Assignment 5: Final – SUR Report – Grunfeld Model: SUR

Predict 411

Section 56

Winter Quarter

****

School of Continuing Studies

Northwestern University

****

Daniel Prusinski

In compliance for Master of Science Predictive Analytics

Bachelor of Science Business Marketing

****

Program Analyst

Wooddale Church

6630 Shady Oak Road

Eden Prairie, MN 55344

Executive Summary

From 1935-1954, the United States faced some of its hardest times as well as best economic growth periods. The Great Depression, World War II, Reconstruction, and the Korean War all took place in this twenty-year span. US companies that were titans of industry prior to The Great Depression were susceptible to the macro world trends. As a result of these major world trends, analyzing financial data from this time period requires different modeling techniques. The familiar Ordinary Least Squares (OLS) model proves insufficient given that the stringent assumptions cannot be validated based on heteroscedasticity amongst residuals between the manufacturing firms being studied. In response to the heteroscedasticity, Seemingly Unrelated Regression (SUR), and Feasible Generalized Least Squares (FGLS) yield models that are more robust and satisfy the general assumptions for regression analysis. The data from this exploratory data analysis (EDA) reflects major world events and demonstrates the financial winners for this time period. General Motors, General Electric, and Chrysler emerged from this time period as winner in respect to growth and increased financial valuation. Through this EDA, economists can better understand contagions and develop strategies to offset such events.

Introduction

From 1929-1939, the Great Depression claimed over 10 million jobs, and through its duration left 1/3 of the non-farmer workforce unemployed (gwu.edu). Literally every American had "firsthand experience" with the Great Depression. The graphic to the right, created by itulip.com, displays the unemployment rate throughout The Great Depression, and each bar represents a year.

As a result of the Great Depression, US companies were hit hard and forced to find new avenues for revenue and growth from traditional business models up until that point in history. A helpful technique for gauging organizations productivity in this time period is to analyze gross investment from Moody's Industrial Manual and annual reports of corporations.

The objective of this report is to conduct an EDA explore the method and concept of seemingly unrelated regression analysis by analyzing a dataset that entails 10 manufacturing firms in the US from 1935 – 1954. Understanding the backdrop of this dataset, the Great Depression, is important to the context of the overall EDA. Through this EDA, I will start with the modeling technique ordinary least squares (OLS) for which I am familiar. As this EDA

develops, I will move on to appropriate modeling techniques that will result in parsimonious and statistically valid results and recommendations.

In the EDA, *I* is the dependent variable and is the gross investment from Moody's Industrial Manual and annual reports of corporations. To be clear, it is desirable to receive gross investment and the more investment one receives the better the organization is perceived to be operating. The independent variables in the data set are: F- the value of the firm from Bank and Quotation Record and Moody's Industrial Manual, C - the stock of plant and equipment from Survey of Current Business. All 10 organizations being analyzed have "hard" assets, and their stock prices are based on actual goods and services.

While The Great Depression was highlighted earlier, World War 2, the Roosevelt Plan, the Marshall Plan, as well as the Korean War all took place during the time that the data was collected. Given that the organizations being studied are manufacturers, I would expect a strong inter-correlation of results. This is justified as a result of The Great Depression taking place as well as major world events that capitalized on manufacturing output both in wartime and postwar rebuilding efforts.

From 2007 – 2009, the Great Recession claimed 8 million jobs and left over 14 million individuals unemployed (VanHorn & Zukin 2010). In a survey conducted by the Heldrich Center in 2010, 73 percent of surveyed Americans stated they had "firsthand experience" with the recession. This study of past organizations in this EDA will give insights and clues as to how companies regain footing and recover from tough economic times. As a result of this study, economists will better understand cross-equation correlation within similar organizations and how to spur macro sized economic recovery analysis. In my brief time as a data scientist in training, OLS has been the training/entry level model for which training wheels are to a bike.

Given that OLS has many assumptions which limit its usefulness, I look forward to exploring alternative methods, such as Seemingly Unrelated Regression (SUR), that are more robust in nature.

Analysis

In order to meet the objective of exploring the relationship between the dependent variable and independent variables, an exploratory data analysis must be conducted. This EDA will start with a basic OLS model. From that model, analysis will be made and if the OLS assumptions do not hold I will move on to SUR. As a data scientist in training, I am inculcating a paradigm of which to study data. While this paradigm is redundant report to report, it is training me to have the correct mindset. I will be utilizing the EDA paradigm and structure put forth by Bruce Ratner found in his book *Statistical and Machine-Learning Data Mining*:

> Problem/Objective: Explore the relationship between *I* (dependent variable) and independent variables F and C. Their interpretations were explained above.
>
> Data: The data has been aggregated and has been supplied from management. There are no missing values.
>
> Analysis: I will describe the data via simple descriptive statistics at first. After the initial analysis, I will start with fitting the data with an OLS model. Pooled regression and simple OLS will be utilized at this step. I am familiar with these two models, and will better understand the data through this process. If/when issues arise from the state of the data as a result of not meeting the OLS assumptions I will select a different model in order to adhere to sound statistical principle.
>
> Model: When issues arise from fitting the data with OLS as a result of endogeneity, management has instructed the use of a SUR model. The advantage of using this model is results that are statistically valid and adhere to the principles of data modeling.
>
> Results/Interpretation: Once the model has been created, an assessment of the model adequacy will be conducted to discern how well the model fits that data and the statistical backing of the model.

A properly executed EDA for management must reflect that the data was the driving force behind constructing the model. The steps outlined above are ordered such that the data drives building the best model, and the analyst's personal bias is mitigated.

