Summer 8-2020

Cyber-Assets at Risk (CAR): Monetary Impact of Personally Identifiable Information Data Breaches on Companies

Omer Ilker Poyraz
Old Dominion University, omerilkerpoyraz@gmail.com

CYBER-ASSETS AT RISK (CAR): MONETARY IMPACT OF PERSONALLY IDENTIFIABLE INFORMATION DATA BREACHES ON COMPANIES

by

Omer Ilker Poyraz
B.A. August 2012, Yeditepe University
M.B.A. August 2015, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

ENGINEERING MANAGEMENT AND SYSTEMS ENGINEERING

OLD DOMINION UNIVERSITY
August 2020

Approved by:

C. Ariel Pinto (Director)

Steven Cotter (Member)

Michael McShane (Member)

ABSTRACT

CYBER-ASSETS AT RISK (CAR): MONETARY IMPACT OF PERSONALLY IDENTIFIABLE INFORMATION DATA BREACHES ON COMPANIES

Omer Ilker Poyraz
Old Dominion University, 2020
Director: Dr. C. Ariel Pinto

Cyber-systems provide convenience, ubiquity, economic advantage, and higher efficiency to both individuals and organizations. However, vulnerabilities of the cyber domain also offer malicious actors with the opportunities to compromise the most sensitive information. Recent cybersecurity incidents show that a group of hackers can cause a massive data breach, resulting in companies losing competitive advantage, reputation, and money. Governments have since taken some actions in protecting individuals and companies from such crime by authorizing federal agencies and developing regulations. To protect the public from losing their most sensitive records, governments have also been compelling companies to follow cybersecurity regulations. If companies are unsuccessfully protecting their customers' records, they are levied by the government agencies. Companies also may face litigation from individuals after the breach. If the company is a public company, then it must provide more details about the incident.

Data breach incidents are one of the significant concerns that organizations have been experiencing for a while. Quantifying the data breach risk into monetary language is a problem that organizations still try to solve due to the unavailability of the data and indirect costs. The cost incurred by personally identifiable information (PII) data breaches may even exceed one billion dollars. Therefore, the monetary cost of a PII data breach is an essential phenomenon that organizations need to forecast and be prepared to mitigate the impact.

The purpose of this study is to identify the correlation between the dependent and independent variables and to develop a predictive model to quantify the monetary value of the PII data breaches with multiple regression.

This study introduces two new categories for personal information; these are PII and sensitive PII. This new taxonomy accentuates the impact of sensitive information, which is more costly than not sensitive personal information. Next, this study also presents significant results that demonstrate the correlations between revenue, PII, SPII, and class-action lawsuits, and the dependent variable, which is the total cost of the data breach. Also, specific models developed in this study are able to predict the responses for new observations.

"*Fate is in love with diligence*."

This thesis is dedicated to my beloved family, for their unwavering love and support.

ACKNOWLEDGMENTS

This dissertation would not be possible without the advising, mentoring, and encouragement of several individuals. They have contributed to this dissertation in different ways. Being a graduate student thousands of miles away from home is sometimes hard. Therefore, a supportive and funny environment encouraging family, friends, and professors is essential to successfully defend the dissertation. I found these throughout my dissertation within the Engineering Management and Systems Engineering department. I would like to thank Dr. Mustafa Canan, Dr. Unal Tatar, my friend Dr. Omer Keskin, Dr. Elnaz Dario, Miguel Toro, Goksel Kucukkaya, Murat Balci, Bora Aslan, Alican Kucukozyigit, Ahmet Aslan, Abdullah Dursun, and Engin Baris for their insights, coffee, and friendship.

I would like to thank Ms. Kim Miller for her candid conversation and warm suggestions to settle in the US.

I would like to thank Dr. Andres Sousa-Poza for guiding, supporting, and teaching me over the years. I will miss his lectures and philosophical ideas.

I would like to thank my committee members Dr. Steven Cotter and Dr. McShane, and my professor Dr. Adrian Gheorghe for their guidance and patience throughout the study.

Above all, I would like to express my gratitude to my advisor, Dr. Cesar Ariel Pinto, for guiding and supporting me over the years. When I wanted to quit the program, his wisdom, calm, and tolerance persuaded me to continue the Ph. D. program. He is an example of an adviser, mentor, and instructor. It was a complete privilege being his advisee.

I would like to acknowledge my former supervisors and colleagues from the Turkish Military Academy. Mainly, I would like to thank (Brig. Gen.) Murat Yetgin, Major Ender

Toydemir, Major Murat Sari, and Colonel Hamza Demirel give their priceless insights, sharing their experience, and encouragements.

Finally, my infinite blessings are to my mother, my father, my wife Elif, and my sisters for patiently being with me during my dissertation.

NOMENCLATURE

AIC	Akaike's Information Criterion
ALE	Annual Loss Expectancy
BIC	Bayesian Information Criterion
CIA	Confidentiality, integrity, availability
CIF	Critical Impact Factors
DoD	Department of Defense
EU	European Union
FAIR	Factor Analysis of Information Risk
FTC	Federal Trade Commission
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
ISRAM	Information Security Risk Analysis Method
ITRC	Identity Theft Resource Center
MFC	Mean Failure Cost
NIST	National Institute of Standards and Technology
OLS	Ordinary Least Square
PII	Personally Identifiable Information
PRC	Privacy Rights Clearinghouse
PRESS	Prediction Error Sum of Squares
RMF	Risk Management Framework
ROI	Return on Investment
SEC	Securities Exchange Commission

SRA	Society of Risks Analysis
SSN	Social Security Number
US	United States
VaR	Value at Risk
VIF	Variation Inflation Factor

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Overview

News of cybersecurity breaches against business or government organizations is becoming frequent in the media. Speed and capability of Information Technology services grow and provide people and organizations ease of use, convenience, and ubiquity. As a result, dependency on the Internet has been increasing and criticality and strength of dependency as well. However, even though cyber-defense mechanisms have been developing, carrying out cyber-attacks is becoming easier while the impact of those attacks has been drastically increasing (Ashford, 2018; Paganini, 2013). Many firms, critical infrastructures, and public services now operate in private, public, or hybrid clouds. Operating in cloud computing enables organizations to have ubiquity, communicate, and make transactions on-demand. However, the Internet and cloud computing also poses an opportunity for malicious actors.

Cyber-systems provide convenience, ubiquity, economic advantage, and higher efficiency to both individuals and organizations. However, vulnerabilities of the cyber domain also provide malicious actors with the opportunities to compromise the most sensitive information of people and organizations. Recent cybersecurity incidents (Ponemon, 2019) show that a group of hackers can cause a massive data breach, which eventually results in companies losing competitive advantage, reputation, and money.

Governments have since taken some actions in protecting individuals and companies from such crime by authorizing federal agencies, developing regulations, or issuing laws. To protect the public from losing their most sensitive records, which are kept in organizations' databases, governments have also been compelling companies to follow cybersecurity

regulations, purchase cybersecurity systems, and employ security experts. If companies are unsuccessful in protecting their customers' records, then they may be fined by the government. Companies also may face litigation from individuals after the breach. If the company is a public company, then it must provide more details about the incident in their yearly reports.

1.2 Purpose of the Study

The purpose of this study is to develop a model to identify the correlation between the monetary impact of personally identifiable information (PII) data breaches and the predictor variables and develop a predictive model to quantify the monetary value of the PII data breaches.

This study will categorize the information according to its criticality, high and low. The model aims to provide a better understanding of the monetary impact of a data breach from companies while they collect personal information. Also, insurance firms can have a better risk estimation of their insureds underwriting accurate cyber insurance premiums for data breach risk.

1.3 Problem Statement

Due to the nature of the cyber domain, any organization can suffer massive data breaches even though they have avant-garde tools. However, organizations may not fully comprehend the type of losses they can incur and are not able to forecast how much money they could lose. Therefore, they may lack an understanding of cyber risk, and as a result, they may not be prudently investing in cybersecurity. For strategic management, top-level management needs to understand the cyber risk in a language that they can speak. When a manager knows the cyber risk in monetary terms, it would be easier to make decisions such as accepting, transferring,

mitigating, or avoiding risk. However, quantifying the cybersecurity risk into monetary language is a problem that organizations still have not solved.

This study will only focus on PII data breach of cyber risk. Risk managers fail to assess the PII data breach risk in a way that people from technical to strategic level positions can communicate. As a result, the lack of understanding of data breach risk may lead to overlooking cybersecurity investment.

Existing data breach datasets do not have a standard structure. As a result, there is not a specific framework to conduct research. The particular impact of data breach incidents is recorded in terms of people; however, in some cases, it is recorded counting the number of data.

Data breach incidents are one of the significant concerns that organizations have been experiencing for a while. Quantifying the data breach risk into monetary language is a problem that organizations still try to solve due to the unavailability of the data and indirect costs. The cost incurred by PII data breach may reach hundreds of million dollars, which can severely affect an organization's financial health. Therefore, the monetary cost of a PII data breach is an essential phenomenon that organizations need to forecast and be prepared to mitigate the impact.

1.4 Research Questions

There are plenty of studies that provide a solution to estimate the technical impact of a cyber incident. However, there are very few studies that attempt to determine the monetary impact of a data breach – both direct and latent costs. A novel attempt will be made to improve the estimate of the monetary impact of a data breach.

The following questions are identified to frame this study:

1. What type of PII is stolen during data breaches?

2. What type of cost results from PII data breaches?

3. What is the monetary impact of PII data breaches?

4. What are the possible independent variables that are related to the cost of data breaches?

1.5 Significance of the Study

The proposed research has several contributions in the fields of cybersecurity and risk management. The contributions of the research are examined under the following categories: (a) Risk Management, (b) Security Economics.

a. Risk Management

The proposed research contributes to the field of risk management in two areas: risk analysis and risk communication.

Risk Analysis

Risk is that future event that yields negative impacts without regarding intent that includes software failures or accidents (Pinto & Magpili, 2015). According to the Society of Risks Analysis (SRA) (Aven et al., 2015), risk analysis is defined as “Systematic process to comprehend the nature of risk and to express the risk, with the available knowledge.” A general risk formula is (Pinto & Garvey, 2012):

Risk = f(probability, impact)

Consequences or impact are other concepts that are used interchangeably with "Impact." The proposed study will contribute to the impact section of the risk analysis of PII data breach.

Risk Communication

According to the SRA, risk communication can be defined as exchanging risk-related information among stakeholders (Aven et al., 2015). In general, decision making performed in three echelons: technical, operational, and strategic. In cybersecurity, at the technical level, capabilities of cybersecurity personnel depend on rapidly adapting existing knowledge into solutions in the complex cybersecurity domain. All the security operations, such as upgrading hardware-software, applying penetration tests, or employing anti-virus tools, can be accomplished by holding a high level of technical expertise. Operation al level decision-makers need to focus on legal, organizations, and the technical intersection of cybersecurity. Governments develop law, frameworks, or regulations to strengthen organizations' cybersecurity preparedness and compel them to follow specific rules and procedures. Therefore, operational level decision-makers need to specialize in these subjects. Decision-makers of the strategic level should be familiar with the impact of cyber threats to the business. Cybersecurity incidents may have a monetary impact on business, and it can jeopardize a business's goal, maximizing the profit.

Risk analysis can develop an intersection for all decision-making levels if a common perception of risk is developed. Developing a common language among echelons will provide a shared understanding and awareness of the risk. In this case, PII data breach can be assessed better if the impact of the PII data breach can be translated into a monetary language. Hence, the monetary impact calculation methodology of the proposed research will help the stakeholders understand and communicate the data breach risk better.

b. Security Economics

Cybersecurity Investment

Cyber risk has been a significant concern for businesses and is listed as one of the top five global risks with significant economic implications (World Economic Forum, 2016). Even some companies such as FICO started to rate cyber risk of businesses, which is now considered in investment decision-making (Lawrence, 2014). Chief Information Security Officers undertake a more critical role in the company's board of management as they are not only accountable for keeping organizations secure from cyber-attacks. Also, they guide member of the strategic management considering the effectiveness and efficiency of cybersecurity investments. Companies invest in cybersecurity are effectively and efficiently observed to have less data breach costs (Ponemon, 2019). To sum up, cyber risk management has become an emerging and vital part of the enterprise risk management. Data breach risk is one of the significant risk items of cyber risk; therefore, understanding the data breach risk and making efficient and effective cybersecurity investment will reduce overall cyber risk and overall cost.

Since the proposed method will help to calculate the monetary impact of the PII data breaches, top-level managers can make better decisions to manage the data breach and choose the most economically profitable risk management strategy (i.e., acceptance, avoidance, transfer or mitigation).

Cyber Insurance

Risk transfer is another option to handle cyber threats. The cyber insurance market has been growing all over the world. The total cybersecurity insurance market in the United States was $3.1 billion in 2017 (Matthews, 2018). In addition to companies, some cities like Dallas, San

Diego, Denver, and Detroit already have cyber insurance to mitigate the cost of after a cybersecurity incident (Calvert & Kamp, 2018).

One of the main issues of cyber risk insurance is the lack of ability of accurate data breach risk calculation, particularly in monetary terms. The impact of the PII data breach model also provides a solution to the underinsurance problem in data breach risk.

CHAPTER 2

BACKGROUND OF THE STUDY

2.1 Introduction

Cyber-crime is criminal activity carried out using computers and the Internet, including downloading intellectual property, identity theft, hacking, or web defacement (Christensson, 2006). Cyber-crimes may cause monetary loss, reputation loss, and business interruption.

Cyber-crime has been carried out since 1973 when a teller at New York's Dime Savings Bank used a computer to embezzle $2 million (Wavefront, n.d.). Since then, the attack sophistication and financial impact of cyber-crime have been increasing.

There have been studies to assess the cyber-risk, such as measuring the financial impact of data breaches or determining the optimal amount of cybersecurity investment. Also, there have been studies to develop standards, frameworks, and regulations to mitigate cyber risks.

2.2 Cybersecurity Risk Management

What is Risk?

Risk is the potential of undesired negative impacts on human life, property, or the environment based on the probability and the impact of the event (Gratt, 1987). Another definition is that risk is that future event that yields negative impacts without regarding intent that includes software failures or accidents (Pinto & Magpili, 2015).

Risk can be formulated as follows (Pinto & Garvey, 2012):

$$\text{Risk} = f(\text{probability}, \text{impact})$$

Kaplan (1997)offered three questions and contributed by (Haimes, Kaplan, & Lambert, 2002) to develop a risk analysis framework. A precursor question is offered in addition to those questions (Pinto, McShane, & Bozkurt, 2012):

0. What are the popular events?
1. What can go wrong?
2. What are the consequences?
3. What is the chance of occurrence?
4. What can be done to manage them?
5. What are the alternatives?
6. What are the effects on future decisions?

