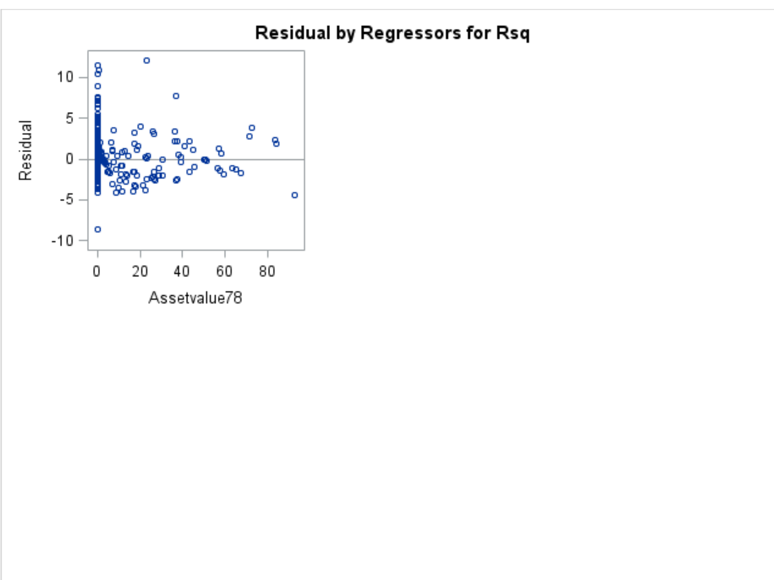
Parameter Estimates										
						Heteroscedasticity Consistent				
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
D	1	2.05922	0.49289	4.18	<.0001	0.42562	4.84	<.0001	0.27399	3.64982
H	1	-0.39340	0.53323	-0.74	0.4611	0.51558	-0.76	0.4459	0.29206	3.42396
Assetvalue2	1	-0.07947	0.02147	-3.70	0.0002	0.02989	-2.66	0.0082	0.25528	3.91728
Assetvalue34	1	-0.47869	0.01464	-32.70	<.0001	0.01266	-37.82	<.0001	0.20430	4.89485
Assetvalue78	1	-0.10980	0.01625	-6.76	<.0001	0.01564	-7.02	<.0001	0.25686	3.89324

9. The next step is to do the diagnostics of residuals.

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
14	8.68	0.8511

P value .0.05, so we fail to reject the H0.It means that the error variances are constant.


Residual by Regressors for Rsq
Residual
AssetValue
B
D
H
Assetvalue2
Assetvalue34
10
5
0
-5
-10
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0


Residual by Regressors for Rsq
Residual
Assetvalue78
10
5
0
-5
-10
0
20
40
60
80

From the diagnostic residuals above, we can see the following:

a) 'Residual' vs 'Predicted Value': Error variance is roughly constant.
b) 'Rstudent' vs 'Predicted Value': Error variance is roughly constant but there are around 15 outliers are spotted.
c) 'Rstudent vs 'Leverage': Generally,most points are clustered except but there are some influencial points and around 15 outliers. May want to check if they are influential points.
d) 'Residual' vs 'Quantile'QQ plot, errors are roughly normal.
e) 'Prestige' vs 'Predicted Value':Clustered around the 45 degree line.
f) 'Cook's D' vs 'Observation' There are around 15 influential points are detected.
g) 'Percent' vs 'Residual'(Residual Histogram):Uni-modal, roughly symmetric.
h) 'Fit-Mean (Residual)' vs 'Proportion Less'-Residual fit plot: the spread in residual is no greater than the spread in centered fit.
i) From the residuals v.s regressor plot, we see no clear pattern. We are safe to utilize linear regression.