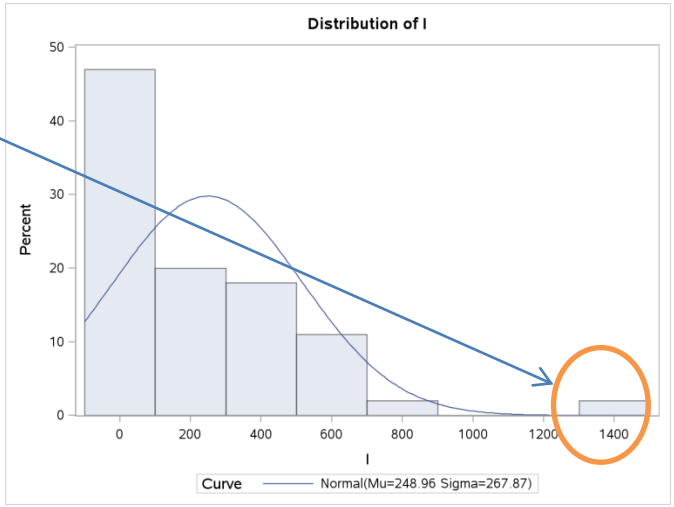
Data

There are a total of 100 observations with 0 completely missing values per row. Management has requested the focus to be on variables *I*, F, and C in this data set. The data is comprised of 5 manufacturing organizations over 20 years, of which three variables are recorded. Thus, each firm has three data values per year and in total there are 300 data values being studied in this EDA. The variables Year and Firm have been explicitly stated to be ignored, and will not be commented on in the EDA. Below you will find general descriptive statistics of the variables and their correlation with the response variable.

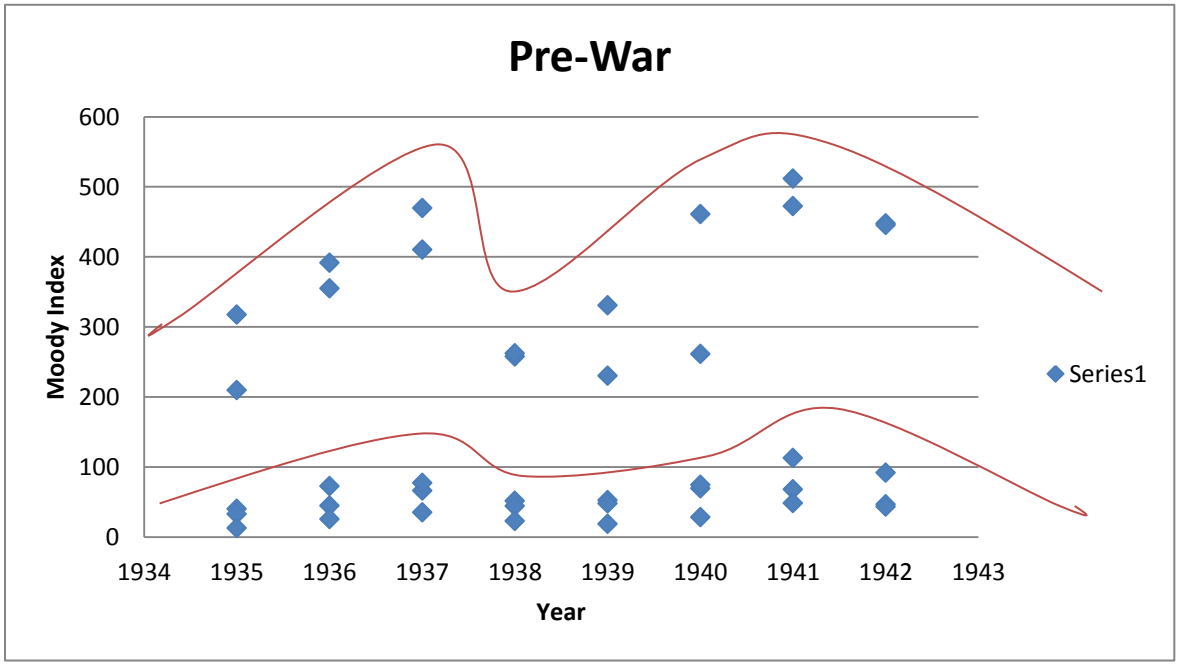
I is the dependent variable and is the gross investment from Moody's Industrial Manual and annual reports of corporations. To be clear, it is desirable to receive gross investment and the more investment one receives the better the organization is perceived to be operating. The baseline stats include 100 observations with a median of 140 and a mean of 249. What this tells me is there are more observations with small values and fewer big values, but the big values pull the skewness of the data by 1.976. The standard deviation is quite large but so too is the range. Visually looking at a histogram is helpful for experiencing the data. The vast majority of the values are between 0 and 100.

Variable I			
N	100	**Median**	140.100
Mean	248.957	**Range**	1474
Std Dev	267.865	**Skew**	1.979

The distribution then tappers off, but notice how there about 2 observations that are 600 points greater than any other value. These points would need to be investigated because they are pulling the distribution of the data.

