2019

A Statistical Investigation of Stock Market Activity and Instances of Financial Crime

Elizabeth G. Biggs

COLUMBUS STATE UNIVERSITY

A STATISTICAL INVESTIGATION OF STOCK MARKET ACTIVITY AND INSTANCES OF FINANCIAL CRIME

A THESIS SUBMITTED TO

THE HONORS COLLEGE

IN PARTIAL FUFILLMENT OF

REQUIREMENTS OF THE HONORS COLLEGE

FOR HONORS IN THE DEGREE OF

BACHELOR OF SCIENCE

DEPARTMENT OF MATHEMATICS

BY

ELIZABETH G. BIGGS

COLUMBUS, GEORGIA

2019

A STATISTICAL INVESTIGATION OF STOCK MARKET ACTIVITY AND INSTANCES OF FINANCIAL CRIME

By

Elizabeth G. Biggs

Committee Chair

Dr. Kristin Seamon Lilly

Committee Members

Dr. Ronald Linton
Dr. Brett Cotten
Dr. Cindy S. Ticknor

Abstract

This project aims to identify a possible statistical relationship between the stock market and incidences of financial crime in Illinois, Texas, and Utah through big data analysis techniques using data from the FBI's Federal Bureau of Investigation's National Incident-Based Reporting System (NIBRS) database. By analyzing the occurrences of financial crime by location, year, and type in relation to the S&P 500's percent change in close price on the first day of the calendar year, it will be possible to determine if a statistical relationship exists between the two. In addition, regression analysis was performed in order to predict financial crime incidence using stock market prices.

Table of Contents

List of Tables

List of Figures

Introduction

Predictive analysis is a statistical technique often used in business analytics but is also useful in predicting the occurrence of crime. The use of analytics to predict crime has been increasing and is incredibly useful to law enforcement since it allows for a smaller number of officers to effectively do their jobs (Bakke, 2019). Using past data, researchers can predict where officers will be needed. Some of the types of crime that have been analyzed include terrorism, organized crime, and tax evasion (Bakke, 2019). Financially, predictive analysis has been used to rank tax returns by the likelihood of tax evasion based upon specific characteristics of the returns to make the detection of tax evasion more efficient (Bakke, 2019). Stock market indicators have been used in the past to predict other financial occurrences, so the linking of the stock market to the occurrences of financial crime would be a reasonable analysis to perform. The purpose of this project is to use regression analysis, one statistical technique used in predictive analysis, to examine the relationship between stock market performance and the reported incidents of financial crime in order to determine if a model could be created to predict the likelihood of financial crime.

Data Analytics, Modeling, and Regression

Data analytics allows researchers to take copious amounts of data and produce applicable results through numerous tools such as modeling and regression (Kte'pi, 2016). Within the field of data analytics, the data analyzed can be used to predict a variable. The input of the data is known as the predictor variable(s), and the output of the data is the response variables or the variable that is to be measured or predicted (James et al., 2017). Predictor variables are the variables that are controlled by the researcher and are the variables that are expected to have a measurable effect on the response variable. The response variable is the dependent variable that is being measured in order to build a model or to predict future values. The goal of predictive analysis is to use the data analyzed to form a predictive model to determine the correct output from the given inputs. Though this is a powerful tool, it is difficult to do since correlation does not always signify causation and the relations between variables are not always adequate for prediction analysis.

According to the *Salem Press Encyclopedia of Science*, data analytics is a comprehensive view on information and large amounts of data that aims to produce results that are comprised of trends and patterns (Kte'pi, 2016). One of the more important parts of data analytics revolves around the data itself. For the results of the study to be accurate and credible, the data needs to be reliable, meaning that the source is trustworthy and that the data itself is formatted so that the variables for analysis are clear and accurate (Piegorsch, 2016). Piegorsch (2016) remarks on the commonality that the data analyst is not always in control of the data they use, and the data can be from another source and direct attention cannot be feasibly maintained over the data collection and entry processes. Therefore, it is crucial to closely examine data for its quality.