Risk Analysis is described by the Society of Risk Analysis as *"Systematic process to comprehend the nature of risk and to express the risk with the available knowledge."* (Aven et al., 2015).

What is cyberspace?

NIST defines cyberspace as "A global domain within the information environment consisting of the interdependent network of information systems infrastructures including the Internet, telecommunications networks, computer systems, and embedded processors and controllers." (Kissel, 2013).

Cybersecurity Actors

There are different types of actors in the cyber domain. Anyone of them can, intentionally or accidentally, cause the unavailability of the service or data breach. Those actors are: Functional users: individuals or organizations for whom the cyber system was meant to be useful.

Security experts: individuals or organizations generate strategies, defense tools, products, and techniques against hackers and malware.

Hackers: individuals or groups use their skills to gain benefits through hacking people, organizations, or governments.

Insiders: employees of an organization may reveal administrative details to hackers or themselves be able to disrupt their organizations' operation as an act of revenge, i.e., disgruntled insider.

Penetration Testers: check security vulnerabilities of web-based applications, networks, and systems with the permission of that organization.

Organized crime: a group of criminals that target victims to demand money and extort information.

Hacktivist: an individual or a group that carries out cyberattacks to draw attention to humanitarian or global problems such as human rights, freedom of speech, or global climate.

Cyber-terrorist: a group of hackers organizes a cyber-attack to cause alarm, fear, or panic with a political agenda.

Competitors: sometimes, a competitor can be the sponsor of an attack, such as hiring a hacker group to conduct a distributed denial-of-service attack to disrupt competitors' service to damage its reputation.

Law enforcement: organizations like INTERPOL, Department of Homeland Security, or National Security Agency monitor cybercrimes.

Nation-States: an attack carried out by state-sponsored hackers.

Cybersecurity Risk Management

Cybersecurity risk management is concerned with risks caused by cyber threats. Cyber risk is caused by a cyber threat (Refsdal, Solhaug, & Stølen, 2015). Cybersecurity risk can be defined as "operational risks to information and technology assets that have consequences affecting the confidentiality, availability, or integrity of information or information systems." (Cebula & Young, 2010). An example of cyber risk is caused by a cyber threat like a virus or denial of service attack.

Confidentiality, integrity, and availability (C-I-A) are three main objectives of cybersecurity, and any incident can have a consequence on each of these objectives or their combinations. In cybersecurity, incidents are categorized by C-I-A objectives. Cybersecurity Risk management also focuses on ensuring these three pillars of cybersecurity (confidentiality, integrity, and availability) by assessing and minimizing risk.

CIA principles are summarized below (Pinto, 2018):

Confidentiality: the principle prevents illegitimate users from accessing the information on a computer or network. Confidentiality breaches cause disclosure of the data to illegitimate users.

Integrity: it ensures that unauthorized actors cannot adjust or destroy information. If changed, it can be found out.

Availability: it enables that only authorized users can access to service or information.

2.3 Information Security Risk Assessment

Information security risk assessment is a primary element of an information security management system that measures the effectiveness of the current security controls to detect vulnerabilities and threats. Then, decide which safeguards to choose to address potential threats

(Landoll, 2011; Shameli-sendi, Ezzati-jivan, Jabbarifar, & Dagenais, 2012). There are two types of risk assessment methods; qualitative and quantitative.

2.3.1 Qualitative Methods

Qualitative risk assessment is measuring an event or regulatory control or security to understand the quality of the operation (Hillestad, 2018). Qualitative risk assessment can be easy and rapid to implement. Assessment of the risk is highly dependent on the assessor background, perception, and environment of the organization. Therefore, a qualitative risk assessment may become biased or subjective. Nevertheless, it is still essential and useful to assess the information security risk.

Although the National Institute of Standards and Technology (NIST) developed a framework for critical infrastructure, any organization can apply the guideline regardless of the size or severity of cybersecurity risk. The Framework empowers organizations to implement risk management best practices and the principles to have robust security and resilience. It considers the cybersecurity risk as part of organizational risk and overall risk management process. The Framework comprises three parts: The Framework Core, the Implementation Tiers, and the Framework Profiles. The Framework Core consists of a set of cybersecurity activities, outcomes, and informative references. Elements of the Core deliver comprehensive guidance for creating individual, organizational Profiles. With the use of Profiles, an organization can arrange and prioritize cybersecurity activities with its business/mission requirements, risk tolerance, and resources. The Tiers offer an instrument for organizations to assess the features of their approach to managing cybersecurity risk (NIST, 2018).

Department of Defense (DoD) has developed a Risk Management Framework (RMF) (DoDI 8010.01) to apply a risk-based approach to cybersecurity implementation, assessment,

decision making, and monitoring. According to DoD (2015), cybersecurity is regarded as a risk-based activity with a mission-driven approach. The RMF is based on the NIST SP-800 series documents. It brings a new method to federal organizations to assess their cyber risk to avoid any mission disruption. DoD RMF has twenty security controls, which are assessed by experts as *"satisfied"* or *"other than satisfied."* DoD RMF is an iterative framework consisting of six steps (DoD, 2015):

- categorize the information system,
- select security controls,
- implement security controls,
- assess security controls,
- authorize the information system
- monitor

Jones (2007) proposes a risk management framework for the management of information risk. He discusses risk management steps and presents a framework that can be employed to develop management structures that can be tried for their efficacy and generality.
Another paper presents an information risk management framework for a better understanding of critical areas of focus in a cloud computing environment to identify threats and vulnerabilities covering cloud service and deployment models (Zhang, Wuwong, Li, & Zhang, 2010). It follows seven processes of Information Security Risk Management.

Organizations have limited resources to protect their assets. Therefore, they should prioritize their assets according to their importance. Scholars offer to utilize a conceptual framework in which security requirements are related to the organization's unique business drivers (Su, Bolzoni, & Van Eck, 2006). The Framework has three parts:

1. Business vision: high-level business goals
2. Critical Impact Factors (CIF): impact of the security violation on business
3. Valuable assets and their security requirements are inventories of security requirements.

They suggest:

- Enumerate useful assets and their security requirements
- Define the organization's CIF and business vision
- Link the security requirements with CIF and business vision

Business impact analysis identifies the organization's critical business function and defines the impact of external and internal CIFs on the various parts of the organization. Figure 1 illustrates the concept offered by the authors (Su et al., 2006).

Figure 1. Business Impact Analysis

2.3.2 Quantitative Methods

Quantitative risk assessment considers factual and measurable data that is based on calculations. Quantitative risk assessment usually considers the impact of the incidents on

economic loss and operational loss and calculates probability. It is more objective and has more generalizability than qualitative risk assessment methods.

Cybersecurity investment or, in a broader sense, the economics of information security has been studied for a long time. Cybersecurity has become a critical investment section, and it has attracted the attention of many industry practitioners and scholars. Therefore, there have been plenty of studies to determine the optimal amount for cybersecurity investment. Scholars suggest different methods to help decision-makers on how much to invest in cybersecurity to protect operational excellence and intellectual property. Certain significant studies focused on cybersecurity investment are summarized below.

Scholars developed and implemented different optimization models on the economics of cybersecurity using game theory, optimization theory, and security controls selection. One of the earlier works (Gordon & Loeb, 2002) utilized optimization to calculate the optimal amount to invest in cybersecurity. They claim that a small fractional amount (37%) of the expected damage loss would be enough to invest in cybersecurity. Lam (2015) employs optimization with a regulatory perspective rather than an enterprise-level analysis. He suggests that the vendor should not burden the full liability of the compromise; instead, it should be shared between the seller and the consumer.

Game theory and optimization are used to compare the two methods of benchmarking the efficiency of cybersecurity investments (Cavusoglu, Raghunathan, & Yue, 2008; Fielder, Panaousis, Malacaria, Hankin, & Smeraldi, 2016).

A Table Top Approach is taken to evaluate the impacts of cyber intrusion events and the benefits of safeguards investments (Garvey, Moynihan, & Servi, 2013). The tabletop approach is designed to *"place light demands on the granularity of inputs"* to analyze the impacts of cyber-

attack events and the perks of cybersecurity investments. The authors merge the Multi-criteria risk and decision-analytic approach and Pareto optimal economic return to estimate the investment amount derived from the impact of cybersecurity incidents and merit points of safeguards.

A risk management approach is suggested for assessing information security products (Arora, Hall, Pinto, Ramsey, & Telang, 2004). They point out that security managers need to consider the risk-based return on investment method to determine how much to invest in cybersecurity due to higher uncertainties in the cyber domain.

A survey-based quantitative approach Information Security Risk Analysis Method (ISRAM) is developed to analyze the security risks of information technologies (Karabacak & Sogukpinar, 2005). The method has seven steps; awareness of the problem, listing and weighing the factors, converting factors into questions and answers, preparation of risk tables, conduction the survey, application of formula and obtaining a single risk value, and assessment of the results. Also, ISRAM is appropriate to calculate the monetary value of cyber risk by using annual loss expectancy (ALE).

Schneier (2008) considers security, not an investment but loss prevention. A company should spend money only on the worth of the problem, not more than that. ALE calculates the cost of a security event in both tangibles and intangibles. Then, it multiplies that by chance, the event will occur in a year. He suggests doing not solely rely on Return on Investment (ROI) or ALE analysis. That gives the amount to spend to mitigate the risk. Challenges for cybersecurity ALE:

- Lack of data on the incident
- High uncertainty and rapid change in the cyber domain

- Extreme and rare events

A user-centered cloud computing risk analysis is explored (Rabai, Jouini, Aissa, & Mili, 2013). They propose a security metric that quantifies the cloud risk for providers and subscribers in economic terms by using mean failure cost (MFC).

A group of scholars discuss the capability of insurance for cyber risk management (Biener, Eling, & Wirfs, 2015). They work on 944 cases of cyber compromise from the operational risk database and assess their statistical outputs. They underscore that cyber risk has unique characteristics and problems due to a lack of available data and information asymmetries.

Bayesian Generalized Linear Models are developed by using the Privacy Rights Clearinghouse (PRC) dataset to study the patterns in data breaches (Edwards, Hofmeyr, & Forrest, 2016). They claim that the size or frequency of data breaches has not increased. However, they see that the heavy-tailed statistical distributions explain the increase. Also, they state that the log-normal distributions better model the size of data breaches.

Cyber costs are identified from an operational risk database and assess these with statistics and actuarial science methods (Eling & Wirfs, 2019). They use 1,579 cyber risk cases from an operational risk dataset. They employ the peaks-over-threshold technique from extreme value theory. Their models can be used to generate reliable risk evaluations based on country, industry, size, and other factors.

Some researchers suggest employing ROI to calculate the optimal amount of information security. Clifton (2015) recommends that the ROI determination should be relevant to the factors associated with the risk of a cybersecurity incident. Several risk assessment organizations use the Cyber Value-at-Risk concept adapted from finance (Sanna, 2016). Cyber VaR model provides a

foundation for quantifying information risk and insert discipline into the quantification process. Cyber VaR models apply probabilities to estimate likely losses from cyber-attacks during a given time-frame. The goal of Cyber VaR modes is two-fold (Sanna, 2016):

- Assist risk and InfoSec professionals to articulate cyber risk in monetary terms
- Empower Chief level managers to make cost-effective decisions and balance between securing the organization and running the business

VaR modeling is a statistical methodology employed to quantify the level of financial risk within an organization or investment portfolio over a specific time frame. VaR is calculated in three variables:

- The amount of potential loss
- The probability of that amount of loss
- The time-frame

2.4 Monetary Impact of Data Breach on Companies

This section will introduce the definition of personally identifiable information, sensitive personally identifiable information, the monetary impact of massive data breaches, summary of literature review, and existing data breach cost models.

Definition of PII and Sensitive PII

The data will be categorized and used in the model is as set forth by the definition of the Department of Homeland Security PII and sensitive PII (SPII). Government or private organizations collect, store, or transfer the data of people's name, address, social security number, driver's license number, mother's maiden name, usernames and passwords, and

credit/debit card and so on. The loss or theft of SPII can result in embarrassment, inconvenience, reputational harm, emotional harm, financial loss, unfairness, and personal safety in danger. DHS (2017) defines personal information as “Personally Identifiable Information” which means that “any information that permits the identity of an individual to be directly or indirectly inferred, including any other information that is linked or linkable to that individual, regardless of whether the individual is a US citizen, legal permanent resident, a visitor to the US, or employee or contractor to the Department."

Sensitive PII is “personally identifiable information which, if lost, compromised or disclosed without authorization, could result in substantial harm, embarrassment, inconvenience or unfairness to an individual (DHS, 2017).″

Examples of PII and SPII are provided below (DHS, 2017; STIP, 2018; WDPI, n.d.):

Table 1. Examples of PII

PII
Name
Account name/ user ID
Password
Email
Address
Telephone number
Education credentials/certificates
Date/place of birth
Vehicle title number

Table 2. Examples of SPII

SPII
Social security numbers
Medical history
Credit/ debit card numbers
Driver's license numbers
Bank account numbers
Passport numbers
Alien registration numbers
Biometric identifiers
Taxpayer identification number

Also, certain information in-combined may pose more threat than standalone. For example, name, zip code, and credit card information may be more sensitive when combined than apart. The most recent data breaches provide more details on the number and type of stolen data. Table 3 provides the details of the Equifax data breach (Owens, 2018).

Table 3. Aftermath of Equifax Data Breach

Stolen data type	Number of records
Name	147 million
Date of birth	147 million
Social security number	146 million
Address	99 million
Gender	27 million
Phone number	20 million
Driver's license number	18 million
Email address	2 million
Credit card number	209,000
Tax ID	97,500
Driver's license state	27,000

Data Breach

A data breach is a cybersecurity incident that causes intentional or unintentional disclosure of data. Exposed data may include personal health information, personally identifiable information, blueprints, intellectual property, or state secrets. The finance industry is one of the primary targets of malicious actors because malicious actors can get credit card numbers, account numbers, or social security numbers (Ponemon, 2019). Since those data are stored in the cyber domain, maintaining the security of data is a must for companies.

Data breaches are the most common cyber incident based on his work on the Advisen dataset (Romanosky, 2016). He found credit cards and medical information were the most stolen data from organizations. As a result, those organizations are more likely to face litigation from individuals.

Data breach risk is a significant concern for organizations that operate in the cyber domain. Cyber systems are now crowded with criminals, hackers, government actors, hacktivists, and other adversaries. Media attention to cybersecurity issues has grown dramatically over the past several years as well. In 2013, about 40 million credit and debit cards were stolen from Target's point of sale terminals (Krebs, 2014). The following year, details on 56 million credit cards were stolen from Home Depot in a similar attack. In February 2015, personal information from about 80 million people was taken from the healthcare company Anthem (Krebs, 2015). Equifax suffered a massive data breach that lost nearly 146 million customers' SSN, passports, or driver's licenses (Johnson, 2018). These examples are massive data breaches. Therefore, the cost may easily exceed $100M. However, on average, the loss per data breach that a compromised company suffers between $2.1M and $3.8M (Eling & Schnell, 2016). Ponemon Institute continuously conducts a study in the cost of data breaches. Its latest report states that the average total cost of a data breach is $3.92M, and the cost per lost record is $150 (Ponemon, 2019).