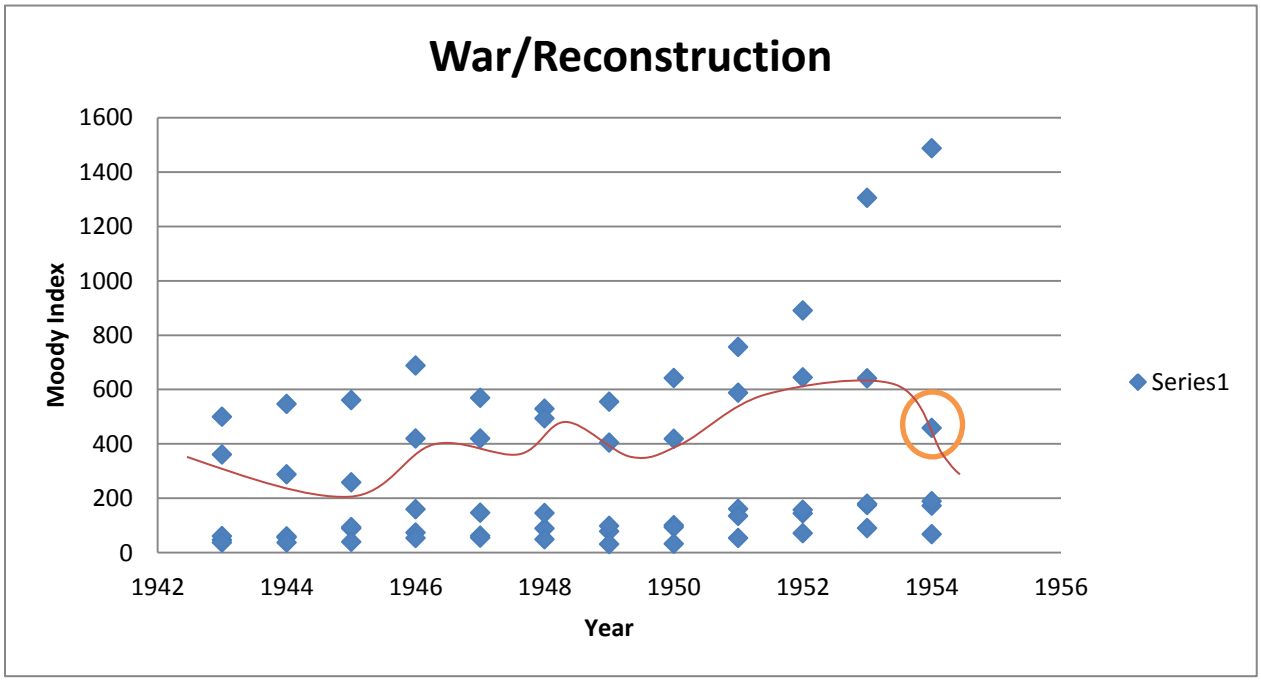


The data covers a timeframe that involved global instability. As a result, I will partition the data by pre-war era (1935-1942) and war/reconstruction era (1943-1955).



The stacked data points show the 5 different companies, and one can discern the overall trend for each company. The trends (red line) show that this time period was turbulent, but all

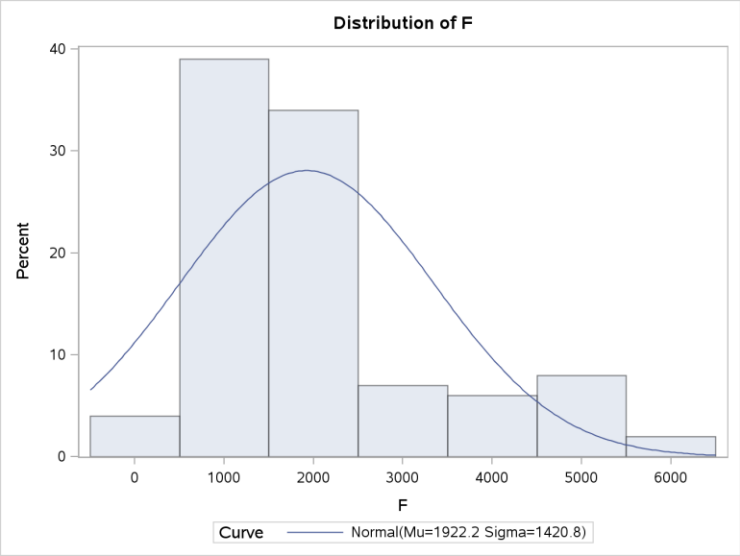
the companies were doing better by 1942 than in 1935. The growth is slow, stagnate, and uncertain.



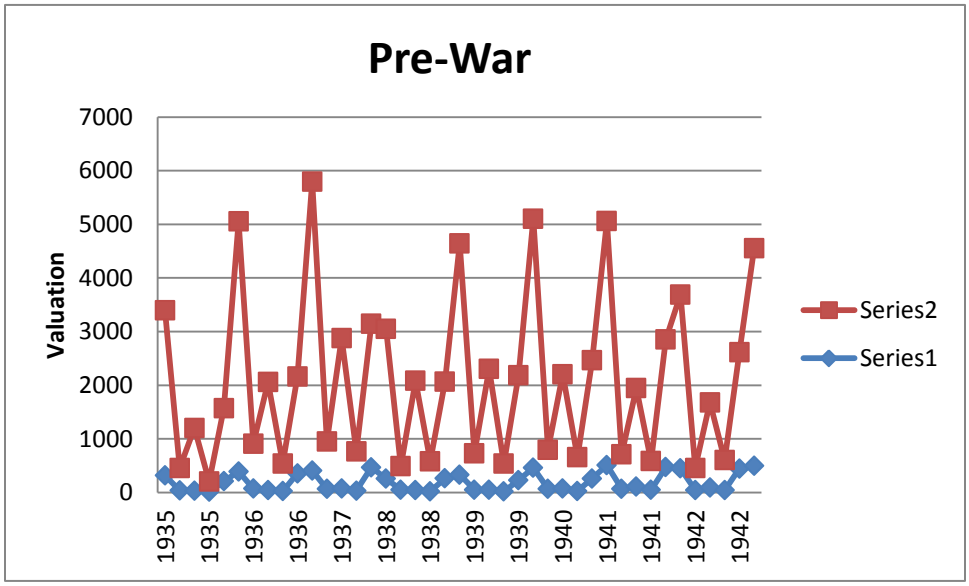
During the economically robust time period, three companies still teetered on disaster. Two companies saw growth, but notice how the smaller company grew stagnate from 1952-1954 and then declined. I would have expected more aggressive growth, and it appears that four of the five companies did not really improve their rating from the overall pre-war to post-war time period.