10. Delete the outliers and run the model again. This is the final model:

Parameter Estimates										
						Heteroscedasticity Consistent				
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	2.91985	0.25740	11.34	<.0001	0.23663	12.34	<.0001	.	0
AssetValue	1	0.82655	0.00839	98.50	<.0001	0.00644	128.30	<.0001	0.25130	3.97928
B	1	4.42292	0.89996	4.91	<.0001	0.36170	12.23	<.0001	0.26538	3.76816
D	1	2.08660	0.32387	6.44	<.0001	0.30696	6.80	<.0001	0.26772	3.73525
H	1	0.04112	0.36286	0.11	0.9098	0.32837	0.13	0.9004	0.27454	3.64243
Assetvalue2	1	-0.08324	0.02431	-3.42	0.0007	0.00955	-8.72	<.0001	0.25216	3.96572
Assetvalue3	1	-0.47847	0.01030	-46.46	<.0001	0.00837	-57.17	<.0001	0.18601	5.37616

Parameter Estimates										
						Heteroscedasticity Consistent				
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
4 Assetvalue78	1	-0.11953	0.01251	-9.55	<.0001	0.00985	-12.13	<.0001	0.26630	3.75522

The final model is as follows:

For assets under the rating category 1, 5 and 6: y(Rsq-hat)=2.91985+0.82655*Asset value

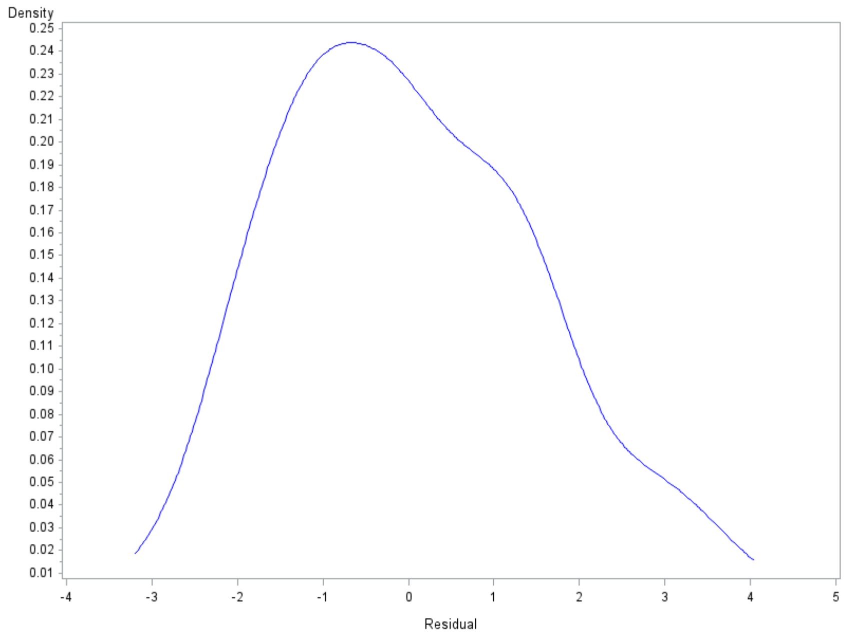
For assets under the rating category 2: y(Rsq)=7.34277+(0.74331) *Asset value

For assets under the rating category 3 and 4: y(Rsq)= 5.00645+(0.34685) *Asset value

For assets under the rating category 7 and 8: y(Rsq)= 2.96097+(0.70702) *Asset value

To test the model, we do a normality test

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.98185	Pr < W	0.0003
Kolmogorov-Smirnov	D	0.054839	Pr > D	0.0163
Cramer-von Mises	W-Sq	0.235628	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.522006	Pr > A-Sq	<0.0050



P value<0.05, reject H0. It means the residuals are not normally distributed. A more visual result can be seen from the density function curve: The residuals are not normally distributed and the data may require the transformation. We can add a second degree term to fit a polynomial function since our model failed the normality assumption.

The following tests are testing the correlations between the residuals and other independent variabes. The 0 correlation result proves the soundness of our model.