Massive data breaches may cause catastrophic damages not only compromise of data but also businesses shutdown resulting in unemployment, legal fees, loss of customer trust, loss of revenue, or a decrease in stock price. However, it is hard to monetize poor public relations, loss of future income, and the value of cyber-assets. We can easily find out the visible or direct costs, such as credit monitoring, investigation, government fee, or litigation. Table 4 shows the victim

organizations, the number of affected people, and how much money they had to spend after the incident. Examples of the massive data breach and its consequences are provided in the table.

Table 4. Infamous Massive Data Breaches

Company/Organization	Number of Affected People	Total Cost
Sony PSN	77M	$193M
Target	40M	$310M
Yahoo	500M	$502M
Equifax	146M	$1,445M
Home Depot	56M	$340M
Uber	57M	$148M
Anthem	80M	$406M

2.4.1 Measuring the Impact of the Data Breach

There are very few practical studies that quantify the monetary value of data breaches. Relevant literature is summarized below.

Romanosky (2016) employed the data collected by Advisen and developed a formula to associate the factors with the monetary impact of a data breach. The mean loss for a data breach is $5.7M. However, the median is $170K.

The data breach impact formula:

$$log\ (cost_{it}) = \beta_0 + \beta_1 * log\ (revenue_{it}) + \beta_2 * log(records_{it}) + \beta_3\ repeat_{it} + \beta_4 * malicious_{it} + \beta_5 * lawsuit_{it} + \alpha * FirmType_{it} + \lambda_t + \rho_{ind} + \mu_{it}$$

The explanation of the variables:

Cost: the total cost of the incident

Revenue: firm's revenue

Records: the number of compromised records

Repeat: binary variable code; 1 if the firm suffered multiple events, and 0 otherwise.

Malicious: a binary variable; 1 if the event was caused by malicious intent

Lawsuit: is a binary variable code; 1 if a legal action resulted

FirmType: a vector of binary variables describing the firm was government agency, nonprofit, private or publicly traded company

λ_t: vectors of years

ρ_{ind} : industry binary variable

μ_{it} : error term supposed to be uncorrelated with the covariates

Romanosky (2016) states that a 10% increase in revenue would increase the cost by 1.3%. He also founds a strong correlation of the number of records compromised with the loss. A 10% increase in the number of records compromised would increase the cost by 2.9%. The R^2 is 0.46.

Another model is developed by using the Ponemon Cost of Data Breach 2014. The author tries a Linear Regression model but concludes that the linear model is inadequate and perform a log-log regression (Jacobs, 2014):

R-Squared is 0.5, and the model is:

Log (impact) = 7.68 +0.76 * log (records)

In this model, he claims that a 10% increase in the lost number of records causes a 7.6% increase in the cost of the data breach.

Another scholar suggested a cybernomics concept to measure the cyber risk. However, her primary intent is to quantify the cyber-assets into the monetary term. She suggests employing MicroMort and VaR to quantify the risk and categorizes the cyber assets (Ruan, 2017).

- Digitized assets
- Assets born-digital
- Operational assets

The total value of cyber assets can be calculated as:

$$V = \sum_{i=1}^{Nc} CVi \; + \sum_{j=1}^{No} OVj$$

V: the total digital value of entity E

CV: the value of core asset c of entity E

OV: the value of the operational asset of O entity E

N_c: the number of core value assets in entity E

N_o: the number of operational assets in entity E

Another study is carried out to calculate the tangible costs of data breaches using two case studies, focusing on a salary guide and ballpark estimation of the work hours of the people who were involved in managing the data breach (Layton & Watters, 2014). They forecast labor costs regarding them as the only tangible cost. Regarding intangible costs, Layton and Waters only consider the loss of reputation. It is interesting to note that they argue that the stock price was not negatively impacted after the announcement of a data breach in the two cases considered.

Kuypers (2017) developed a total cost of cyber incidents at an organization per year. He runs a Monte Carlo simulation in which the data comes from historical events and scenarios. He modeled different attack types, their frequencies, and their impacts. However, his model takes privacy information loss as a variable in the equation. The total cost of each incident is gathered by adding each impact category; investigation cost, direct costs, business interruption, reputation damage, credit monitoring, and loss of intellectual property.

Factor analysis of information risk (FAIR) is a cyber risk assessment approach and provides a well-reasoned and logical assessment framework (Freund & Jones, 2015). The main focus is developing probabilities for frequency and magnitude of confidentiality, integrity, or availability breach.

Ponemon Institute consistently conducts data breach surveys and publish the results as in their reports. The report provides a sample from the populations and gives a perspective by describing the facts such as the number of records exposed, total-average-per record data breach cost. According to the latest Ponemon (2019) data breach report,

- The average total cost of a data breach: $3.92 million
- The average size of a data breach: 25,575 records
- Cost per lost record: $150
- The Healthcare industry has the highest average cost of a data breach: $6.45 million.

However, the report does not provide any model or prediction; instead, it only provides descriptive statistics of the sample.

Related literature on the monetary impact of a data breach is tabulated and presented in Table 5. The table is divided into seven sections: "Source" column indicates the article's

reference; "Qualitative" and "Quantitative" columns indicate the approach; "Monetary Impact" column indicates if the study covers the monetary aspect of the cyber risk; "Type of Cyber Risk" column shows the type of cyber risk if it is overall or specific; "Data Classification" column shows if the study classified the data; "Model/Method" column explains the proposed method.

Table 5. Summary of Literature Review

Source	Qualitative	Quantitative	Cost	Type of Cyber Risk	Data Classification	Model/Method
(Gordon & Loeb, 2002)		x	x	overall		Gordon-Loeb model
(Arora et al., 2004)		x	x	overall		RMF and ROI
(Karabacak & Sogukpinar, 2005)	x	x	x	overall		ALE
(Su et al., 2006)	x			overall		Framework
(Jones, 2007)	x			overall		Framework
(Schneier, 2008)		x		overall		ALE
(Cavusoglu et al., 2008)		x	x	overall		Game theory
(Zhang et al., 2010)	x			cloud		Framework
(Rabai et al., 2013)		x		cloud		MFC
(Garvey et al., 2013)		x	x	overall		Multi-criteria risk and decision analytics
(Jacobs, 2014)		x	x	data breach		Linear regression
(Biener et al., 2015)	x	x	x	overall		Statistical tests
(Clifton, 2015)		x	x	overall		ROI
(Lam, 2015)	x			overall		Optimization
(Freund & Jones, 2015)	x	x	x	overall		FAIR
(Romanosky, 2016)		x	x	data breach		Multiple regression
(Fielder et al., 2016)		x	x	Overall		Game theory
(Sanna, 2016)		x	x	overall		Cyber VaR
(Kuypers, 2017)		x		overall		Statistical tests
(Edwards et al., 2016)		x	x	data breach		Bayesian Generalized Linear Model
(Ruan, 2017)		x	x	overall		Cybernomics/ MicroMort and VaR
(Eling & Loperfido, 2017)		x		data breach		Statistical tests
(Eling & Wirfs, 2019)		x	x	overall		Actuarial models
(NetDiligence, 2018)			x	data breach		Descriptive
(Ponemon, 2019)			x	data breach		Descriptive

2.4.2 Impact on Stock-prices

The impact of cybersecurity incidents on public companies has been a point of interest for scholars for a long time. Most of the researchers find out that cybersecurity incidents have temporary negative impacts on stock prices. However, it is hard to reach a robust conclusion due to a lack of the number of incidents, categorization per industry, and elimination of factors other than cybersecurity incidents. Nevertheless, the studies are shown in the table below state that data breaches have a negative impact on stock prices. Also, SEC may issue fines to the public companies that fail to comply with the data security regulations and fail to notify the customers.

The impact of cybersecurity incidents on stock prices may increase in time as it appears more in mass media, and governments issue higher fees than before. The Securities Exchange Commission (SEC) enforces public companies to reveal the cost of a data breach. However, SEC compels the public companies since 2018. Before 2018, companies were not supposed to share the total cost in their yearly reports. A literature review of the impact of data breaches on stock prices is illustrated below in Table 6. The table shows the number of events, timeframe, and event windows.

Table 6. Impact of Data Breaches on Stock-prices

Author	number of events	sample period	Event windows
(K. Campbell, Gordon, Loeb, & Zhou, 2003)	43	1995-2000	[-1,1]
(Ko & Dorantes, 2006)	19	1997-2003	subsequent four quarters
(Goel & Shawky, 2009)	168	2004-2008	[-2,1]
(Bolster, Pantalone, & Trahan, 2010)	93	2000-2007	[-1,0] [-1,1] [1,30]

(Gatzlaff & McCullough, 2010)	77	2004-2006	[0,1] [0,35]
(Yayla & Hu, 2011)	123	1994-2006	[-1,1] [-1,5] [-1,10]
(Modi, Wiles, & Mishra, 2015)	146	2005-2010	[-2,2]
(Hinz, Nofer, Schiereck, & Trillig, 2015)	6	2011-2012	[0,1] [0,2] ,[0,3]
(Schatz & Bashroush, 2016)	50	2005-2013	[-121,-3], [-2,2]
(Poyraz, Serttas, Keskin, Tatar, & Pinto, 2018)	27	2006-2018	[-7, -3, -1,0,1,3,7]

2.5 Cyber-cost Taxonomy

Several studies shed light on the types of costs incurred after cyber-attacks and data breaches. Cyber-cost has two categories that are direct and indirect costs. The cyber-cost taxonomy is summarized in table 7 (Kopp, Kaffenberger, & Wilson, 2017; Kuypers, 2017; NetDiligence, 2018; Ponemon, 2019; Romanosky, 2016).

Table 7. Cyber-cost Taxonomy

Phase	Direct Costs	Indirect Costs
Prevention (continuous)	• Safeguards • Regulatory compliance cost	Opportunity cost
Reaction (immediate)	• Technical investigation • Stop intrusion and initiate the recovery of systems	• Cost of operational disruption • Opportunity costs • Loss in revenue • Loss in equity value
Impact management (short-term)	• Adjustment to infrastructure and processes • System and data recovery • Damage reduction • Post-breach customer protection • Initiation of cyber audit • Attorney and litigation cost	• Opportunity costs • Loss in revenue • Loss in equity value • Customer loss
Business recovery and remediation (medium to long term)	• Credit monitoring • Class action lawsuits	• Increased funding costs

	• Fees • Penalties • Discounts for future products and services	• Lower future demand for breached firm's services • Redesign of business processes and systems • Rebuilding relationships, reputation and brand value • Investment in better security systems and preparedness capabilities

2.6 Data Protection Frameworks, Regulations, and Guidelines

State and federal governments have been working on improving the data breach legislation in favor of citizens. Companies must notify the customers within the given time-frame and take steps to reduce the impact of the breach (DWT, 2018). There are specific regulations and frameworks for industries to follow the Federal Trade Commission (FTC), Securities and Exchange Commission (SEC), General Data Protection Regulation (GDPR), and Health Insurance Portability and Accountability Act (HIPAA).

2.6.1 Federal Trade Commission Data Breach Guide

FTC is one of the major government organizations that monitor data breach or privacy rights violations. It is authorized to issue fines and enforce companies to follow specific procedures. FTC recommends five fundamental principles to have a robust data security plan (FTC, 2016):

1. Take stock: know what personal information is held have in inventory.
    a. Check who sends sensitive information
    b. Check how personal information is received
    c. What kind of information is collected

   d. Where the information is kept

2. Scale down: keeping only what is required, do not collect the sensitive information that is not needed
3. Lock it: protect the information that is kept
   a. Ensure physical security
   b. Ensure electronic security
   c. Ensure cybersecurity
   d. Conduct continuous cyber risk assessments
   e. Employee training
   f. Monitor third-party risks
4. Pitch it: properly dispose of what is no longer needed
5. Plan ahead: create a plan to respond to security incidents

2.6.2 Securities and Exchange Commission Guidance on Cybersecurity

SEC issued about the cybersecurity incident disclosure of public companies. The statement adds two rules to its previous guidance in 2011 that are (SEC, 2018):

- establishing and maintaining appropriate and effective disclosure controls and procedures
- prohibiting company personnel from insider trading before appropriate disclosure of cybersecurity incidents

Disclosures should include:

- Frequency of cyber events, based on experience
- Probability and magnitude of incidents (costs, in financial terms)
- Adequacy of controls

- Third-party suppliers' risk
- Amount of insurance coverage
- Potential reputational harm
- Relevant laws and regulations
- Potential fines and judgments from cybersecurity incidents

Cybersecurity controls and procedures should identify cybersecurity risks, incidents, impacts on business. Also, there should be a shared understanding between technical experts and disclosure advisors to provide timely disclosures of risks that may not yet have been the target of a cyber-attack and incidents.

Nevertheless, SEC does not recommend making detailed disclosures of the cybersecurity incidents or system features to prevent malicious actors from penetrating the organizations' security. The guidance aims to protect investors' interest by keeping them knowledgeable about the companies that investors put money.

Because cybersecurity incidents yield monetary loss such as investigation, breach notification, or loss of revenue, these financial impacts should be incorporated into companies' financial statements.

A failure to appropriately disclose the cybersecurity incidents may yield more financial damage to companies. For example, Yahoo was fined $35 million by SEC for under-reporting its cybersecurity incidents (Michaels, 2018).

2.6.3 General Data Protection Regulation

General Data Protection Regulation (GDPR) has been issued by the European Union (EU) in 2016 and has been active since 2018. The goal of the GDPR is to ensure the privacy of all EU citizens. All companies operating in EU borders are subject to GDPR regardless of the company's location. Therefore, GDPR applies to companies that are not located in the EU, too. Failure to complying with GDPR can be fined up to 4% of annual global turnover or €20 million (whichever is greater). All member states are required to notify data breaches where a data breach can "result in a risk for the rights and freedom of individuals." Data breach notification must be done within 72 hours of first having become aware of the breach. Critical items of GDPR:

- Consent be freely given
- New individual rights given to data subjects
- Data protection impact assessments for large scale processing
- Notifying within in 72 hours of the data breach
- Appointment of Data Protection Officers to manage privacy framework
- Considering privacy in developing business processes and new systems
- Accountability for personal data

There are two different types of data-handlers the GDPR applies to: "processors" and "controller." Whereas a controller is a "person, public authority or another body which determines the purposes and means of the processing of personal data," a processor is a "person, public authority, agency or another body which processes personal data on behalf of the controller." Both processors and controllers are obliged to comply with GDPR.