According to Piegorsch (2016), two manners of data quality distortion are prevalent: individual and collective. The first is "individual" distortion, meaning that the majority of the data is good, though some individual errors are present either because of collection, entry, or other means. Many of the errors that would be labeled as "individual" errors would be a typography error in the entry of the data such as transposing numbers or double tapping a key to enter the wrong value. The occurrence of individual errors can be significant, since the results could be reflecting the error(s) as a trend when it is not a correct representation of the data, though under closer scrutiny, the errors can be detected, and the trend can be determined to be false (Piegorsch, 2016). The second is “collective,” meaning that the majority of the data is marred by errors categorized by the sampling or the identification of data. According to Piegorsch (2016), collective data errors corrupt the quality of the data overall and can disorient the data frame, so measures such as checking for outliers are put in place to reduce or negate collective errors. A data frame is the specific population the data is being taken from, meaning that the data frame is all of the individuals or events that could be in the study based on the set characteristics of the study. For example, the data frame for a study on autism in adolescents would have a data frame of people with the characteristic of the individuals being under 18 years of age, since that is the population to be represented in the study. It is important in data collection to have a data frame that accurately represents the population to be analyzed in order to have good results.

Regression is the use of mathematical modeling to predict the value of the response variables based on the values of the predictor variable or variables (Wienclaw, 2013). With regression analysis, a predictive model is formed based on the data in the form of an equation of best fit. The best-fit equation is determined by what equation is most representative of the data and would have the most accurate predictions of the response variable. One of the most frequently used methods to find a best-fit equation is the least-squares method, which aims to minimize the sum of the squared residuals. Residuals are important because the analysis of the residuals determines the accuracy of the best-fit equation. The residual is the difference between the observed and predicted values of a response variable. Residual analysis can be used to determine which equation has the best-fit equation when compared to the model’s other regression equations. For residual analysis, the focus is on the equation’s predicted response values and the actual response values. For least-squares regression, the residual plot is interpreted based on the sum of squares of the residuals, which evaluates the amount that the predicted response values deviate from the actual values. The smaller the difference of the residuals, the better the fit of the model. In addition, residual plots should be uniform, meaning that the plotted points should be equally distributed about the x-axis versus showing a heavy skew, identifiable wave patterns, or other unequal distribution (Weinclaw, 2013).

Within regression-based data analytics projects, the goal is to determine what predictor variables best predict the response variable by determining which variable(s) explain most of the response variable's variation. Statistically, this is done using the coefficient of determination, denoted as the "R^2" value. The R^2 value measures how close the actual data points are to the estimated fitted regression line. If a model has a high R^2 value such that it is close to 1, it means that the dependent variations in the model is predicted by the majority of the variation within the model and, therefore, the model shows a high correlation between the input and output variables. Correlation is important but having a correlation between the predictor variable(s) and the response variable is only part of the analysis necessary to predict response values accurately (Wienclaw, 2013). In addition to having a strong correlation coefficient, a regression analysis is also necessary. Regression analysis is the determination of what independent variables influence a dependent variable and the strength of the impact of the independent variables. It is always important to remember that though there may be a strong correlation between the predictor variables and the response variables, an understanding of the field is also necessary to determine whether there is a logical or possible connection between the predictor variables and the response variables to ensure that the results of the data analytics study are valid.

Regression, according to Wienclaw (2013), requires the assumptions that the data is reliable. In application, these assumptions are not always met, since real world situations are not commonly ideal. Problems with data projects can arise from this. In some cases, the function form can be incorrect, and therefore the variable correlation is also unreliable (Weinclaw, 2013). Least-squares regression is primarily used on normalized data where there is not a skew or outliers since least-squares regression would be a good estimation of non-skewed data. For skewed data, Least Absolute Deviation (LAD) regression may be more appropriate (Bloomfield and Steiger, 1980). Regardless, the type of regression used is determined by the scatter plot of the data itself.

García, Ramírez-Gallego, Luengo, Benítez, and Herrera (2016) explained that data sets with a large quantity of predictor values can encounter an issue regarding the project's amount of processing required. One of the ways to reduce the processing costs of a project is to reduce the number of predictor variables used in the regression analysis. Removing predictor variables haphazardly is reckless in an analytics project since the predictor variables are used to predict the response variable so that the true variations can be visualized. One of the methods García et al. (2016) describes is feature selection, which reduces the quantity of predictor variables in the analysis in order to reduce the computation cost as well as potentially better the algorithms determined in the project. This is done by lowering the likelihood of overfitting, so the variation of the errors is modeled in addition to or in place of the response variable's variation, making feature selection valuable since the errors detract from the variation of the response variable that we want to generalize and explain (Garcia et al., 2016). Also, dimensionality reduction can be achieved through space transformations which transform the original predictor variables into new predictor variables by combining some of the initial variables into new variables, and therefore reducing the total number of predictor variables to be used in the regression model (Garcia et al., 2016).