The independent variable F is the value of the firm from Bank and Quotation Record and Moody's Industrial Manual. A company desires to have a high valuation from this independent variable. The median and mean are rather close for having such a vast range. The skew is not as severe as the dependent variable. I would have expected a more modest standard deviation, but the histogram will better visualize the overall distribution.

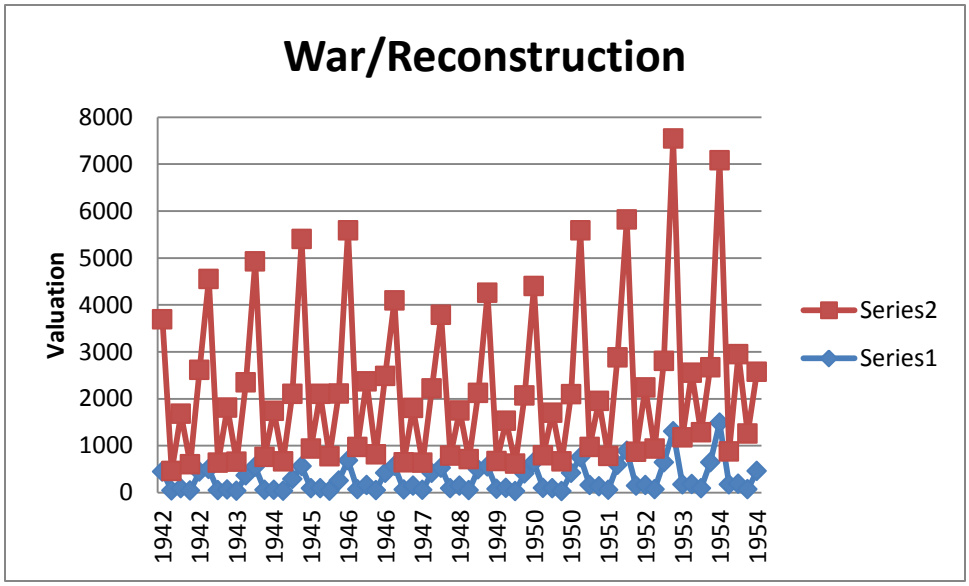
Variable F			
N	100	**Median**	1682.300
Mean	1922.223	**Range**	6050
Std Dev	1420.783	**Skew**	1.129



65 percent of the data points for F fall within 1,000 and 2,000. This leads me to believe that majority of the companies had this valuation throughout the data's time period. In addition, I want to further investigate the higher end valuations and whether or not they pertain to a specific company rather than a pooled value from all the companies. The valuation variable as a whole has a greater range, which leads me to believe it was more chaotic than the response variable.



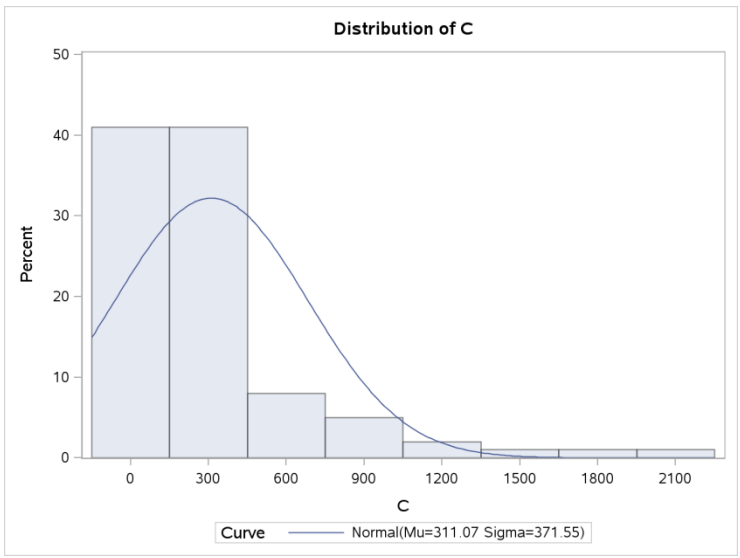
In this graph, series 2 represents the independent variable F, and series 1 represents the independent variable. It can be seen that while F has much larger swings, both variables emulate the same overall flow. At this point in the EDA, I would be led to believe that a mild to strong correlative relationship exists between F and *I*.