So far, the largest GDPR fine issued is €50 million to Google that the company was not following GDPR (Palmer, 2019). British Airways parent company is faced with paying $230 million due to last year's data breach (Wall & Olson, 2019). Nevertheless, the fine has not been finalized.

2.6.4 Health Insurance Portability and Accountability Act Privacy Rule

The HIPAA enforce security provisions and data privacy to protect patients' medical records. HIPAA Privacy Rule is established to protect patients' "individually identifiable health information." The requirement for notifying individuals of a data breach of their information is active since 2009 with the Breach Notification Rule. Examples of individually identifiable health information (OCR, 2003):

- Past, present, future physical or mental health or condition of individuals
- Provision of healthcare to the individual
- Payment information of individuals

The HIPAA Privacy Rule has two primary purposes:

- Defining limitations on the allowable uses and disclosures of protected health information, instructing when, with whom, and under what conditions, health information could be shared.
- Giving patients access to their health data on demand.

The goal of the HIPAA Security Rule is to provide electronic health data that is appropriately secured, and accessibility to electronic health data is controlled. The final goal is the auditable trail of personal health information activity is maintained. HIPAA violations can

cause severe costs for a healthcare organization. In addition to the investigation, remediation, and notification costs, the Office of Civil Rights may also fine the organization.

2.7 Limitations

The study of the monetary impact of a data breach is very recent; as a result, there is a certain lack of methodology and data to develop further models to predict the monetary impact. The limitations are listed below:

- Lack of available datasets that categorizes the information as PII and SPII
- Lack of available data to estimate the total cost of a data breach
- The non-random structure of missing data prohibits the development of an unbiased set of regression models either from using only full data cases or estimating missing data (Little & Rubin, 2002). The dataset is developed based on the availability of information about the variables. Therefore, random sampling is not applied. The details of the data collection will be explained in the next chapter.
- The risk preference, i.e., vNM utility theory of risk, is not considered in this study
- This study will only focus on massive data breaches in which the amount of stolen data is in millions of affected people
- Majority of the companies in the dataset are public companies due to the data availability

2.8 Knowledge Gap

Cyber-risk, cybersecurity investment, or impact of cyber-attacks on firms have been well studied. There have been plenty of studies that propose qualitative or quantitative cyber risk

assessment methods or cybersecurity investment models. However, in recent years, state and federal governments have developed rules, regulations, and laws to strictly enforce organizations to take measurements against PII data breaches to protect the privacy rights of citizens. These government agencies have been issuing fees to any company that fails to protect PII by imposing high penalties. In addition to that costs, there are also indirect costs, such as customer notifications, credit monitoring, and class action lawsuit. As a result, a massive data breach that includes PII or SPII may yield financial consequences that can jeopardize the profit of the companies. However, there is very little empirical research that focuses on the monetary impact of PII or SPII data breaches.

A few studies consider the monetary impact of data breaches, but the criticality of the information or classification of the information has not been studied. Also, current datasets regard the incidents in terms of affected people, not the number of stolen data or the sensitivity of the data. SEC (2018) compels public companies to disclose their loss due to cyber-attack and adequately report the cybersecurity incidents and material cyber risk in their reports. However, only the data breaches that occurred after SEC's update in cybersecurity reporting have more detail in about the impact of data breaches.

As mentioned by many scholars that data breach incidents are not normally distributed (Eling & Loperfido, 2017; Wheatley, Hofmann, & Sornette, 2019). The distribution is heavy-tailed. The difference between the median and mean of data breach cost is wide (Ponemon, 2019). Also, current models consider the only number of affected people or the number of records disregarding the type of information. As a result, those models have very low accuracy in guessing the impact of a new data breach. Therefore, data breach cost models need segmentation to estimate the cost, such as cases where the number of affected people is between:

- 1-100,000
- 100,000 – 1,000,000
- 1,000,000 and more

A study is required to identify the relevant independent variables and develop a new data breach dataset that will help develop predictive models using different algorithms.

The outcome of the study will contribute to the impact of data breach risk assessment. This study will introduce new variables that calculate the monetary cost of a massive data breach. Statistical tests will be employed, which will provide the validity and generalizability of the model.

CHAPTER 3

METHODOLOGY

3.1 Introduction

The nature of a research problem defines the methodology to be used. The unique methodology for this research adopts an empirical approach to statistically test the existence of a relationship of a variable (cost of PII data breach) among other variables. The main focus of this research is to develop a regression model that predicts the cost of PII data breach and how the independent variables are related to the dependent variable. The problem requires the availability of the data for the type of data compromised, total cost due to a data breach, lawsuits, and revenue. Although there are datasets that give an idea about the size of the breach such as Advisen, Privacy Right Clearinghouse, Ponemon reports, or Identity Theft Resource Center, they do not categorize the stolen information per type. Also, only Advisen among those has information about the victim company, such as revenue, cost of the breach, or lawsuits.

Moreover, datasets regarding cybersecurity incidents, either financial or technical datasets, have not reached a consensus on what data needs to be collected, how to categorize it, and to what extent it should be available to the public. Therefore, there is not a complete dataset to observe the total impact of a cybersecurity incident, either financial or technical, yet. Another problem with the datasets is the subjectivity of the data collected. The datasets have been formed based on the interviews, surveys, or yearly reports of organizations. Therefore, it is not entirely objective or correct. However, they are good enough to shed light on the consequences of cybersecurity incidents.

This chapter covers the overall research methodology, research questions, methods, datasets, and validity of the research. The rationale of the research methodology is also discussed in this chapter.

3.2 Type of Reasoning in Research

In general, researchers adopt a research methodology to develop an argument. The research methodology depends on the nature of the problem. In this section, the definition of inductive, deductive, and abductive reasoning will be provided.

3.2.1 Inductive Reasoning

Kerlinger (1986) defined theory as "A set of interrelated constructs, definitions, and propositions that presents a systematic view of phenomena by specifying relations between variables, to explain natural phenomena."

Researchers widely employ inductive research because it provides to identify a new theory, a broad explanation for behaviors, patterns, or attitudes. Inductive reasoning enables developing from the data to broad perspectives to a generalized model. The approach is applied where the hypotheses do not help to develop a generalized model or a theory.

Inductive reasoning brings out theories towards the end of the research process after observations are carried out (Goddard & Melville, 2004). Inductive approach "involves the search for a pattern from observation and the development of explanations – theories – for those patterns through a series of hypotheses" (Bernard, 2006). A researcher may have no theory at the beginning of the study, yet, theories may evolve as a result of the research.

Inductive reasoning is also referred to as a bottom-up approach where a researcher uses observations to develop an abstraction or to define an image of the phenomenon which is already

studied (Lodico, Spaulding, & Voegtle, 2010). "Inductive approach aims to generate meanings from the data set collected to identify patterns and relationships to build a theory; however, the inductive approach does not prevent the researcher from using existing theory to formulate the research question to be explored." (Saunders, Lewis, & Thornhill, 2012).

The typical steps of the inductive approach are explained below (Creswell, 2009):

1. The researcher gathers information (interviews, observations)
2. The researcher asks open-ended questions of participants or record field notes
3. The researcher analyzes data to form themes or categories
4. The researcher looks for broad patterns, generalizations, or theories from themes or categories
5. The Researcher pose generalizations or theories from past experiences and literature

3.2.2 Deductive Reasoning

Another approach that is widely used by scholars is the deductive approach or deductive reasoning. A deductive argument runs from a general statement to conclusions about the specifics. A generalized model or a theory is generated through; stipulation of a theory or hypotheses, justification of that theory, or hypotheses on specific observations. Unlike inductive reasoning, the theory is placed toward the beginning of the proposal for a study. The objective here is to test or verify the theory. A researcher develops a theory, gathers data to test it, and contemplate on its confirmation or disconfirmation by the results (Creswell, 2009). The researcher tests or verifies a theory by examining hypotheses or questions generated from it. Typical steps of deductive reasoning are given below (Creswell, 2009):

The researcher:

1. tests or verifies a theory

2. tests hypotheses or research questions from the theory
3. defines and operationalizes variables derived from the theory
4. measures or observes variables using an instrument to obtain scores

3.2.3 Abductive Reasoning

Abductive reasoning is a type of logical inference form termed by Charles Sanders Peirce in the 19th century. Abductive reasoning differs both inductive and deductive reasoning. Deductive reasoning starts with a rule, proceeds from there to a specific solution that either shows the acceptability of the assertion or falsifies it (Tavory & Timmermans, 2014). Example for deduction:

- All X is Z
- Y is X
- Thus, Y is Z

On the other hand, induction starts with observations that are limited and specific in scope and moves to a generalized conclusion that is not certain (Butte.edu, 2013). Example of induction (Tavory & Timmermans, 2014):

- All observed swans are white
- Thus, all swans are white

On the other hand, abductive reasoning begins with an observation or set of observations and then proceeds to generate the most straightforward and most likely conclusion from the experiences. Abduction is an inference to the best explanation (Douven, 2017). Example for abduction (Tavory & Timmermans, 2014):

- The surprising fact C is observed

- But if A were true, C would be a matter
- Hence, there is a reason to suspect that A is true

3.3 Research Design and Methodology

Based on the characteristics of the research problem, the choice of statistical tests as a primary research method can vary. Since the proposed model will be a follow-up study of Romanosky (2016), and due to the small sample size, multiple regression methods will be proper to apply.

The details of the methodology of this research are given in Figure 2.

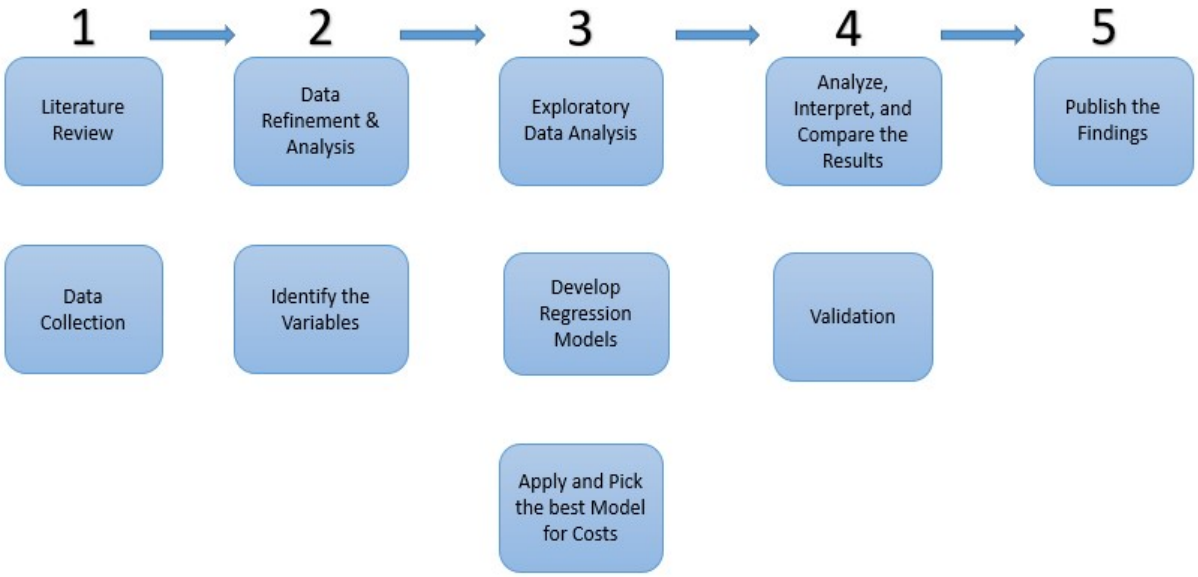

Figure 2. Research Methodology

The methodology of the proposed research is as follows:

Phase I

- Literature Review: This sub-phase includes a literature review to show existing studies and methods which are pertinent to the research problem of the research proposes.
- Data Collection: This sub-phase provides a review of case studies PII data breaches, comparison of datasets such as Advisen, PRC, Ponemon, news, media outlets, or yearly reports of victim organizations

Phase II

- Data refinement and analysis: This sub-phase organizes and refines the data collected from the resources.
- Determine the independent variables: This sub-phase includes the categorization of the type of data stolen according to criticality and determine independent variables.

Phase III

- Perform exploratory data analysis: analysis of the dataset, visualization of the variables.
- Develop regression models: During this sub-phase, regression analysis will be completed for each independent variable. After ensuring multiple regression analysis assumptions, hypotheses for multiple regression analysis will be developed.
- Apply and pick the best model: This sub-phase includes data analysis by using the statistical report outputs such as R^2, adjusted R^2, predicted R^2, and T-statistic, p-value, Mallow Cp, and F-test scores. Thus, it will help determine the proposed predictors explain the response variable.

Phase IV

- Analyze, interpret, and compare the result: In this sub-phase, a detailed analysis of the regression analysis will be executed. The outputs that will be interested are R^2, predicted

R^2, adjusted R^2, Mallow Cp, T-statistic, p-value, and F-test score; in this sub-phase previous models and proposed model will be compared.

- Validation: This sub-phase includes the validation of the model by capitalizing on the statistical significance and generalizability of the model.

Phase V

- Publish the findings: This phase will be close out of the research, and conclusions of the research will be reported with the various deliverables.

3.4 Model Development

Measuring impacts of cybersecurity incidents has been a significant challenge. The proposed research will categorize the data according to criticality, which is elucidated by the Department of Homeland Security (DHS, 2017).

Companies tend to under-report cybersecurity events to evade fees or loss of reputation (Ferran, 2016; IT-Online, 2016). As a result, organizations keep records of their cyber incident data in-house. However, this may not be the situation all the time. For example, depending on the scope of incidents, some of these incidents may have to become public.

There are not many sources that collect and organize data breach incidents through surveying companies. The available datasets will be explained in the following section.

3.4.1 Data Collection

There are cybersecurity incidents datasets as Advisen, PRC, and ITRC. Each of these dataset sources will be briefly described. Furthermore, this study will elaborate on why the datasets stand-alone is not adequate to use the data in the proposed model. These datasets are

essential for the proposed model because each dataset will contribute to the development of the new dataset, which will be used to determine the variables of the proposed model. Besides, case studies, news websites, quarterly, or yearly reports of organizations will be used to develop the new dataset. Typically, the categorization of the data is based on the number of affected people and the type of stolen information.

The goal of the proposed research is to introduce a categorization that is based on the criticality of the stolen data.

Advisen

This database is developed and sold by Advisen Ltd, which is a leading data provider for the commercial property and casualty insurance market. The dataset is more comprehensive than the other available data breach datasets. It includes more than 40,000 cyber incidents. It includes (Advisen, 2019):

- Case information (type, legal status, accident date)
- Number of affected people
- Type and amount of monetary loss
- Victim company (revenue, number of employees, industry code, geography)
- Actor

Although Advisen cyber loss dataset is one of the most comprehensive datasets, it does not regard the type of records. This means that the Advisen dataset only considers the number of affected people. Therefore, the dataset alone is not suitable to use in the proposed model.