When developing a regression model, an essential step is to graph the data points. When graphed, data points can appear as several different overall shapes that imply a type of regression modeling will be a better fit than other types of regression modeling. When data points appear as a simple linear or straight line, the analyst will use linear regression in order to find the line of best fit. According to Wienclaw (2013), simple linear regression is the modeling of one predictor variable to one response variable to predict the response variable's value, so that a single variable is used as the input to form the line of best fit for the response data points. Non-linear regressions are also common. If, when looking at the data points graphically, the relationship appears to be quadratic or exponential, then the analyst should perform regression based on the general appearance of the data. If the relationship between two variables is, for example, quadratic, linear regression or multiple linear regression is unlikely to accurately fit the model that is best to predict the response value. For predictor and response variables with a quadratic relationship, multiple curvilinear regression or multivariate polynomial regression is likely to be a better fit for regression analysis for the data (Wienclaw, 2013). In addition, even if the model seems to fit the data based on the R^2 value, if it does not match the appearance of the data, the regression has potential to be a false or misleading result. It is important to have thorough knowledge of the subject matter to have a fully formed analysis.

Variables Used in the Analysis

To investigate the relationship between financial crime and the stock market performance, I used a single predictor variable instead of multiple variables for my analysis. For my predictor variable, I used a stock market index, the S&P 500 annual close price as reported by the ticker symbol ^GSPC on January 1st for years 2006 through 2017. The S&P 500 is a U. S.

stock market index based on a diverse selection of 500 large companies that have stock traded on the NYSE or NASDAQ. The S&P 500's is considered representative of the U. S. stock market and as a predictor of the U. S. economy ("S&P 500 (GSPC) Price History", 2019).

For the response variable, I used federally reported crime data within the to the Federal Bureau of Investigation's National Incident-Based Reporting System (NIBRS). The NIBRS database is a comprehensive collection of law enforcement's crime data. This data is particularly extensive, as it encompassed numerous details about the incident that is not always recorded in other crime data bases, especially local databases. Some of the more unique data variables reported include information regarding the victim(s), relationships between the victim(s) and perpetrator(s), and the property involved in the crime. The NIBRS database is able to provide more context when compared to other crime databases, because it provides information regarding specific events versus a general overview. The next most detailed database, the Summary Reporting System (SRS) provides an aggregate monthly tally of crimes, which would be beneficial to determine a possible relationship between the financial indicators and crime in the United States, but would not be able to determine whether a trend related to day of the week or day of the month exists within the data. Due to the ability to go more in depth when looking for a relationship, the NIBRS database is the better source of data for this analysis. The NIBRS is not an immaculate database, though. Due to it being the new format that law enforcement agencies are in the process of converting to, it is not all encompassing. Fourteen of the fifty states within the United States are not represented in the database. In addition, some states have more years of quality data available than others.

Procedures

Even though the NIBRS database is incomplete for all states from 2007 to 2017, Illinois, Texas, and Utah had the files for all years from 2007 to 2017. Though more states were available, I only used the data from years 2007 to 2017 from Illinois, Texas, and Utah due to time constrictions. My original intent was to use all available states, though technical equipment failure prohibited me from encompassing more states in my analysis. My data included Illinois, Texas, and Utah since their data files were complete. If there are identifiable patterns in the analysis for these three states, then in the future analysis can include all available states.

The crime files downloaded as a zip folder for each of the states and years. In each zip folder, there were several files, but of the files within each zip folder, I only needed the incident file, the offense file, and the offense type file for each state and year. In the Incident file, I manually added columns for the State and the Year and the other columns were agency_id, incident_id, nibrs_month_id, incident_number, cargo_theft_flag, submission_date, incident_date, report_date_flag, incident_hour, cleared_except_id, cleared_except_date, incident_status, data_home, ddocname, orig_format, and ff_line_number. From the Offense file, I used the columns offense_id, incident_id, offense_type_id, attempt_complete_flag, location_id, num_premises_entered, method_entry_code, and ff_line_number. From the Offense Type file, I used the columns offense_type_id, offense_code, offense_name, crime_against, ct_flag, hc_flag, hc_code, and offense_category_name.

I saved each of these three files for each state and year as .csv files. Next, I connected the data files for each state and year through file joins. For joining, unioning, and cleaning the data, I used Tableau Prep. Tableau Prep is a tool created to look at data to see what information is contained within the data and to pinpoint what part of the data needs to be cleaned (Tableau

Software, 2018). It is good for merging columns that are separate due to the column header name being incorrect, or in the case of the unioned file, columns that have inconsistent header names, such as “INCIDENT_ID” and “incident_id”. In addition, Tableau Prep also shows the user the values that each column contains with the count of instances of that value in a frequency graph. With this, the user can see the values to look for typos and such and correct the value. For example, if a year was mistyped so that instead of saying 2007 for the year the value was 20007, then the user can quickly correct the input error. In addition, if the values were inconsistent in capitalization or symbolization, it can also be easily changed to be consistent so that the values that are equivalent can be read by analysis programs correctly or to rename the columns for the user’s convenience. Also, Tableau Prep also shows the user what columns have null values and the user can treat the null values based on the user’s discretion. Due to its highly visual nature, Tableau Prep’s cleaning functions were advantageous to this project.