Pearson Correlation Coefficients, N = 332
Prob > |r| under H0: Rho=0

	residual	AssetValue	B	D	H	Assetvalue2	Assetvalue34	Assetvalue78
residual	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Residual		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
AssetValue	0.00000	1.00000	0.07449	0.04663	-0.08918	0.16160	0.55545	0.25142
	1.0000		0.1757	0.3971	0.1048	0.0031	<.0001	<.0001
B	0.00000	0.07449	1.00000	-0.16102	-0.10859	0.84482	-0.10759	-0.07800
	1.0000	0.1757		0.0033	0.0480	<.0001	0.0502	0.1562
D	0.00000	0.04663	-0.16102	1.00000	-0.51027	-0.13603	0.66816	-0.36650
	1.0000	0.3971	0.0033		<.0001	0.0131	<.0001	<.0001

Pearson Correlation Coefficients, N = 332
Prob > |r| under H0: Rho=0

	residual	AssetValue	B	D	H	Assetvalue2	Assetvalue34	Assetvalue78
H	0.00000	-0.08918	-0.10859	-0.51027	1.00000	-0.09174	-0.34094	0.71824
	1.0000	0.1048	0.0480	<.0001		0.0952	<.0001	<.0001
Assetvalue2	0.00000	0.16160	0.84482	-0.13603	-0.09174	1.00000	-0.09089	-0.06589
	1.0000	0.0031	<.0001	0.0131	0.0952		0.0983	0.2312
Assetvalue34	0.00000	0.55545	-0.10759	0.66816	-0.34094	-0.09089	1.00000	-0.24488
	1.0000	<.0001	0.0502	<.0001	<.0001	0.0983		<.0001
Assetvalue78	0.00000	0.25142	-0.07800	-0.36650	0.71824	-0.06589	-0.24488	1.00000
	1.0000	<.0001	0.1562	<.0001	<.0001	0.2312	<.0001	