Privacy Rights Clearinghouse (PRC)

PRC data breach dataset is a publicly free dataset that includes the data breaches from 2005 to 2019 that consists of 9,000 incidents. It provides the following information:

- Date
- Company (name, location, type of organization)
- Number of affected people

Unlike Advisen, the PRC dataset does not mention about the total cost of the data breach or revenue of the company or legal status of the case. Besides, it does not categorize the criticality of information and only considers the number of people (PRC, 2019).

Identity Theft Resource Center

ITRC has been recording publicly available disclosures of data breaches since 2005. The types of data they track are; social security number, credit/debit number, email-password-user name, protected health information, driver's license numbers, and financial accounts. ITRC categorizes the data as sensitive and non-sensitive. The ITRC dataset includes the information of victim company, date of the breach, records. However, it does not mention about the cost the data breach caused but only calculates the total breached records (ITRC, 2019).

3.4.2 Multiple Regression Definition

Correlations are complex computations that measure the degree of association between two or more variables, using exact scores instead of rough categories. The calculation produces a single number called a correlation coefficient, which summarizes the relationship.

Linear regression is a statistical procedure used to estimate the amount of change in a dependent variable that can be expected for a given change in an independent variable. Simple regression involves one dependent variable and one independent variable. A linear equation of a regression model can be shown as follows:

$$Y = \alpha + \beta X$$

The α is the value of Y when X equals zero. It is also called Y-intercept.

The β is referred to as the regression coefficient. It is also known as the slope of the regression line. The β gives the number of units change in Y that can be expected for a one-unit change in X.

It is essential to distinguish between the actual Y values that do not fall on the regression line and the corresponding estimated E(Y) values that we would estimate based on a given respondent's X value. The discrepancy between the actual Y value and E(Y) value represents the prediction error. When the Y values tend to cluster very close to the regression line, E(Y) and Y values will be very similar, and the error in prediction will be small. However, when the Y values tend to deviate markedly from the regression line, the Y and E(Y) values will be quite different, the error in prediction will be high.

Multiple regression

Multiple regression is a frequently used statistical method for analyzing data when there are more than one independent variable and one dependent variable. Independent variables are also called predictors, and the dependent variable can be called criterion, outcome, or response variable.

Multiple regression is an extension of simple regression. A general multiple regression formula can be defined with the following linear equation:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

The coefficients β_n values in multiple regression equations are referred to as partial-regression coefficients. These coefficients represent the degree of change in the dependent variable that we would estimate for a one-unit change in the specified predictor.

The multiple correlation coefficient (R) is used to summarize the accuracy of our prediction equation. The difference between Y and E(Y) stands for the error in the prediction. If we have selected a set of predictors that yield accurate estimates of Y, then the difference between Y and E(Y) values will be small, and multiple correlations will be high. If, however, we have selected a set of predictors that yields poor estimates of Y, then the difference between E(Y) and Y values would be larger, and the multiple correlations would be small.

R is the correlation of the combination of the independent variables with the dependent variable. The R tells the strength of the correlation exists between the predictor variables and the criterion variable. The goal is to find a linear combination of independent variables that explains the most variance in the dependent variable. Multiple regression is used to predict or explain the relationship between the linear combination of the independent variables and the dependent variables. As with correlation, even a high R does not mean that the independent variables caused the change in the dependent variable.

The R^2 gives the proportion of the variance in the response variable, which is accounted for by the predictors in the regression. If R = 0.9 then $R^2 = 0.81$; it determines that the independent variables are considered account 81% of the variance in the response variable.

Conditions and assumptions of multiple regression

The assumptions to employ multiple regression include the follow (Gliner, Morgan, & Leech, 2017; Statistics Solutions, n.d.):

- The relationship between each of the predictor variables and the dependent variable is linear
- The error or residual is normally distributed and uncorrelated with the predictors
- There should be no high correlations between independent variables. In the case of multicollinearity, two or more predictors are highly correlated. Variance Inflation Factor or a Correlation matrix can be applied to see if there is multicollinearity between independent variables
- The dependent variable should be approximately normally distributed.

The validity of the model

The value of R^2 means how well the prediction fits the data. R^2 can be used in F statistics to determine the data provide sufficient evidence about how the overall model contributes information to predict the response variable. The value of R^2 will increase as more variables are included in the model. R^2 can be forced to approach to one '1' though the model provides no information for the prediction of Y.

Unlike R^2, adjusted R^2 considers both the sample size and the number of β parameters in the model. Adjusted R^2 will always be smaller than $R^{2,}$ and, cannot be one by adding more independent variables to the model. Some researchers prefer adjusted R^2 while choosing a measure of model adequacy.

However, what both R^2 and adjusted R^2 provide are useful, considering only their result is not enough to claim that the model helps predict. Predicted R^2 is calculated to define how well a regression model makes a prediction. It is helpful to determine whether there is an overfitting in the model. If there is a substantial difference between adjusted and predicted R^2, that means the model has overfitting. Predicted R^2 is calculated by (Minitab, 2013):

- excluding a data point from the dataset
- computing the regression equation
- calculating how well the model estimates the removed observation
- repeating this for each data points
- and calculating aggregated R^2

Applying T-tests on each β parameter can provide a better idea about the adequacy of the model. Also, the F-test can be used to make inferences about the overall adequacy of the multiple regression model. Another method can be Mallows' C_p, Akaike's Information Criterion (AIC), and Bayesian Information Criterion (BIC) test to find the best model among subset models.

Checking the utility of a multiple regression model

1. Conduct a test of overall model adequacy using the F test that is

   H_0: $\beta_1=\beta_2=\beta_3=\beta_4= 0$

   H_A: At least one $\beta_i <> 0$

   If the model seems adequate to go to step 2
2. Conduct T-tests on those β parameters that are of interest (most important β). Number of β should be limited to avoid type I error
3. Examine the values of adjusted R^2 and predicted R^2
4. Examine the p-values of the variables
5. Compare each model's:
   a. the C_p values
   b. Akaike's Information Criterion

c. Bayesian Information Criterion

d. PRESS (Prediction error sum of squares)

3.4.3 Model Development

In this dissertation, an incident-driven model is used to assess data breaches to predict monetary impact. Current datasets supply the data breach in terms of the number of records stolen, which only focuses on the number of people affected. Due to the lack of available datasets, we will employ the data collected from yearly reports of organizations, case studies, and other datasets like PRC, Ponemon, websites, and news. This study will only focus on the impact leg of risk assessment.

The dataset will include the variables as follows:

- The total cost of the data breach
- Revenue of the organization
- Total number of data stolen in PII category
- Total number of data stolen in SPII category
- Class-action lawsuits

Once the data are obtained, they need to be analyzed in a way that enables the calculation of the monetary impact of the loss of records. So far, there are only two statistical models that predict the monetary impact of cybersecurity incidents, which were covered above (Jacobs, 2014; Romanosky, 2016).

The right model for this study will be multiple regression analysis to predict the PII data breach cost because of:

- the continuity of earlier studies (Jacobs, 2014; Romanosky, 2016)

- small sample size
- measuring the association between dependent and independent variables

Since multiple regression estimates the correlation between dependent and independent variables, we cannot infer that there is a cause-effect relation. The variables are shown in the table below.

Table 8. Description of the Variables

Variable	Type	Denotation	Definition
Total Cost	Dependent variable	Y	The estimated cost of a data breach
Revenue	Independent variable	X_1	Yearly revenue of the company at $
High critical data	Independent variable	X_2	Total number of SPII (e.g., SSN, credit card numbers) in numbers
Low critical data	Independent variable	X_3	Total number of PII (e.g., name, address) in numbers
Class-action lawsuit	Independent variable	X_4	Binary variable (1 or 0; 1 means if there is a lawsuit)

Hypothesis

- The independent variable X_1 relates to the dependent variable (Total cost)
- The independent variable X_2 relates to the dependent variable
- The independent variable X_3 relate to the dependent variable
- The independent variable X_4 relate to the dependent variable

After identifying the dependent and independent variables, Multiple Regression assumptions will be checked, which are:

1. The linear relationship between the dependent and independent variables: to check this condition, scatterplots will be used, or simple regression will be done for each independent variable.
2. Residuals are normally distributed: the errors between observed and estimated values are normally distributed. It can be checked with a histogram.
3. No multicollinearity: no high correlation among independent variables. This can be checked via a correlation matrix or the Variation Inflation Factor (VIF). Low VIF values are desirable. If there is multicollinearity in the data, one of the independent variables can be taken out of the equation.
4. Homoscedasticity: one way to check homoscedasticity is by creating a scatterplot of residuals versus predicted values. We look for that there is no clear pattern in the distribution to satisfy homoscedasticity.

   $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

   E(Y): Expected total cost of PII data breach in $

   β_n: partial correlation coefficient per each independent variable

   β_0: constant where X equals to zero

   X_1: Revenue, in $

   X_2: High critical data were stolen: Combined data of PII (e.g., SSN and name) in number

   X_3: Low critical data were stolen: PII (e.g., name, address) in numbers

   X_4: Binary variable (1 or 0; 1 means if there is a lawsuit)

Backward Elimination

Initially, the backward elimination method will be applied to find the best model that explains the total cost of a data breach. The backward stepwise selection offers an efficient way to identify the statistically non-significant variables. The first model includes all independent variables. In each run, a statistically non-significant independent variable is removed. The backward selection method requires that the number of sample size is larger than the number of independent variables.

The output will give the results of R^2, predicted R^2, adjusted R^2, p-value, PRESS statistic, T-tests score, F-test, AIC, and BIC. Predicted R^2, and adjusted R^2 alone is not enough to claim the overall adequacy of the predicting model. Therefore, we need to look at Cook's D, AIC, BIC, C_p.

Statistical Software

The data is deployed in Python 3 environment. Python is a programming language that allows researchers to work more quickly and more effectively. It provides an environment for web development, data science, or scripting. It has a very rich library for data science such as Pandas, NumPy, StatsModels, and Scikit-learn. A multiple regression analysis will be held using data science libraries in the Python environment. Since Python libraries do not have predicted R^2 and PRESS statistic calculations, they will be calculated in Minitab.

T-test:

A T-test is an inferential statistic utilized to find if there is a significant difference between the means of two groups. It is used to test hypotheses that allow testing of an assumption applicable to a population.

A T-test looks at the t-statistic, t-distribution, and the degrees of freedom to find the probability of difference between two sets of data. Our α value will be 0.05. Then, we will define the t-critical value from looking at the T-table. According to our T-statistic, we will interpret that our results are statistically significant, and there is a pattern. Our null hypothesis will be that there is no correlation between the independent and dependent variables. An alternative explanation will be that there is a linear relationship between independent and response variables.

H_0: $\beta = 0$

H_A: $\beta > 0$ or $\beta < 0$

Figure 3 illustrates an example of two-tailed t-distribution. Our expected t-statistic should be in the blue area, which states that there is a linear correlation between response and predictor variables. If our t-statistic occurs in the white area, then we must accept the null hypothesis, which states that there is no correlation between the independent and dependent variables.

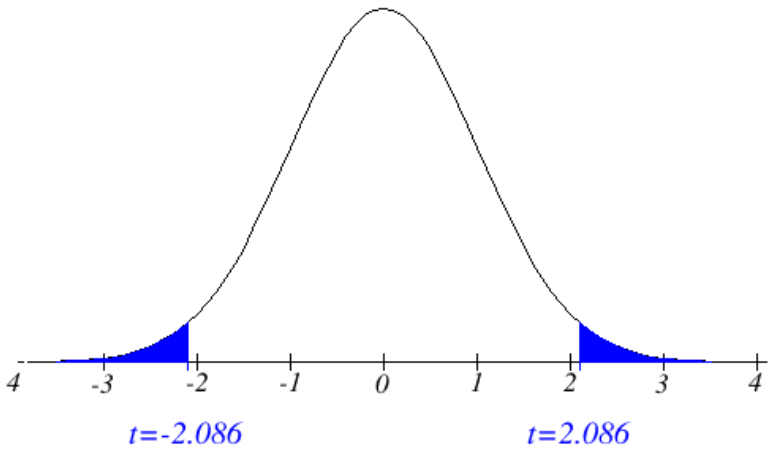

Figure 3. Example of Two-tailed t- distribution

F statistic: is the ratio of among estimate of variance and the within the estimate of variance. The output will give us F-statistic. We will determine α value as 0.05. Therefore, we will expect our F-the statistic will be higher than F-critical value depending on the sample size. If so, then our hypothesis will be correct, which is there is a correlation between independent and dependent variables, and the results are statistically significant.

Mallows' C_p Criterion

Mallows' C_p is developed by Colin Mallows to evaluate the fit of a regression model, which is estimated using ordinary least squares. It compares the predictive ability of subset models to that of the full model. Mallows' C_p-statistics assess the size of the bias that is introduced into the predicted responses by having an underspecified model (N/A, 2018). A small value of C_p where C_p is near p tells that the model is more precise.

$C_p = RSS_p/s^2 – (n-2p)$

RSS= residual sum of squares

p= number of parameters including β_0 (intercept)

s^2= residual mean square from the largest equation

Akaike's Information Criterion (AIC)

AIC compares the quality of a set of regression models to each other. Alone AIC number does not mean anything about the model. It compares each model and assigns them a value. It provides the best model out of a set of models. The best model has the lowest AIC value among the subset models (James, Witten, Hastie, & Tibshirani, 2017). The formula is:

AIC= -2(log-likelihood) +2p

P= number of parameters including β_0 (intercept)

Log-likelihood = measure of model fit. A higher number of log-likelihood means a better fit

If the sample size is small, $n/K < \approx 40$, then use the following AIC formula:

AIC_c = -2(log-likelihood) +2K + (2K(K+1)/(n-K-1))

N= sample size

K= number of model parameters

Log-likelihood = measure of model fit

Bayesian Information Criterion

BIC is developed from a Bayesian perspective. *The BIC tends to take on a small value for a model with a low-test error,* therefore, select the model that has the lowest BIC value like AIC and C_p to determine the best model (James et al., 2017).

BIC= -2*(log-likelihood) + k*log(N)

K= number of parameters

N=sample size

Predicted Residual Error Sum of Squares (PRESS)

PRESS statistics is a cross-validation technique to provide how a model fits the data in regression. It is calculated similarly to Predicted R^2. In this one, we compute the sums of the squares of the prediction residuals for the removed data points. The lowest value of PRESS statistic can be interpreted as the best model that fits the data in the same dataset (Minitab, 2019).