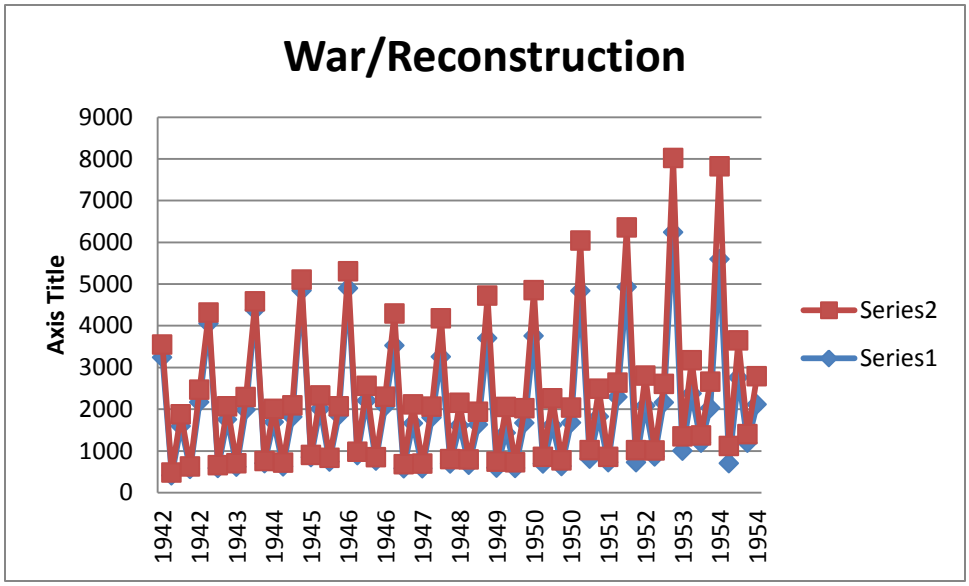
The pattern for the post war era follows the same trend as the pre-war graph. What is shocking about this graphical representation of the data is how volitlae the stock valuations were during that time period. Valuations jump and fall by over 50 percent in a two year span, which means fortunes could be made and lost in a day as well.

C is the other independent variable, and represents the stock of plant and equipment from the Survey of Current Business. This variables mean is 311 and the median is 205, which is rather skewed. The range is not as sporadic as variable F but still has quite a difference. Out of all the variables, C has the largest skew which demonstrated quite a few lower stock valuations.

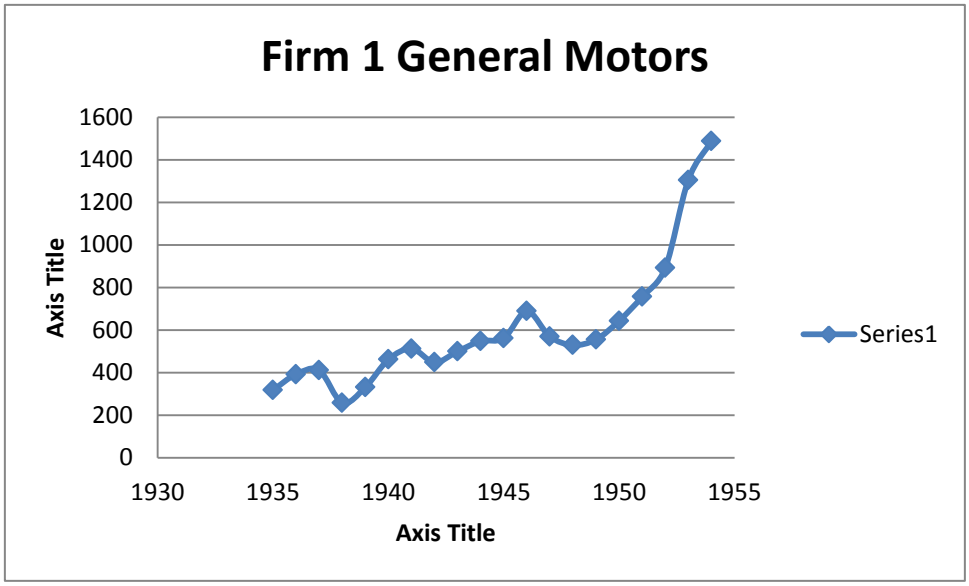
Variable C			
N	100	**Median**	205.350
Mean	311.067	**Range**	2226
Std Dev	371.552	**Skew**	2.695