Appendix: SAS code

```
%global Inpath Outpath;
```

```
%let Inpath= \\vfacad.yorku.yorku.ca\uit-
ugrad2\8\73\tsobhani\ADMS4370;
%let Outpath= \\vfacad.yorku.yorku.ca\uit-
ugrad2\8\73\tsobhani\ADMS4370\Result;

LIBNAME mylib "&Inpath.";

PROC IMPORT OUT=MYLIB.ACORR
DATAFILE="&INPATH.\PrivateFirmAssetCorr.CSV"
DBMS=CSV REPLACE;
GETNAMES=YES;
RUN;

/* q1*/
data Mylib.dummy;
  set Mylib.ACORR;
  if RatingCategory='1' then do
    B=0;
    C=0;
D=0;
E=0;
F=0;
G=0;
H=0;
  end;
  else if RatingCategory='2' then do
    B=1;
    C=0;
D=0;
E=0;
F=0;
G=0;
H=0;
  end;
  else if RatingCategory='3' then do
    B=0;
    C=1;
D=0;
E=0;
F=0;
G=0;
H=0;
  end;
  else if RatingCategory='4' then do
    B=0;
    C=0;
D=1;
E=0;
F=0;
G=0;
H=0;
end;
```

```
  else if RatingCategory='5' then do
    B=0;
    C=0;
D=0;
E=1;
F=0;
G=0;
H=0;
end;
  else if RatingCategory='6' then do
  B=0;
    C=0;
D=0;
E=0;
F=1;
G=0;
H=0;
end;
  else if RatingCategory='7' then do
    B=0;
    C=0;
D=0;
E=0;
F=0;
G=1;
H=0;
end;
  else if RatingCategory='8' then do
   B=0;
    C=0;
D=0;
E=0;
F=0;
G=0;
H=1;
  end;
  Assetvalue2=AssetValue*B;
  Assetvalue3=AssetValue*C;
  Assetvalue4=AssetValue*D;
  Assetvalue5=AssetValue*E;
  Assetvalue6=AssetValue*F;
  Assetvalue7=AssetValue*G;
  Assetvalue8=AssetValue*H;
  run;
  proc sgscatter data=mylib.dummy;
          plot (Rsq )*(Assetvalue  Assetvalue2 Assetvalue3 Assetvalue4
Assetvalue5 Assetvalue6 Assetvalue7 Assetvalue8 );
run;


  proc sgscatter data=mylib.DUMMY;
     compare Y=Rsq X=(AssetValue)/group=Ratingcategory reg;
```

```
run;

PROC SORT DATA=Mylib.dummy;
BY RatingCategory;
RUN;

PROC BOXPLOT DATA=Mylib.dummy;
    PLOT (assetvalue rsq)*ratingcategory/boxstyle=schematicid;
    insetgroup mean Q2 stddev/
           header='Statistics by Type';
     insetgroup min max/
           header='Extreme Values by Type';
run;

/*test if rsq is diff for each category*/
proc anova data=mylib.dummy;
 class RatingCategory;
model Rsq = RatingCategory;
run;

proc univariate data=Mylib.dummy normal;
var Rsq;
histogram rsq/normal;
qqplot rsq;
run;


proc corr data=Mylib.dummy pearson spearman;
var AssetValue RatingCategory Rsq;
run;


/*q2*/
data mylib.q2;
set Mylib.dummy;

AR=Assetvalue*ratingcategory;
  run;



oDS GRAPHICS ON;
proc reg data=Mylib.q2;
model Rsq=Assetvalue ratingcategory  AR/sse influence r spec vif tol
white;
run;
test RatingCategory,ar;
quit;
ODS GRAPHICS OFF;
/*drop AR becuse of  multicollinlity*/

ODS GRAPHICS on;
```

```
proc reg data=Mylib.q2;
model Rsq=Assetvalue ratingcategory  /sse influence r spec vif tol
white;
run;
test RatingCategory;
quit;
ODS GRAPHICS off;


/* should also drop RatingCategory as well*/
ODS GRAPHICS on;
proc reg data=Mylib.q2;
model Rsq=Assetvalue/sse influence r spec white;
OUTPUT OUT=Mylib.PREDq2 PREDICTED=PRED RESIDUAL=residual  UCL=UCL
LCL=LCL  rstudent=studentizedres  h=lev cookd=cookd;
run;

ODS GRAPHICS OFF;



proc univariate data=mylib.PREDq2 mu0=0 normal;
var residual ;
histogram residual/normal;
qqplot residual ;
run;

proc kde data=mylib.PREDq2 out=mylib.denq2;
  var residual;
run;

proc sort data=mylib.denq2;
  by residual;
run;

goptions reset=all;
symbol1 c=blue i=join v=none height=1;