Skewness

Skewness tells about how symmetrical the residual distribution is. The skewness of a normal distribution is equal to zero. Negative values show that the data is skewed left, and positive skew values indicate that the data is skewed right. If the data is left-skewed, the data is concentrated on the left side of the distribution rather than in the middle. If the data is right-

skewed, vice versa. As a rule of thumb, if the skewness is between -0.5 and 0.5, the distribution of the data is reasonably symmetrical (McNeese, 2016).

Kurtosis

Kurtosis does not tell about the shape of the peak; instead, it is the interpretation of tail extremity (Westfall, 2014). It is "a measure of the combined weight of the tails relative to the rest of the distribution." (Wheeler, 2011). Therefore, kurtosis tells about the tails of the residual distribution. If (NIST/SEMATECH, 2012):

- kurtosis > 3, heavy-tailed distribution
- kurtosis <3, light-tailed distribution

Omnibus and Prob(Omnibus)

Omnibus value tells about the skewness and kurtosis of the residuals. The ideal value should be close to zero to indicate the normality of the residuals. The Prob(Omnibus) suggests that the probability of the residuals is normally distributed. In an ideal case, it is expected to be close to one (McCarty, 2018).

Durbin-Watson

The Durbin-Watson value says the measure of autocorrelation. The Durbin-Watson test ranges from zero to four. If the value is (Kenton, 2019):

- 2, no autocorrelation
- Between 0 and 2, positive autocorrelation
- Between 2 and 4, negative autocorrelation

The Durbin-Watson test also tells about the homoscedasticity; the value is expected to be between one and two (McCarty, 2018).

Jarque-Bera/ Prob(JB)

Jarque-Bera test is a measure of the normality of the residuals. It tests both skewness and kurtosis. It follows a similar pattern with Omnibus values (McCarty, 2018).

3.5 Generalizability of the Research

The outputs of research are as useful as used by others. The generalizability of research defines the effectiveness and usefulness of the research. According to Polit & Beck (2010) generalizability is a study of reasoning which deduces broad inferences from specific observations.

Generalizability is a fundamental element of research and convinces the usefulness of research (Lee & Baskerville, 2003). Ensuring the generalizability is a way to broaden the applicability of the findings of a research.

A statistical generalization is a common form of generalization methods. The sample to be used should satisfactorily represent the population to have successful generalizability. Random sampling can increase the chance of the sample representing the population.

Gliner et al. (2017) define research validity as the merit of the whole study, includes measurement reliability and statistics, internal validity, overall measurement validity of the constructs, and external validity.

They describe the research validity as below:

1. Measurement reliability and statistics
    a. Test-retest reliability
    b. Parallel forms reliability
    c. Internal consistency reliability
    d. Interrater reliability
2. Internal Validity

    a. Equivalence of groups on participant characteristics
    b. Control of extraneous experience or environmental variables
3. Measurement validity and generalizability of the constructs
    a. Face validity
    b. Content validity
    c. Criterion-related validity
    d. Construct validity
4. External validity
    a. Population validity
    b. Ecological validity

Reliability discusses if score to items on an instrument is consistent, stable over time, and there is a consistency.

Internal validity can be defined as "the approximate validity which we can infer that a relationship is causal" (Campbell & Cook, 1979). Internal validity is shaped by the strength or soundness of the design and influences if a researcher can deduce that the predictor variable caused change on the response variable.

Measurement Validity and generalizability of the constructs

Validity is an establishment of evidence for the use of a specific instrument in a particular setting. An instrument may have high reliability, yet, it may not be valid. The instrument should measure what it is supposed to measure. The authors cover four different types of evidence for validity.

1. Face validity: an instrument has face validity if the content seems to be suitable for the instrument. Face validity does not define the content.

2. Content validity: it denotes the actual content of the instrument. The content creates the instrument that is representative of the concept that one is trying to measure. The first step of establishing content validity is a definition of the concept that the researcher is trying to measure. A second step to content validity is a literature search to find out how this concept is embodied in the literature.
3. Criterion-related validity: It is the validation of the instrument against the external criterion. This validation method contains establishing a correlation coefficient between the instrument and the external criterion. There are two types of criterion validity:
    a. Predictive evidence: examines the relationship between the response and predictor variables to predict future performance.
    b. Concurrent evidence: examines the relationship between variables
4. Construct validity: construct validation is a process where researchers carry out studies to demonstrate that the instrument is measuring a construct.

External validity

A research study should have a high rate of external validity, or the researcher should at least be cautious about generalizing the findings to other measures, populations, and settings. External validity seeks the question of generalizability: "To what populations, settings, treatment variables, and measurement variables can this effect be generalized." (D. T. Campbell & Stanley, 1967).

The representativeness of the sample determines the external validity. However, the sampling design or the type of sampling does not directly affect the internal validity of a study. External population validity is affected by sampling design; however, internal validity depends on how subjects get into groups.

Population external validity should be based on rating on the following criteria

1. representativeness of accessible population vis-à-vis theoretical population
2. adequacy of sampling method from the available population
3. sufficiency of the response or return rate

Ecological external validity should be based on:

1. the naturalness of setting or conditions
2. adequacy of rapport with testers or observers
3. the naturalness of procedures or tasks
4. appropriateness of timing and length of treatment
5. the extent to which results are restricted to a specific time in history

This research aims to develop an approach to predict the cost of PII data breach by employing multiple regression. First, the scarcity of available data is an issue for this study due to the lack of formality in data breach reporting. The nature of the study requires criterion-related validity to have valid research. To achieve generalizability, we will select all available data of massive data breach and check statistical tests as adjusted and predicted R^2, PRESS, Cook's D, F-test, t-test, Mallows' C_p to make inference about the population.

CHAPTER 4

MODEL DEVELOPMENT

4.1 Introduction

This chapter covers how the dataset is developed, existing datasets are compared, characteristics of the developed dataset, how variables are defined, why multiple regression is used, and the outputs of the developed models.

4.2 Data Collection and Variables

This study only focuses on data breaches, where the number of affected people is more than one million. A multiple regression model will be developed for the data breach impact will be based on the developed dataset. The incidents are acquired from the PRC dataset as of November 4, 2018. The developed dataset does not include government organizations, non-profits, educational institutions, or unknown organizations in the PRC dataset, which were listed as 'EDU,' 'GOV,' 'NGO,' 'UNKN.' This study will include data breach incidents that happened in Finance, Insurance, Retail, Online Retail, Healthcare industries, and other businesses. Those are listed as 'BSF,' 'BSO,' 'BSR,' and 'MED.' Also, this study includes any type of PII disclosure, either intentional or accidental. Then, the data is filtered according to the criteria above from 2005 to November 4, 2018. As a result, 133 distinct data breach incidents occurred between those dates. In addition to that, the AOL data breach happened in 2004 is added into the dataset. As a result, the dataset includes 134 distinct data breach cases in total in which the number of affected people is more than one million. Next, the information sought for each case for the following parameters per incident:

- Cost types and its amount in $

- Type and number of stolen data
    - PII: the sum of number of names, address, phone number, date of birth, and so on
    - Sensitive PII (SPII): the sum of number of social security number, driver's license numbers, passport numbers, and so on
- The legal status of the case: the class-action lawsuit is dismissed or concluded
- Revenue for the year in $ that data breach happened

This study compares the PRC numbers with ITRC and website sources to be sure about the number of affected people. Company quarterly and annual reports, SEC filings, news media, websites, and case studies are reviewed. The number of affected people for 'Neiman Marcus' data breach was reported for more than one million. However, this study finds out that the number of affected people is later reported 370,000. Therefore, the Neiman Marcus data breach case is excluded from the list. Regarding the parameters mentioned above, 31 data breach cases have information about the parameters. Therefore, our data breach sample size is 31.

This study mostly focuses on the impact of the data breach on the U.S. citizens, and the companies are publicly traded in the U.S. stock markets. The reason is that the U.S. has well-defined laws, regulations, and agencies to monitor data breach incidents. However, for the Marriot data breach, this study added the fine of GDPR for the company to the total cost because there has not been data breach settlement or fines issued by U.S. courts or agencies. Moreover, it is not clear how many U.S. citizens are affected by that data breach. Therefore, all number of stolen records are included in the table. Next, this study only considers the number of U.S. and Israel Yahoo account users because of the data breach settlement and fees cover only those two countries. It is not difficult to reach the sought information for public companies. However, there is limited information regarding the data breach for private companies or companies that are not

traded in U.S. stock markets. For example, Uber was not a public company when the breach happened; therefore, we only know the cost of settlement, which is $148 million. Also, for older cases, we have not found the revenue information of the company for the breach year due to bankruptcy or acquisition. Therefore, we have recorded the next available revenue information or parent firm revenue.

4.2.1 Calculation of number of the PII and the SPII

The current data breach cases are reported based on the number of affected people. However, in this model, we will categorize the stolen data into PII, which is low critical data, and SPII, which is highly critical data. Also, in most cases, the victim companies report only the number of affected people and categorically type of stolen information. Therefore, in many cases, we do not have the exact number of stolen data for each PII or SPII type. However, we know the number of affected people and the type of stolen PII and SPII. Therefore, we take each PII or SPII type equal to the number of affected people unless the number of records for each PII or SPII is stated. For example, in Anthem data breach, we know the number of affected people is 78, 800,000. We also know the stolen type of PII and SPII for Anthem:

- Number of records for the name: 78, 800,000
- Number of records for address: 78, 800,000
- Number of records for SSN: 78, 800,000

For the example purpose, the total number of records is calculated below:

- Total number of PII records exposed in Anthem breach: 157,600,000
- Total number of SPII records exposed in Anthem breach: 78,800,000

Table 9. Dependent and Independent Variables

Variable	Type	Denotation	Definition
Total Cost	Dependent variable	Y	The estimated cost of a data breach
Revenue	Independent variable	X_1	Yearly revenue of the company at $
High critical data	Independent variable	X_2	Total number of SPII (e.g., SSN, credit card numbers) in numbers
Low critical data	Independent variable	X_3	Total number of PII (e.g., name, address) in numbers
Class-action lawsuit	Independent variable	X_4	Binary variable (1 or 0; 1 means if there is a lawsuit)

4.3 Multiple Regression Models

This study uses multiple regression model to define the association between dependent and independent variables and also to develop a predictive model to estimate the monetary impact of a PII data breach. Developing a multiple regression model has several benefits (James et al., 2017):

- Interpretability of the model
- A small number of predictor variables
- Small sample size
- If there is a linear relationship between response and predictor variables

Multiple regression modelling provides so many advantages. However, on the other hand, to develop a multiple regression model, the assumptions below must be met (James et al., 2017):

- Less flexibility of the model
- The relation between response and predictor variable must be linear
- Residuals need to be normally distributed

- There should be no multicollinearity among predictor variables
- In the case of heteroscedasticity, the different transformation of the dependent variable is sought.

In this study, the Ordinary Least Square (OLS) is used to develop models. OLS models presume that the analysis is fitting a model of a linear relationship between independent and dependent variables that minimizes the sum of square error (Zdaniuk, 2014).

Backward elimination is that running the model with all independent variables, then, in each run, removing a statistically non-significant variable.

$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

E(Y): Expected total cost of PII data breach in $

β_0: constant where X equals to zero

X_1: Revenue, in $

X_2: High critical data were stolen: Combined data of SPII in number

X_3: Low critical data were stolen: PII in numbers

X_4: Binary variable (1 or 0; 1 means if there is a lawsuit)

4.3.1 Exploratory Data Analysis

Initially, exploratory data analysis is performed to see the linearity and distribution of the variables and characteristics of the dataset.

Summary of the Dataset

There are 31 massive data breach incidents in the table. The dataset covers the incident from 2004 to 2018. The industries included in the dataset are; medical, finance, retail, online

retail, and other businesses. Also, the dataset provides information about the class-action lawsuit status of the companies. Swarm plots below illustrate the data breach cost per industry, class-action lawsuit per industry, and the data breach per industry.

Figure 4 shows that the cost of a data breach per industry. In the medical industry, the Anthem case has the highest cost amount. For BSO- other businesses- is Yahoo, and for the finance industry, Equifax is the outlier.

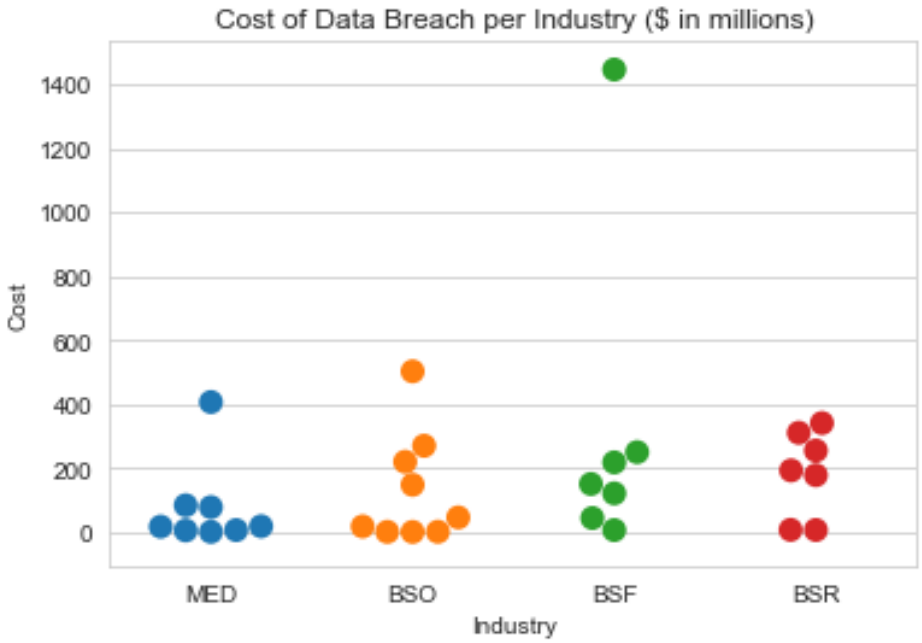

Figure 4. Cost of Data Breach per Industry ($ in millions)

Figure 5 shows that 19 cases have a class action lawsuit concluded. In 12 cases, there currently are not any class-action lawsuit.