When combining the data, I began by taking the three files for each state and year and, using Tableau Prep, I joined all three of the files together using an inner join. I used an inner join because if the information in the three files did not have information from the other files, the row would not be usable in my analysis and an inner join excludes rows where the row is not represented in both of the columns used in the join clause, meaning that the if the row is does not have corresponding information in both files, them the row is not included in the final inner join file. The data rows that do not incorporate all three files are useless to the analysis because I would not be able to determine the type of crime to see whether it was financial in nature or to determine when the incident occurred. The files join based on a user-specified join clause. A join clause is comprised one column from each file being joined and is used to identify which rows should be together in the resulting file. The only file that I was not able to use all of the data from was the Utah 2008 file due to an incomplete file about the offense type. Though I had approximately 140,000 incidents reported in Utah in 2008, I was only able to join the incidents to their correct offense type for approximately 80,000 incidents. Since I was not able to connect about 60,000 incidents to their offense type to determine whether or not the incidents were a financial crime, I was not able to use the 60,000 or so incidents without joinable information and, therefore, they were not included in the final file of data since the inner join removed them. For all other files for Utah, Texas, and Illinois for all years from 2007 to 2017, the files joined fully and did not have any mismatched rows of data that lead to a loss of rows.

After joining the files for each year in Tableau Prep, I then unioned all of the joined files to create one large file with all of the data from all three states and all years from 2007 to 2017. A union combines all rows by the header name. If the header names do not exactly match, then a new column is created through the union by Tableau Prep with the unlike header name as the new column's header. Though the files contained the same data, since Tableau Prep is case-sensitive in regards to column names, the unioning of 33 files of data created several duplicate columns for variables that were actually identical but were determined to be different by the program. For example, Tableau Prep read "Incident_ID" and "INCIDENT_ID" as two different variables and created a new column, while the two columns in actuality contain the same variable's data. Since several new columns were unnecessarily created, I went through the columns in the file created through the unions and merged all files that were identical in the information they contained. After I merged the columns that were equivalent in content, I went through the remaining columns and judged the quality of the data they held. For several columns, all of the rows were null. Since these columns held not value or information, I removed the columns since they were unnecessary to my analysis. Cleaning the data and merging these equivalent columns is essential to have good variables to use in analysis in order to have accurate results.

When cleaning the file, I removed all incidents of crimes that were not financial in nature, such as kidnapping, assault, and robbery since I focused on only financial crime in my analysis. The financial crimes I kept in the data to use in the analysis were embezzlement, false pretenses/swindle/confidence game, credit card/automated teller machine fraud, impersonation, welfare fraud, wire fraud, identity theft, counterfeiting/forgery, and extortion/blackmail. All other crimes were removed from the crime data to be excluded from the analysis for their lack of financial nature.

After cleaning the data with Tableau Prep, I exported the cleaned data from Tableau Prep to a .csv file. I then opened the .csv file using Excel in order to summarize the crime data. I used Excel's pivot table function to summarize the data by determining the count of financial crime incidents of each type by year as well as the total count of all financial crime as seen in Table 1.

Table 1. *This table is a summary of the occurrences of financial crimes divided by year and type crime as well as a Grand Total of financial crime occurrences. Made using Excel.*

Year	Counterfeiting /Forgery	Credit Card /Automated Teller Machine Fraud	Embezzlement	Extortion /Blackmail	False Pretenses /Swindle /Confidence Game	Identity Theft	Impersonation	Welfare Fraud	Wire Fraud	Grand Total
2006	12185	9024	1136	42	5955	-	7750	14	374	36480
2007	11646	10529	1291	29	6739	-	7650	4	374	38262
2008	7950	8322	1315	24	5869	-	6879	4	302	30665
2009	8813	9473	1219	26	6331	-	7641	12	466	33981
2010	7959	10294	885	35	6159	-	6950	30	386	32698
2011	7017	8759	922	41	6141	-	6259	9	409	29557
2012	6143	9022	842	46	6420	-	4938	18	373	27802
2013	6126	8772	848	44	7070	-	4428	10	449	27747
2014	6239	8202	914	39	7446	-	5508	9	423	28780
2015	6727	9388	869	82	7970	-	8551	10	559	34156
2016	8128	11272	921	102	8185	156	7236	16	709	36725
2017	9174	12327	914	128	9723	2490	5091	24	656	40527