proc gplot data=mylib.den;
  plot density*residual=1;
run;
quit;


/*Resideual assumption: Independent*/

proc corr data=mylib.PREDq2;
var AssetValue residual;
run;


proc reg data=mylib.PREDq2;
model  AssetValue = residual;
```

```
run;


/*Resideual assumption: constant variance*/


proc sgscatter data=Mylib.PREDq2;
     plot residual*pred;
run;

/*Resideual assumption: linearlity*/
proc sgscatter data=Mylib.PREDq2;
     plot residual*assetvalue;
run;
/*+ white test above*/





/*with  outliters delete*/
DATA MYLIB.q2delete;
set mylib.predq2;
if cookd>1/400
then delete;
run;


oDS GRAPHICS ON;
proc reg data=Mylib.q2delete;
model Rsq=Assetvalue/sse influence r spec  white;
OUTPUT OUT=Mylib.predq2delete PREDICTED=PRED RESIDUAL=residual
UCL=UCL LCL=LCL  rstudent=studentizedres  h=lev cookd=cookd;
run;

ODS GRAPHICS OFF;
/*drop AR becuse of  multicollinlity*/



proc univariate data=mylib.predq2delete mu0=0 normal;
var residual ;
histogram residual /normal;
qqplot residual ;
run;

proc kde data=mylib.predq2delete out=mylib.denq2delete;
  var residual;
run;

proc sort data=mylib.denq2delete;
```

```
  by residual;
run;

goptions reset=all;
symbol1 c=blue i=join v=none height=1;
proc gplot data=mylib.den;
  plot density*residual=1;
run;
quit;


/*Resideual assumption: Independent*/

proc corr data=mylib.predq2delete;
var AssetValue residual;
run;


proc reg data=mylib.predq2delete;
model  AssetValue = residual;
run;



/*Resideual assumption: constant variance*/
proc sgscatter data=Mylib.predq2delete;
     plot residual*pred;
run;

/*Resideual assumption: linearlity*/
proc sgscatter data=Mylib.predq2delete;
     plot residual*assetvalue;
run;


/*Q3*/
proc robustreg data=Mylib.q2;
   model Rsq=Assetvalue/ diagnostics;
run;


/*q4*/
/* Stepwise. bad modle!!!!*/
ODS GRAPHICS ON;
proc reg data=Mylib.dummy;
model Rsq=Assetvalue B C D E F G H   Assetvalue2 Assetvalue3
Assetvalue4 Assetvalue5 Assetvalue6 Assetvalue7
Assetvalue8/slstay=0.05 slentry=0.05 selection=stepwise sse influence
r spec vif tol white;
OUTPUT OUT=Mylib.PREDbad PREDICTED=PRED RESIDUAL=residual  UCL=UCL
LCL=LCL  rstudent=studentizedres  h=lev cookd=cookd;
run;
```

```
ODS GRAPHICS OFF;

/* DIY our good model*/

proc reg data=Mylib.dummy;
model Rsq=Assetvalue B C D E F G H   Assetvalue2 Assetvalue3
Assetvalue4 Assetvalue5 Assetvalue6 Assetvalue7 Assetvalue8 /sse;

run;
/*test ocerall AssetValue*/

proc reg data=Mylib.dummy;
model Rsq=Assetvalue B C D E F G H   Assetvalue2 Assetvalue3
Assetvalue4 Assetvalue5 Assetvalue6 Assetvalue7 Assetvalue8 /sse;

run;

testnosig: test AssetValue, Assetvalue2 ,Assetvalue3 ,Assetvalue4
,Assetvalue5, Assetvalue6, Assetvalue7 ,Assetvalue8;/*overall*/
testnosig: test  Assetvalue2 ,Assetvalue3 ,Assetvalue4 ,Assetvalue5,
Assetvalue6, Assetvalue7 ,Assetvalue8;/*diff than 0 slope exist*/
testsame: test  Assetvalue2-Assetvalue3-Assetvalue4-Assetvalue5-
Assetvalue6-Assetvalue7-Assetvalue8;/* diff slope*/
testrc5: test   Assetvalue5, e ;/*test overall rc5's effectst*/
/*fail!!*/

quit;
/*it is sig*/

/*drop rc5*/

proc reg data=Mylib.dummy;
model Rsq=Assetvalue B C D  F G H   Assetvalue2 Assetvalue3
Assetvalue4  Assetvalue6 Assetvalue7 Assetvalue8 /sse;

run;