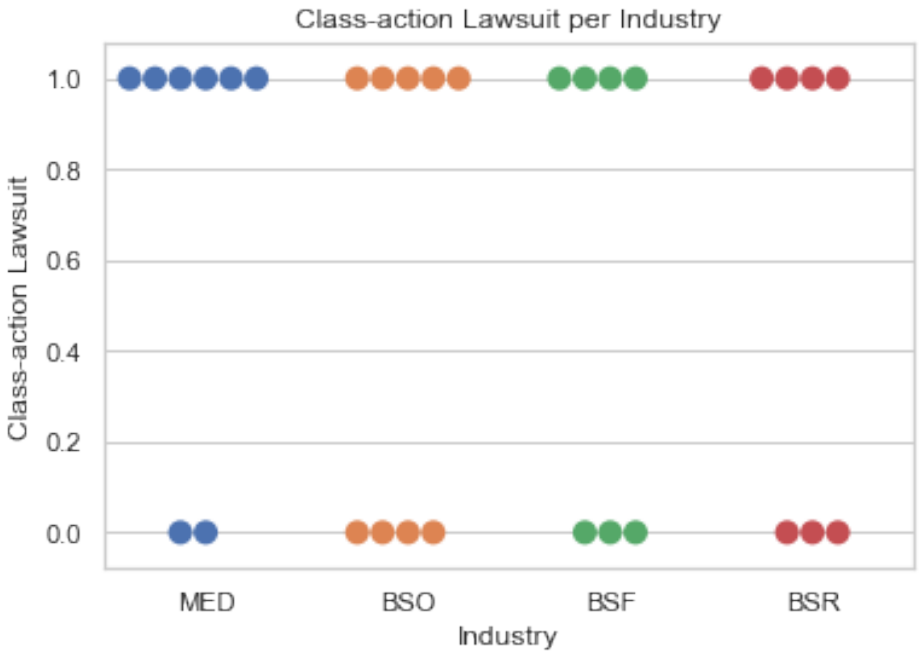

Figure 5. Class-action Lawsuit per Industry

Figure 6 shows the massive data breach per industries over the years. It seems the medical industry was the point of interest of hackers in 2015. However, since 2015, any massive data breach in the medical industry is not observed. The number of incidents per industry is close to each other.

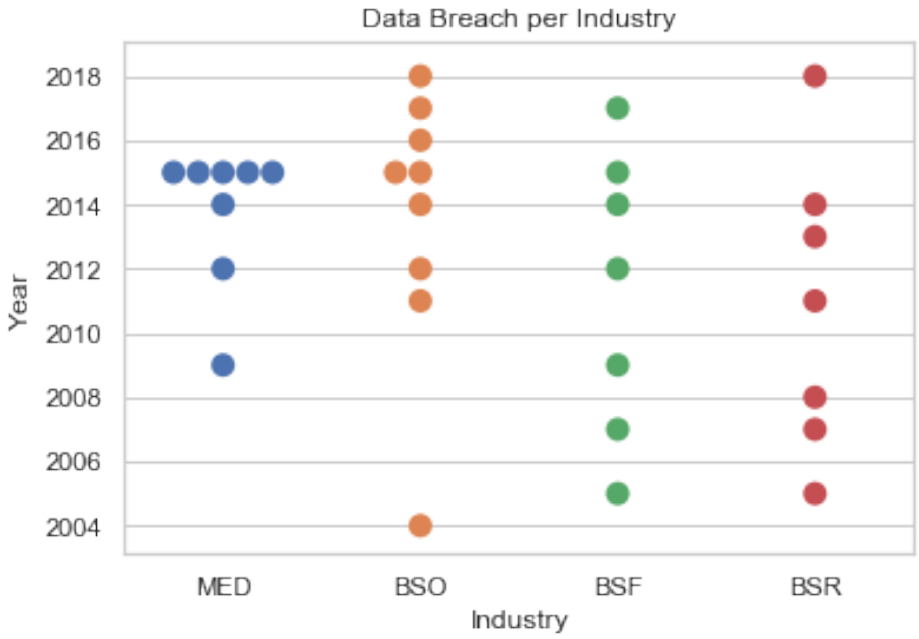

Figure 6. Data Breach per Industry

The discrepancy between data breach cost, the revenue of the companies, and the number of records stolen are very high. The summary of the numeric variables provided in table 10.

Table 10. Summary Statistics

	Revenue ($)	Total number of PII records	Total number of SPII records	Total Cost ($)
Mean	17,832,960,000	165,991,300	22,838,880	172,640,000
Std. deviation	28,440,830,000	267,261,000	39,396,670	273,120,000
Min	9,000,000	0	0	700,000
25%	1,312,000,000	3,050,600	0	7,050,000
50%	5,169,000,000	40,000,000	2,800,000	84,000,000
75%	17,871,500,000	213,800,000	36,425,000	235,000,000
Max	94,205,000,000	970,000,000	164,306,500	1,445,200,000

The pair plot of the variables provided below shows the distribution and the relationship between the variables. The ‘Equifax’ case stands far away from the average.

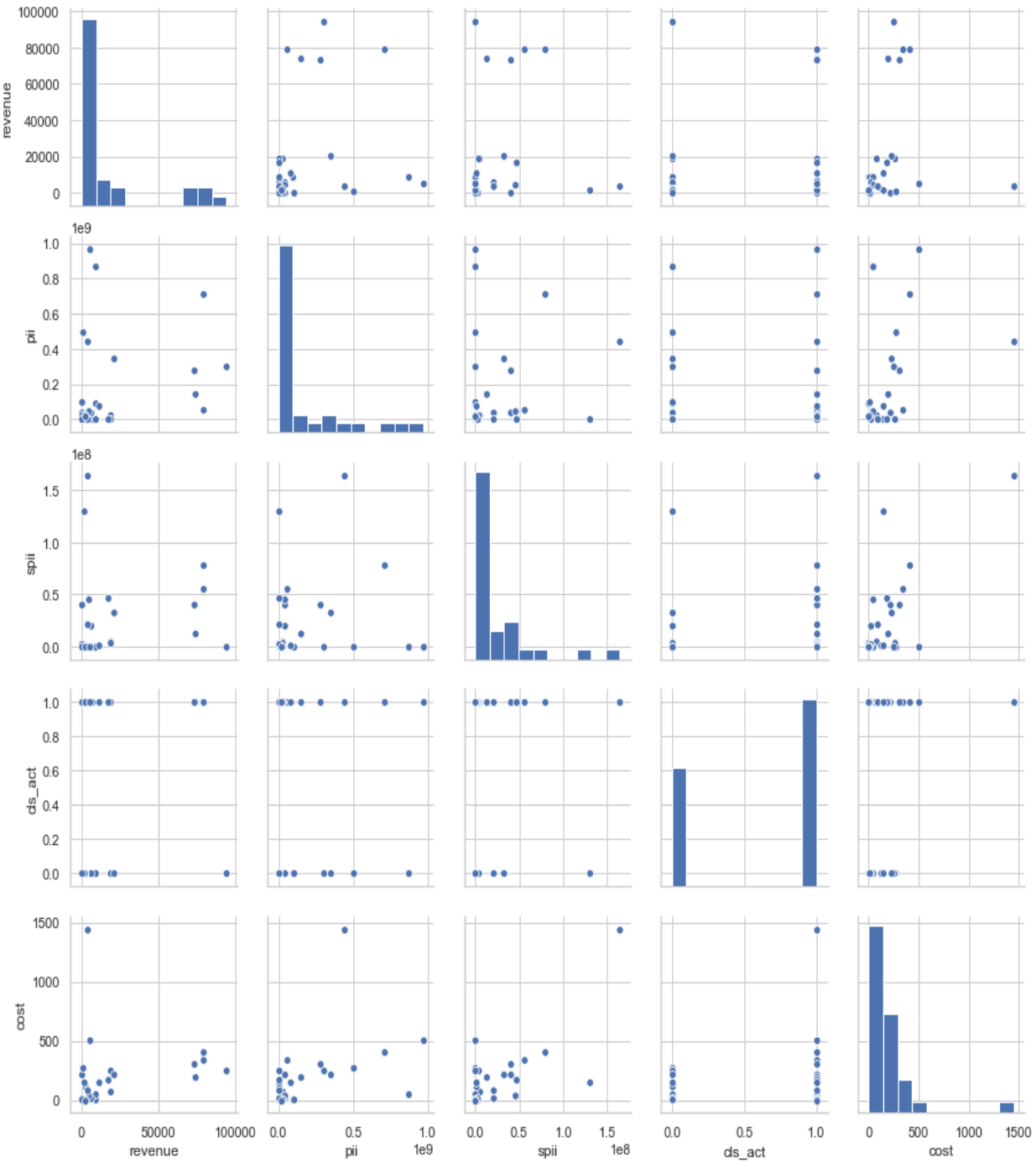

Figure 7. Pair Plot of the Variables

The next figure shows the histogram of the data breach cost. Equifax is a unique case in terms of the total cost. The majority of the data breach cost is concentrated between o and $300 million.

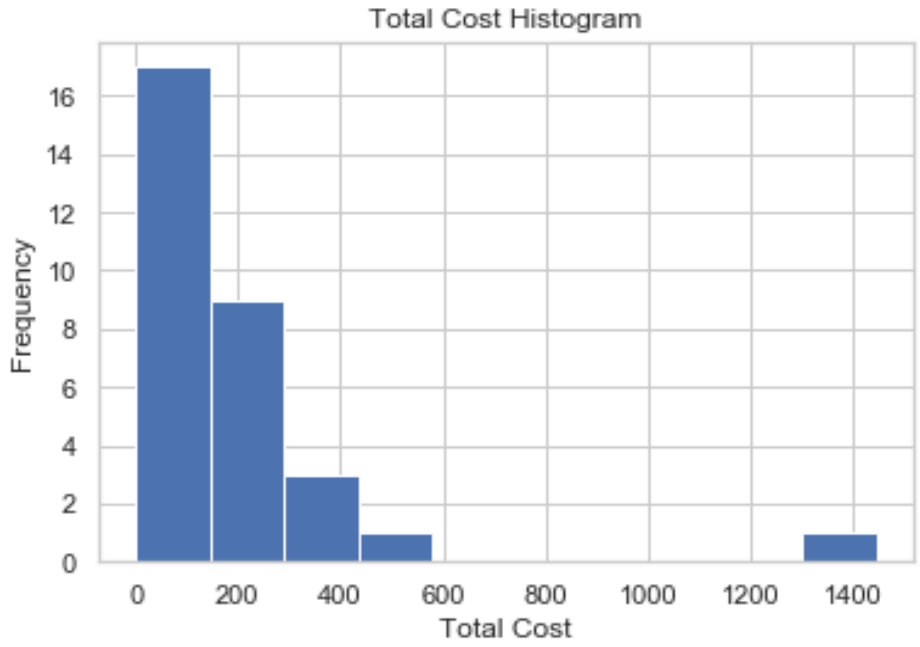

Figure 8. Histogram of Total Cost

Figure 9 shows that most of the companies have revenue from $9 million to 20 billion. Nevertheless, very few data points have revenue between $70 billion to $94 billion.

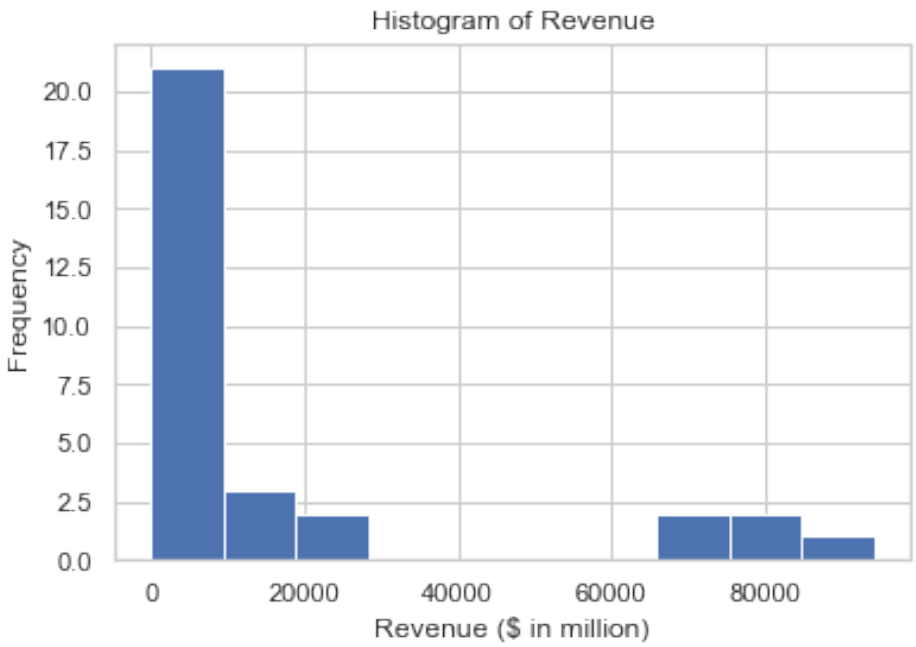

Figure 9. Histogram of Revenue

Figure 10 illustrates the distribution of the number of stolen PII records. The number of stolen PII records is shown in scientific notation. The majority of the data breach is concentrated between 0 and 100 million.

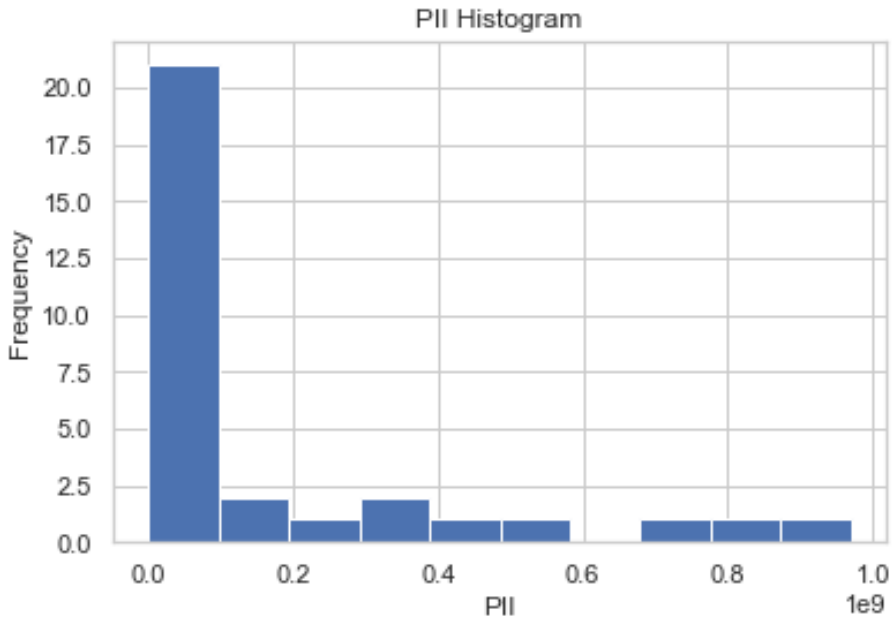

Figure 10. Histogram of PII

Figure 11 shows the distribution of the number of stolen SPII records. The number of stolen SPII records are shown in scientific notation. The majority of the data is concentrated between 0 and 25 million.