After creating this summary file, I then needed to add the S&P 500 financial data. I got the close price of ^GSPC for the S&P 500 from Yahoo Finance. I added a column to my summary file that contained my calculated percent change of the S&P 500’s close price from January 1st of year for the years 2006 to 2017. I used the S&P 500’s annual percent change as the predictor variable in my analysis. I used the percent change of the close price rather than the actual closing price because it provides a better indicator of the stock market performance when comparing specific time periods. In other words, a 10% change in today's market would be approximately 300 points but would be 34 points in the year 1990.

Data Analysis

For the data analysis, I used R, a statistical program that can run an assortment of tests and measures in addition to generating graphs and other visuals (R Core Team, 2017). After creating a good summary file of the data that included both the financial crime and the S&P 500 predictor variable data, I used R to create a scatter plot of the data as shown in Figure 1. I used R's "scatter.smooth" function to create a scatter plot of the data with the x-axis being the Percent Change of the S&P 500's annual close price and the y-axis being the count of incidents of all financial crimes from all three states.

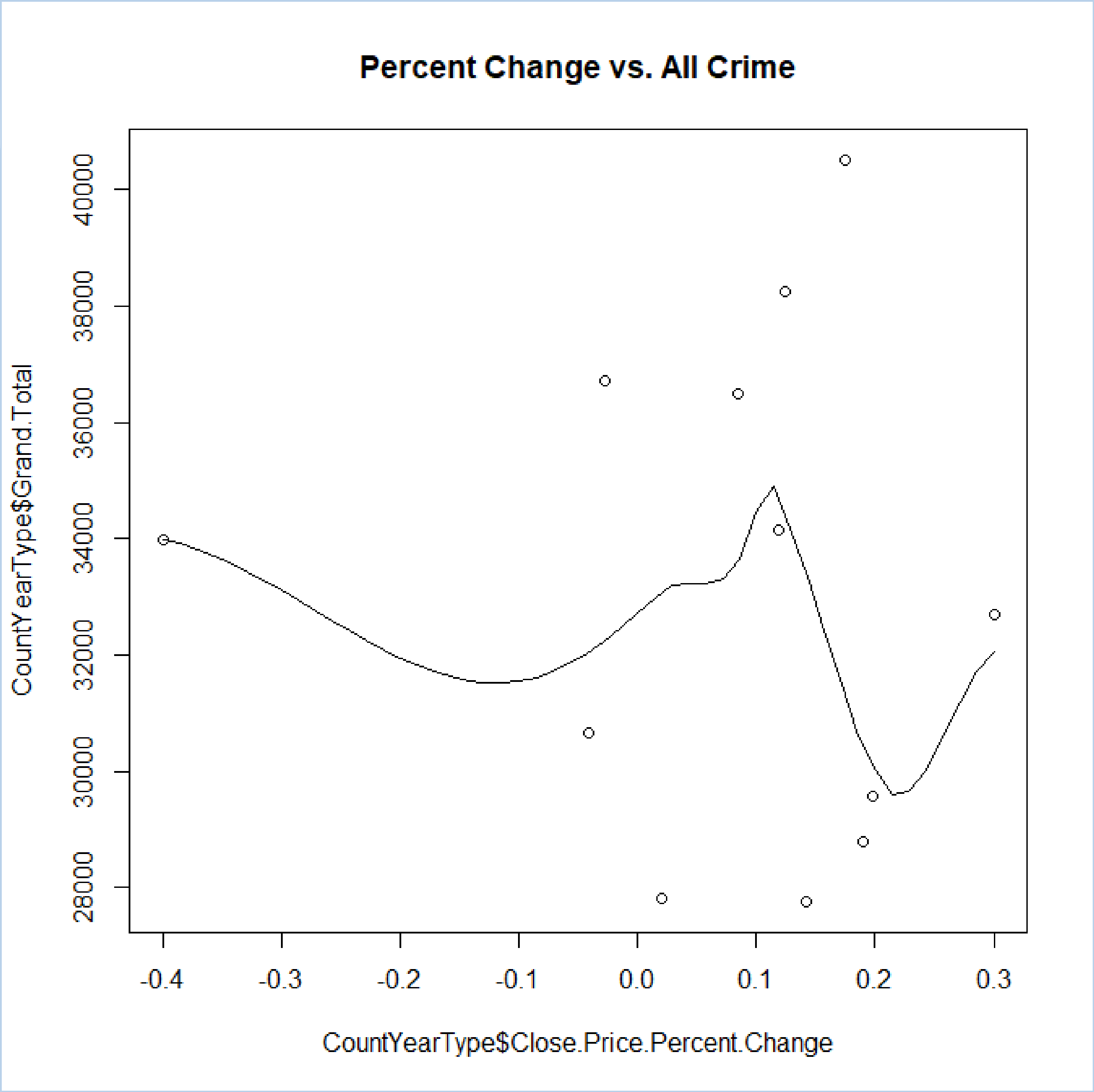
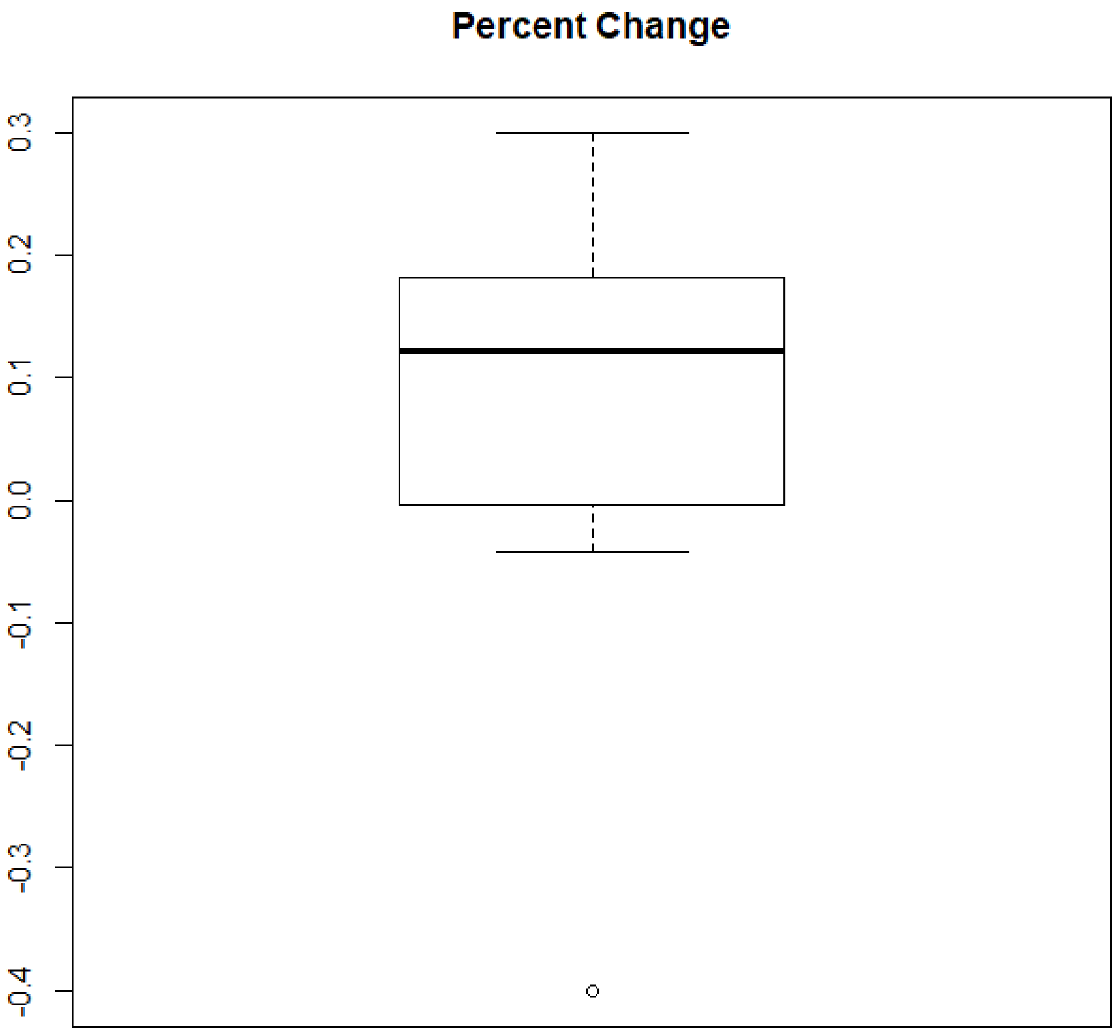


Figure 1. *This graph plots the percent change of the annual close price of the S&P 500 with the total count of financial crime for the corresponding year. Made using open source program developed by R Core Team (2017).*

When looking at the scatter plot, there was not a discernible pattern among the points that would lead me to use a certain type of regression. There appeared to be a poor correlation and a possible outlier in the S&P 500 Percent Change.