testrc6: test   Assetvalue6,f ;/*test overall rc6's effectst*/
/*fail!!*/
/*drop rc6*/

proc reg data=Mylib.dummy;
model Rsq=Assetvalue B C D   G H   Assetvalue2 Assetvalue3 Assetvalue4
Assetvalue7 Assetvalue8 /sse;

run;
TEST  G,Assetvalue7;
test h,Assetvalue8;


test Assetvalue7-Assetvalue8;
```

```
test g-h;
test Assetvalue3-Assetvalue4;
test c-d;
run;
/*combine: 3+4=3_4*/
/* 7+8=7_8*/


/*combining data*/
PROC IMPORT OUT=MYLIB.ACORR1
DATAFILE="&INPATH.\PrivateFirmAssetCorr.CSV"
DBMS=CSV REPLACE;
GETNAMES=YES;
RUN;


data Mylib.comb;
  set Mylib.ACORR1;
  if RatingCategory='1' then do
    B=0;
    C=0;
D=0;
E=0;
F=0;
G=0;
H=0;
  end;
  else if RatingCategory='2' then do
    B=1;
    C=0;
D=0;
E=0;
F=0;
G=0;
H=0;
  end;

  else if RatingCategory in ('3','4') then do
    B=0;
    C=0;
D=1;
E=0;
F=0;
G=0;
H=0;
end;
  else if RatingCategory='5' then do
    B=0;
    C=0;
D=0;
E=1;
F=0;
```

```
G=0;
H=0;
end;
  else if RatingCategory='6' then do
  B=0;
    C=0;
D=0;
E=0;
F=1;
G=0;
H=0;
end;

  else if RatingCategory IN('7','8') then do
   B=0;
    C=0;
D=0;
E=0;
F=0;
G=0;
H=1;
  end;
  Assetvalue2=AssetValue*B;
  Assetvalue34=AssetValue*D;
  Assetvalue5=AssetValue*E;
  Assetvalue6=AssetValue*F;

  Assetvalue78=AssetValue*H;
  run;


/*D is C+D, H=H+G combining 34 and 78*/
  ods graphics on;
proc reg data=Mylib.COMB;
model Rsq=Assetvalue B D   H   Assetvalue2  Assetvalue34
Assetvalue78 /sse influence r spec vif tol white;
OUTPUT OUT=Mylib.PRED PREDICTED=PRED RESIDUAL=r  UCL=UCL LCL=LCL
rstudent=studentizedres  h=lev cookd=cookd;
run;
  ods graphics off;


/*test normarmilty*/

proc univariate data=mylib.pred mu0=0 normal;
var r;
histogram r/normal;
qqplot r/normal;
run;

/* delete delete outliers*/
DATA MYLIB.delete;
```

```
set mylib.pred;
if cookd>1/400
then delete;
run;

ods graphics on;

proc reg data=Mylib.delete;
model Rsq=Assetvalue B D   H   Assetvalue2  Assetvalue34
Assetvalue78 /sse influence r spec vif tol white collinoint;

OUTPUT OUT=Mylib.PREDdelete PREDICTED=PRED RESIDUAL=residual  UCL=UCL
LCL=LCL  rstudent=studentizedres  h=lev cookd=cookd;
plot rstudent.*p. / noline;

run;
test h ,Assetvalue78;
quit;
ODS GRAPHICS OFF;
/*white test above chisq right before the big table for constant
variance*/



/*Resideual assumption: normality*/
/*test normarmilty*/
proc univariate data=mylib.PREDdelete mu0=0 normal;
var residual ;
histogram residual /normal;
qqplot residual ;
run;

proc kde data=mylib.PREDdelete out=mylib.den;
  var residual;
run;

proc sort data=mylib.den;
  by residual;
run;

goptions reset=all;
symbol1 c=blue i=join v=none height=1;
proc gplot data=mylib.den;
  plot density*residual=1;
run;
quit;


/*Resideual assumption: Independent*/

proc corr data=mylib.PREDdelete;
```

```
var residual AssetValue  B D   H   Assetvalue2  Assetvalue34
Assetvalue78;
run;


proc reg data=mylib.PREDdelete;
model  AssetValue = residual;
run;



/*Resideual assumption: constant variance*/


proc sgscatter data=Mylib.preddelete;
     plot residual*pred;
run;
/*+ white test above*/



/*Resideual assumption: linearlity*/
proc sgscatter data=Mylib.preddelete;
     plot residual*(AssetValue  B D   H   Assetvalue2  Assetvalue34
Assetvalue78);
run;
```