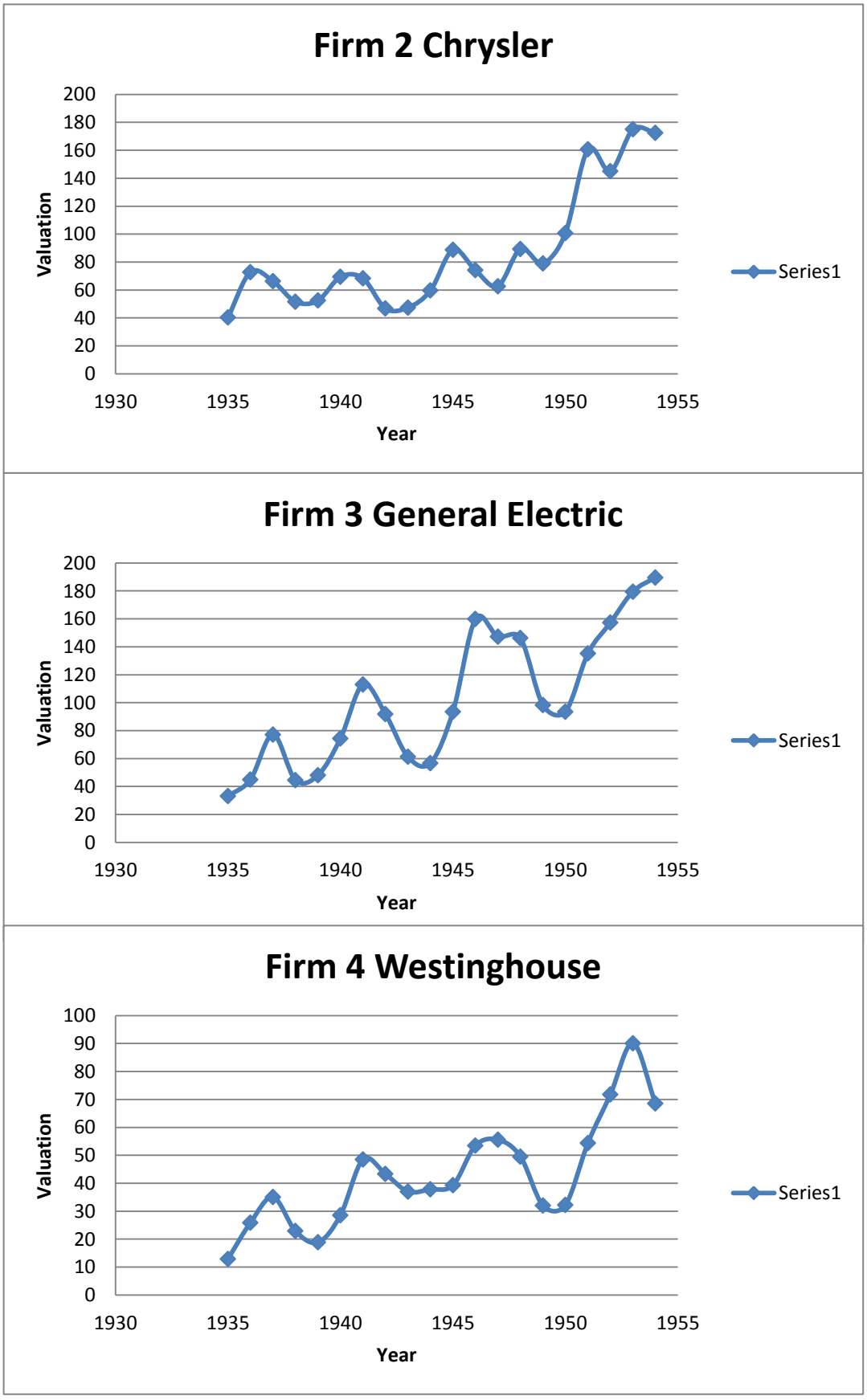
The histogram nicely demonstrates that 80% of the observations are less than 600, which leads me to believe that only a few companies saw there stock grow throuhgout this time period.



Series 2 represents variable C and Series 1 represents variable F. It can clearly be seen that these two variables are almost identical in their magnitude as well as trends. This raises a red flag because these two variables could have some collinear association. Given the graph from variable F, it is an intuituve exercise to realize variable C interacts in the same manner with the dependent varible. At this point in the EDA, I am concluding that the two indepenent variables will most likely by correlative with the response variable. In addition, I expect to see specific manufaturing companies that did well throughout the time span, as well as some others that stayed stagnate throughout the entire period.

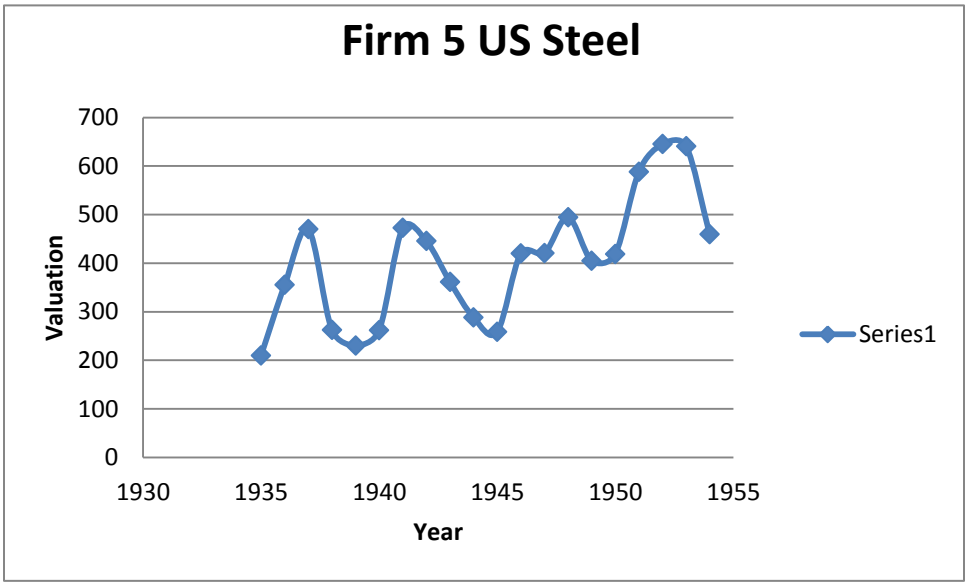


Each of these graphs represent the valuation of *I* throughout the time span. Firm 1 is a winner.



The growth is at almost 400 percent. Firm 2 is also a winner, its valuation is also around 400 percent greater at the end of the study.

The best company to own is Firm 2, it makes a very strong turnaround at the end of the series.



Firms four and five are the losers that stay rather stagnant, but still grow. The ROI for these two companies in not nearly as rewarding as the other companies.