To investigate the outlier, I used R to create a box plot of the S&P 500 Percent Change in annual close price. In Figure 2 is the box plot of the annual percent change of the S&P 500 close price. There is a slight skew that indicates that the data is not a perfect normal distribution and the point that appeared to be a possible outlier in the scatter plot is identified as an outlier in the boxplot, which was the Percent Change of the Annual S&P 500 Close Price from 2009, which is likely to be an outlier due to the recession from the late 2000s.



Outlier rows: -0.400906768

Figure 2. *This plot is a box plot of the Percent Change of the Annual S&P 500 Close Price from 2007 through 2017. Made using open source program developed by R Core Team (2017).*

After removing the outlier and creating a second the scatter plot, there is still a lack of identifiable pattern that could lead to the identification of a regression to use on the data, as seen in Figure 3. The scatter plot in Figure 3 does not present a clear form that would determine possible regression (e.g. linear, exponential, polynomial, etc.) that could validly be performed.

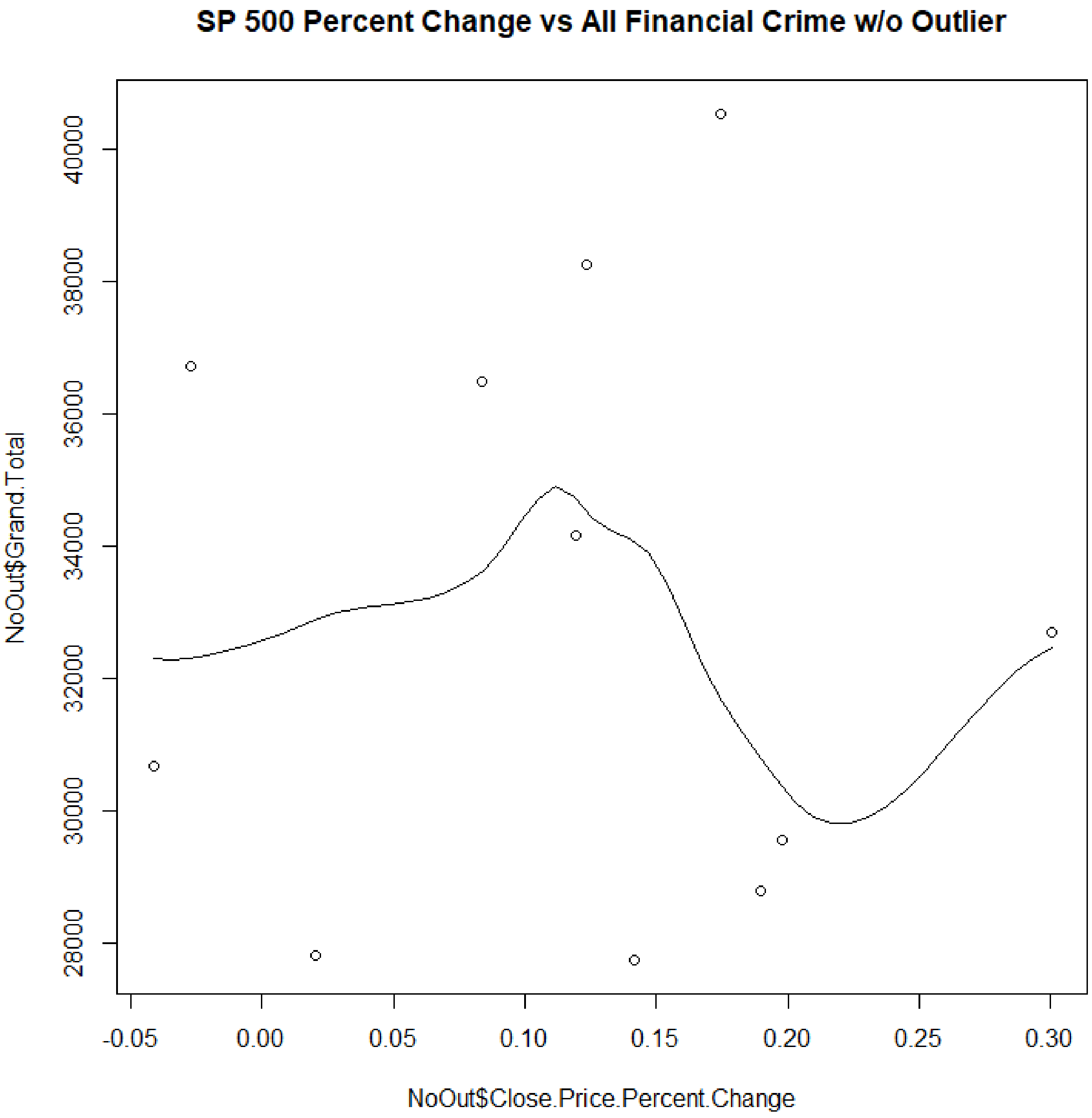


Figure 3. *This graph plots the percent change of the annual close price of the S&P 500 with the total count of financial crime for the corresponding year with the outlier from the S&P 500 percent change of annual close price removed. Made using open source program developed by R Core Team (2017).*

Although the correlation appeared poor, using R, I performed a linear regression to confirm my observations. The regression had a multiple R^2 of .0003095, which is a poor correlation since the model explains .03095% of the variation of the crime instances. In addition, the p-value was 0.9591 for the f-statistic of 0.002786 on 1 and 9 degrees of freedom, which implies randomness in the data points.

Discussion & Limitations

When annually analyzed, the financial crime incidents from Illinois, Texas, and Utah for the years 2007 through 2017 do not seem to correlate with the percent change of the annual S&P 500 annual close price. With such a high p-value, the implication is that when using the S&P 500 percent change in close price as the independent variable, the dependent variable, the number of occurrences of financial crime is more random in nature that predictable. Since I began with an annual-based analysis, the poor results may be attributed to the loss of detail of the data since monthly data can provide a more descriptive pattern. Since I summarized the data and condensed it into one row per year, it is likely that my analysis did not encompass enough detail to have meaningful results.

Conclusions and Future Work