Before diving into dealing with the heteroscedasticity, one needs to establish the reasoning behind the heteroscedasticity. One can draw conclusions from the specific time period that major macro world events influenced global markets and subsequently influenced industries. From basic intuition, General Motors and Chrysler produce similar product, as does General Electric and Westinghouse, and US Steel does not have a partner company in this data set. I would expect the macro world events to influence all the companies, but I would expect a stronger correlation amongst the companies based on their industries. Moving forward with this EDA, I expect strong correlation in general, and more correlation between the like firms.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5532554	2766277	170.81	<.0001
Error	97	1570884	16195		
Corr Total	99	7103438			

Root MSE	127.258	R-Square	0.778
Dependent Mean	248.957	Adj R-Sq	0.774
Coeff Var	51.116		

Parameter Estimates

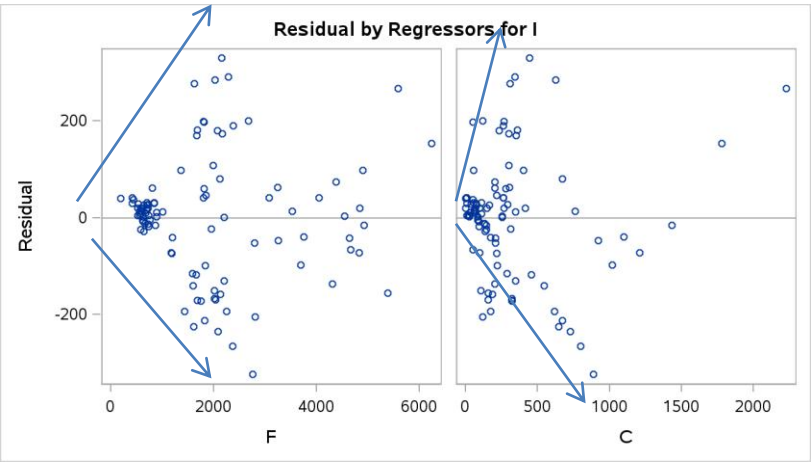
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	VIFS
Inter	Inter	1	-48.029	21.48017	-2.24	0.027	0
F	F	1	0.105	0.011	9.24	<.0001	1.597
C	C	1	0.305	0.043	7.02	<.0001	1.597

Results

From the data analysis, management has encouraged to start with a pooled OLS model for

each variable and then combine the variables into a multiple regression model. Preliminarily, the model has a strong R-squared, but the Adj R-squared is preferred given that the model has more than one variable. The F-value is very significant based on two degrees of freedom. This can be interpreted as at least one variable is explanative of the dependent variable in the model. Statistically the variables are significant. The variance inflation factors (VIFs) do not warrant concern for multi-collinearity. If any of the VIFs had been above 6 for any variable, the model would need to be adjusted for the effects of collinearity. Interpreting the coefficients are key for understanding how the model will predict. The coefficient for variable F can be interpreted as a one unit change in F equals a .105 unit change for dependent variable *I* holding all other things constant. Similarly, the coefficient for variable C can be interpreted as a one unit change in C equals a .305 unit change for dependent variable *I* holding all other things constant. Notice how the intercept is barely significant at the .95 percent level. This raises a concern, given that the intercept is very large.

Besides an outlier, the analysis of the residuals highlights the area of concern.



At this point in the EDA, checking the models diagnostics is done to validate the OLS assumptions for further assessment of the model adequacy. Specifically honing in on the residuals is vital for assessing the presence of heteroscedasticity. Utilizing scatterplots of the variables versus the residuals is a great initial process for detecting heteroscedasticity. The scatter plots capture the relationship of the residuals with the predicted variables. Notice the funnel/cone like scatter of the plots. Scatter

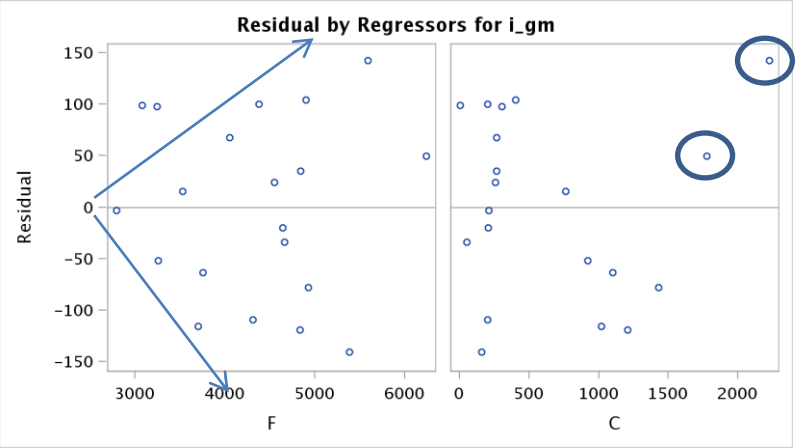
plots can be interpreted subjectively, but it is rather apparent that this plot suffers from heteroscedasticity. The next step is to conduct individual regression models to better understand the issues as hand.

Each of the five firms needs to be analyzed, and each will have its own regression equation. This is a new concept in regression modeling for my background. Please refer to the last paragraph of the data section, I fully expect a correlation of the random error terms between the equations. As a result of this inter-correlation, OLS is rendered useless based on violations in the OSL assumptions that result in outputs that are not valid. Feasible Generalized Least Squares (FGLS) is a modeling technique that takes multiple equations and satisfies the standard regression assumptions. Notice that FGLS does not satisfy the OLS assumption but rather the standard regression assumptions. Through the next section, I will demonstrate OLS and FGLS for each equation. The overarching goal is to utilize a modeling technique that is appropriate for dealing with the heteroscedasticity of the equations, and produce viable output to draw conclusions and recommendations.

Firm 1: General Motors (GM)

In the dataset, GM was hit hard by the initial effects of The Great Depression, but grew strongly at the end of the time period. Initially, consumers could not purchase the primary products supplied by GM but as time progressed the US government purchased large amounts of tanks, vehicles, and other GM products. As a result, this jump started profitability for GM and sustained as consumers gained back the purchasing power post World War II. The initial OLS regression output demonstrates a strong model. In appendix 1, one will find the SAS output that shows that the initial p-value for the f-value is strong as well as the f-value itself. In addition, the

r-square is strong and the coefficients are statistically significant. On the surface this OLS model looks great, but one must analyze the diagnostics to validate the assumptions of OLS.