In the future, I would like to delve deeper into the data and analyze on a monthly level instead of an annual scale. In addition, I think that a time-based analysis such as time regression would be more applicable to the data. I would also be interested in using a different independent variable such as the U. S. Treasury yield since it is highly used to predict other financial occurrences. In addition, expanding the states from three to all of the available states in the data would be intriguing.

References

Bakke, E. (2019). Predictive Policing: The Argument for Public Transparency. New York *University Annual Survey of American Law, 74*(1), 131–172. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=lgs&AN=136819438&site=eds-live&scope=site

Federal Bureau of Investigation Uniform Crime Reporting Program. (2018). *Illinois 2007-2017* [Data file]. Available from the National Incident-Based Reporting System (NIBRS) Crime Data Explorer

Federal Bureau of Investigation Uniform Crime Reporting Program. (2018). *Texas 2007-2017* [Data file]. Available from the National Incident-Based Reporting System (NIBRS) Crime Data Explorer

Federal Bureau of Investigation Uniform Crime Reporting Program. (2018). *Utah 2007-2017* [Data file]. Available from the National Incident-Based Reporting System (NIBRS) Crime Data Explorer

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016, November 01). Big data preprocessing: Methods and prospects. Retrieved November 27, 2018, from https://bdataanalytics.biomedcentral.com/articles/10.1186/s41044-016-0014-0

ICE Data Services. (2019). *S&P 500 (^GSPC)* [Data file]. Accessed through Yahoo! Finance. Retrieved from https://finance.yahoo.com/quote/%5EGSPC/history?period1=1104555600&period2=1546318800&interval=1mo&filter=history&frequency=1mo

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning*(8th ed.). New York, NY: Springer. doi:https://www-bcf.usc.edu/~gareth/ISL/ISLR Seventh Printing.pdf

Kte'pi, B. M. (2016). Data analytics (DA). *Salem Press Encyclopedia of Science.* Retrieved from http://ezproxy.columbusstate.edu:2048/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=113931286&site=eds-live&scope=site

Piegorsch, W. W. (2016). *Statistical data analytics: Foundations for data mining, informatics, and knowledge discovery.* Chichester, England : Wiley, 2016. Retrieved from http://ezproxy.columbusstate.edu:2048/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat06563a&AN=csu.9915190105802931&site=eds-live&scope=site

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.R-project.org/

S&P 500 (GSPC) Price History. (2019, December 9). Retrieved December 9, 2019, from https://wallmine.com/index/gspc.

Tableau Software (2018). Tableau Prep: A visual data prep software. Tableau Software, Seattle, Washington. Available from https://www.tableau.com/pricing/individual

Wienclaw, R. A. (2013). *Regression Analysis (Sociology).* Research Starters: Sociology (Online Edition). Retrieved from http://ezproxy.columbusstate.edu:2048/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=89185668&site=eds-live&scope=site