The residual plot shows heteroscedasticity for the variable F, and notice the two outliers in variable C. While the model is very appealing at first glance, it can be concluded that the model suffers from heteroscedasticity and a different approach will be needed. Furthermore appendix 1 shows the output from Proc Reg in SAS, and visually demonstrates that all the OLS fitted models suffer from heteroscedasticity when analyzing the scatter plots of the residuals.

Cross Model Covariance					
	GM	CH	GE	WE	US
GM	8423.88	-332.655	714.74	148.443	-2614.2
CH	-332.65	176.320	-25.15	15.655	491.9
GE	714.74	-25.148	777.45	207.587	1064.6
WE	148.44	15.655	207.59	104.308	642.6
US	-2614.19	491.857	1064.65	642.571	10466.4

SUR has been used to analyze the correlation of the random errors, which saves time from having to analyze each variable by itself. Essentially, the variables are stacked in a matric and the covariance is calculated. The cross model covariance is the Mean Square error of each variable along with the covariance between the variables. The cross model correlation is where I draw the conclusion about the stacked models being correlated. As can be seen from the highlighted cells, US steel is mildly correlated with Chrysler and General Electric. This

Cross Model Correlation					
	GM	CH	GE	WE	US
GM	1.00000	-0.27295	0.27929	0.15836	-0.27841
CH	-0.27295	1.00000	-0.06792	0.11544	0.36207
GE	0.27929	-0.06792	1.00000	0.72896	0.37323
WE	0.15836	0.11544	0.72896	1.00000	0.61499
US	-0.27841	0.36207	0.37323	0.61499	1.00000

comes as no surprise based on the world events at the time. US steel supplied the raw metals for GE and WE to use in production during the war. In addition, General Electric and Westinghouse have a strong inter-correlation. I expected this for two reasons; first, they both compete in the same industry, secondly, both supplied major components during the war and both were strong brands after the war in the reconstruction efforts. At first glance, I would have expected GM and CH to be related. In reality, the companies did play a major role in this time period, but they supplied different goods which led to differentiation. The seemingly unrelated regression model has a strong adjusted r-squared and proves to be a solid model.

Feasible Generalized Least Squares (FGLS) will be used to fit the model in order to relax the assumptions for regression analysis. Sigma is often unknown and needs to be estimated, this is where FGLS estimates the coefficients. Appendix 2 shows the output from fitting the FGLS to each variable. It can be seen that for the GM model the intercept is smaller and the variables have more impact. The Chrysler model also has a smaller intercept, which gives the variables more impact in the model. For the General Electric, the FGLS model does not do much. The standard errors decrease somewhat, but the model does not improve that much. Similarly, WE has a lower standard error but the model does not improve. The standard errors for the US steel model decrease more significantly, but the models remain similar. By fitting the data to a FGLS model, it relaxes the regression assumptions and the model is validated. This is a major accomplishment because the data drove the modeling decision process, which is a core objective for any EDA.

Future Work

Further recommendations on how this study can be improved upon are the following:

- Time permitting, it would be great to utilize the heteroscedasticity diagnostic tests to further validate this data set.
- It would be an absolute party to analyze other companies from different countries and compare their data output during this time period. Countries of interest would include Japan, Germany, Australia, New Zealand, and Switzerland. From this study, one could ascertain from an economic standpoint the winning and losing economies.
- Economists often compare the Great Recession to The Great Depression, and one could analyze the same companies data for these different time periods and look for similar trends.

Through this initial EDA, coupled with the future work recommendations, economists would gather pertinent information in regard to economic recovery macroeconomic events.

References

Ajmani, V. (2009). *Applied Econometrics Using the SASSystem*. Hoboken: John Wiley & Sons.

Allison, Paul David. *Logistic regression using SAS theory and application, second edition*. 2nd ed. Cary, N.C.: SAS Institute, 2012. Print.

Doyle, Alison. "Unemployment Benefits." *Job Search, Interview & Employment Advice from About.com*. N.p., n.d. Web. 22 Jan. 2013. <http://jobsearch.about.com/cs/unemployment/a/unemployment.htm>.

Van Horn, Carl, and Cliff Zukin. "Unemployed Workers and the Great Recession." *Work Trends Reports, 2009-2010* 1.1 (2010): 1-19. *John J. Heldrich Center for Workforce Development Rutgers University*. Web. 22 Jan. 2013.

"The Anguish of Unemployment." *Rutgers*. Version 1. Rutgers State University, 1 Sept. 2009. Web. 22 Jan. 2013. <http://www.heldrich.rutgers.edu/sites/default/files/content/Heldrich_Work_Trends_Anguish_Unemployment.pdf>.

Janzen, Eric. "FIRE Economy Explosion Fallout -- Part I: Recession ends, depression begins ." *Itulip*. N.p., n.d. Web. 7 Feb. 2013. <http://www.itulip.com/forums/showthread.php/11043-FIRE-Economy-Explosion-Fallout-Part-I-Recession-ends-depression-begins-Eric-Janszen>.

"The Great Depression (1929-1939)." *The George Washington University*. N.p., n.d. Web. 9 Feb. 2013. <http://www.gwu.edu/~erpapers/teachinger/glossary/great-depression.cfm>.

Leuchtenburg, William E. *Franklin D. Roosevelt and the New Deal, 1932-1940.* New York: Harper Torchbooks, 1963, passim.

McElvaine, Robert S. *The Great Depression: America, 1929-1941*. New York: Times Books, 1993, passim.