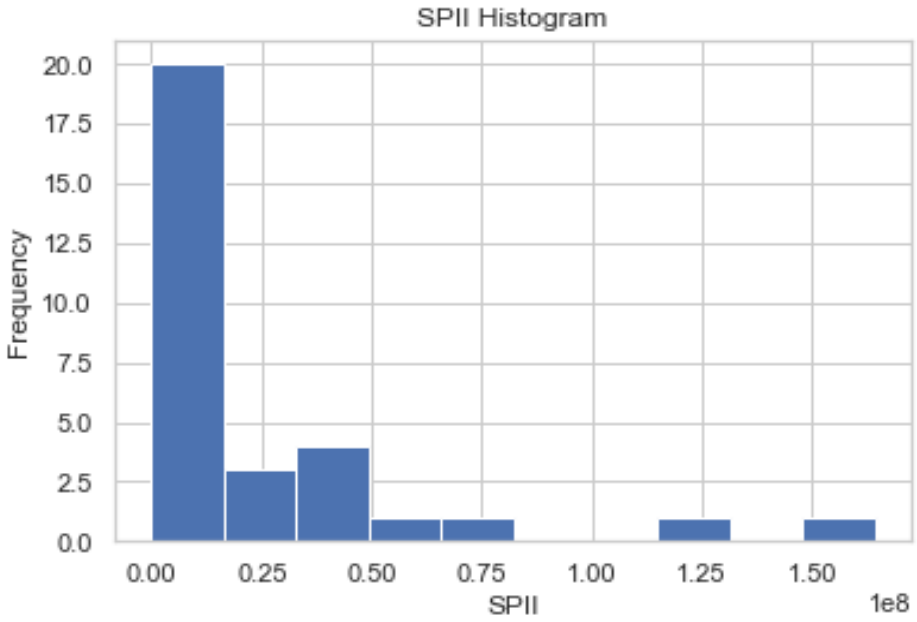

Figure 11. Histogram of SPII

The proposed model has one categorical variable, which is the class-action lawsuit. To include the categorical variable into the model, it is transformed into a binary variable. Concluded cases are coded as 1. On the other hand, dismissed cases are coded as 0. As a result, the class-action lawsuit variable has 19 ones (concluded case) but also 12 zeros (dismissed). Figure 12 shows the distribution of the class-action lawsuits.

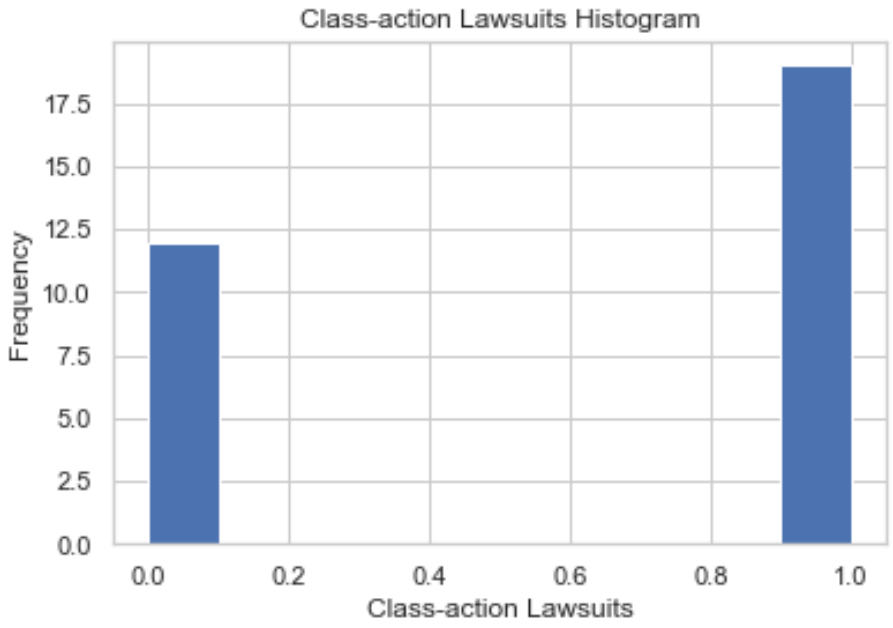

Figure 12. Histogram of Class-Action Lawsuits

The linearity of the relationship

The figures below show a scatter plot of cost vs. revenue, PII, and SPII, respectively. There is an acceptable level of linearity between cost and revenue. Only the Equifax case is far away from the linear line.

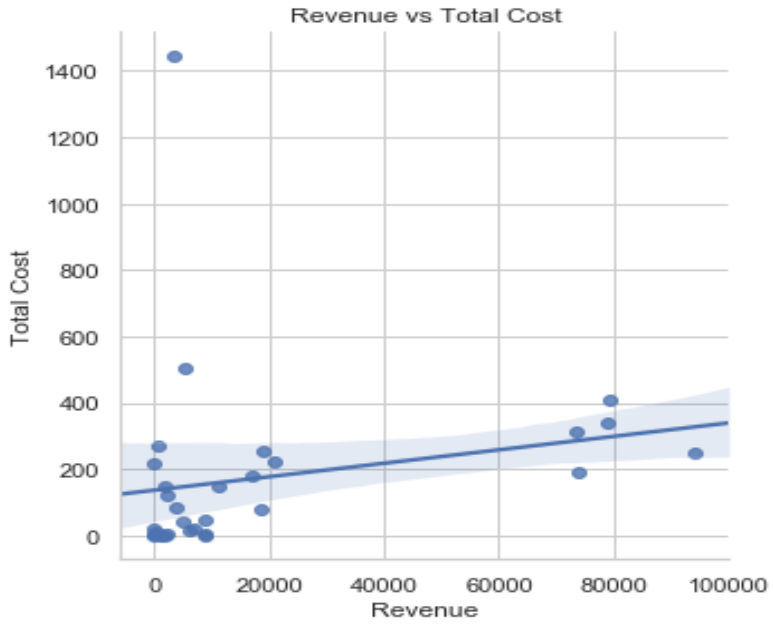

Figure 13. Revenue vs. Total Cost

The figure below shows the linear relationships between PII and the total cost. In this figure, the Equifax is again, far away from the linear line. The number of stolen PII records are shown in scientific notation. The scatter plot shows there is a linear relationship between PII and the total cost.

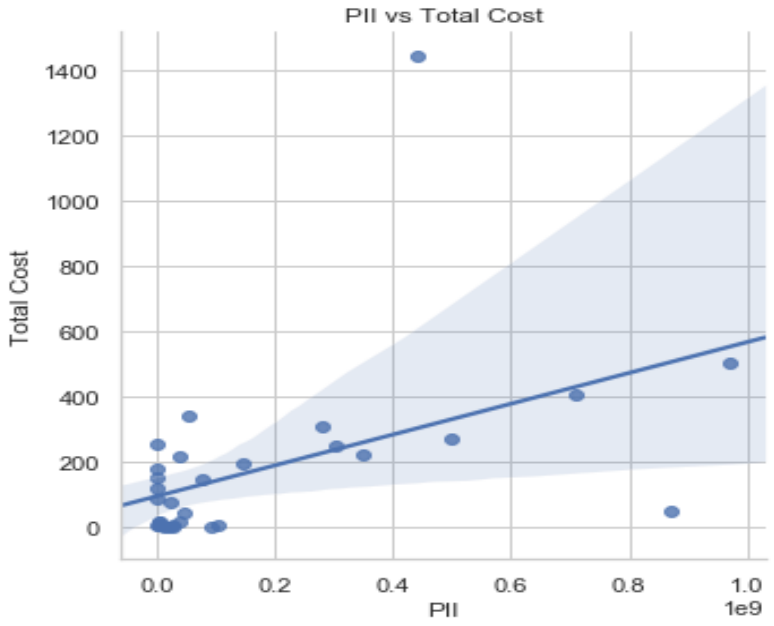

Figure 14. PII vs. Total Cost

The figure below shows the linear relationship between the SPII and the total cost. In this figure, Equifax is again far away from the linear line. The number of stolen SPII records are shown in scientific notation. The scatter plot shows there is a linear relationship between SPII and the total cost.

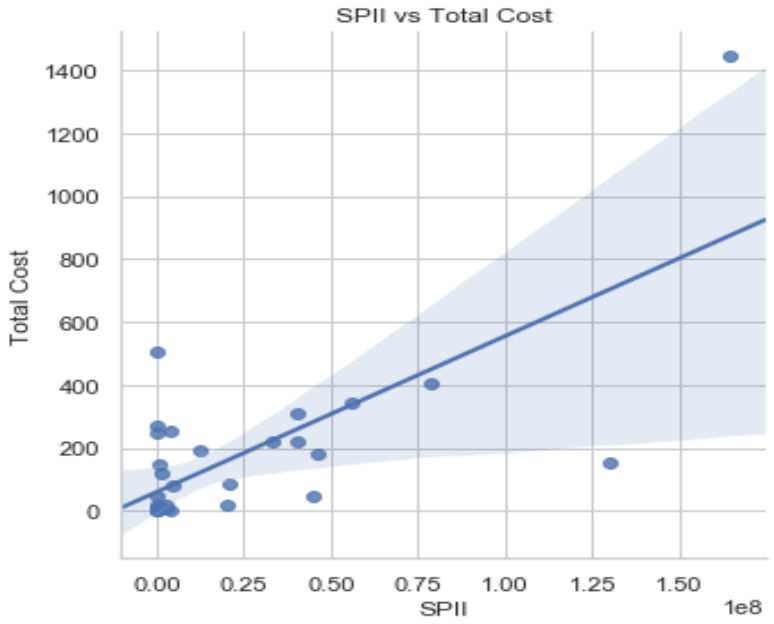

Figure 15. SPII vs. Total Cost

Correlation Matrix and Variance Inflation Factor (VIF)

The next multiple regression assumption is checking if there is multicollinearity among the independent variables. Pearson correlations of the variables and VIF values are calculated and shown below.

Table 11. Pearson Correlation Matrix

	Revenue	PII	SPII	Class-action Lawsuit	Cost
Revenue	1	0.2332	0.1397	0.0991	0.2111
PII	0.2332	1	0.1366	-0.0694	0.4628
SPII	0.1397	0.1366	1	0.1435	0.7147
Class-action Lawsuit	0.0991	-0.0694	0.1435	1	0.1775
Cost	0.2111	0.4628	0.7147	0.1775	1

VIF Values intercept included:

Table 12. VIF values

	Intercept	Revenue	PII	SPII	Class-action lawsuit
VIF	3.3	1.08	1.08	1.05	1.04

As a rule of thumb, VIF values higher than five may pose multicollinearity. All VIF values are below five. Both the correlation matrix and VIF table show that there is no sign of multicollinearity among the independent variables. The next multiple regression assumptions will be checked according to the model results.

4.4 Models

A comparison of the valid models will be made in the next chapter. In this chapter, the outputs of the hypothesized models and a summary are provided.

$E(cost) = \beta_0 + \beta_1 Revenue + \beta_2 Pii + \beta_3 Spii + \beta_4 Class\text{-}action$

OLS Regression Results

Table 13. Model 1 Outputs

R-squared	0.659	Sample size	31	Jarque-Bera	12.432	
Adj. R-squared	0.607	Df Residuals	26	Prob(JB)	0.0002	
Predicted R-squared	0	Df Model	4	Omnibus	8.789	
AIC	1268	Skew	0.485	Prob(Omnibus)	0.012	
BIC	1275	Kurtosis	5.947	Log-likelihood	-628.98	
Mallow Cp	5	Durbin-Watson	2.184	F-statistic	12.58	
PRESS	2.24464E+18					
	coefficient	std err	t	p-value	0.025	0.975
intercept	-34,028,000	55,900,000	-0.609	0.548	-149,000,000	80,900,000
revenue	0.0002	0.001	0.187	0.853	-0.002	0.003
pii	0.3854	0.122	3.165	0.004	.135	.636
spii	4.468	0.815	5.484	0.000	2.793	6.142
Class-action	60,096,800	64.421	0.933	0.359	-72,300,000	192,516,000

In the first multiple regression model, revenue, intercept, and class-action lawsuit variable are statistically non-significant. The backward elimination rule suggests removing the variable with the highest p-value, which is revenue in this case.

In the 2nd run, revenue is eliminated due to high p-value.

Table 14. Model 2 Outputs

R-squared	0.659	Sample size	31	Jarque-Bera	11.015	
Adj. R-squared	0.621	Df Residuals	27	Prob(JB)	0.004	
Pred. R-squared	0.136	Df Model	3	Omnibus	8.151	
AIC	1266	Skew	0.439	Prob(Omnibus)	0.017	
BIC	1272	Kurtosis	5.785	Log-likelihood	-629	
Mallow Cp	3	Durbin-Watson	2.21	F-statistic	17.38	
PRESS	1.93228E+18					
	coefficient	std err	t	p-value	0.025	0.975
intercept	-32,159,234	54,000,000	-0.595	0.557	-143,000,000	78,670,000
pii	0.391	0.116	3.354	0.002	0.152	0.63
spii	4.482	0.796	5.629	0.000	2.848	6.116
Class-action	61,340,000	62,920,000	0.975	0.338	-67,757,000	190,444,000

In model 2, adjusted and predicted R-squared slightly increase. However, there are still statistically not significant variables that intercept and class-action lawsuit variables.

Table 15. Model 3 Outputs

R-squared	0.755	Sample size	31	Jarque-Bera	17.61	
Adj. R-squared	0.729	Df Residuals	28	Prob(JB)	0.000	
Pred. R-squared	0.432	Df Model	3	Omnibus	10.649	
AIC	1264	Skew	0.582	Prob(Omnibus)	0.005	
BIC	1269	Kurtosis	6.505	Log-likelihood	-629	
Mallow Cp	2	Durbin-Watson	2.199	F-statistic	28.81	
PRESS	1.79373e+18					
	coefficient	std err	t	p-value	0.025	0.975
pii	0.3648	0.107	3.415	0.002	1.46e-07	5.84e-07
spii	4.4003	0.775	5.676	0.000	2.81e-06	5.99e-06
Class-action	35,330,000	44,750,000	0.789	0.436	-56,300,000	127,000,000

In model 3, adjusted and predicted R-squared substantially increased; however, there is still a statistically not significant variable that is the class action lawsuit.

Table 16. Model 4 Outputs

R-squared	0.847	Sample size	31	Jarque-Bera	1.979	
Adj. R-squared	0.824	Df Residuals	27	Prob(JB)	0.379	
Pred. R-squared	0.563	Df Model	4	Omnibus	3.393	
AIC	1263	Skew	-0.425	Prob(Omnibus)	0.183	
BIC	1266	Kurtosis	3.899	Log-likelihood	-629	
Mallow Cp	3	Durbin-Watson	1.763	F-statistic	37.23	
PRESS	1.38116e+18					
	coefficient	std err	t	p-value	0.025	0.975
pii	0.3255	0.087	3.754	0.001	0.148	0.503
spii	1.3257	0.990	1.339	0.192	-0.706	3.357
Class-action	-12,680,000	38,000,000	-0.333	0.741	-90,723,000	65,364,000
Spii*class-action	5.054	1.26	4.006	0.000	2.465	7.64

In model 4, the hypothesis is that there is an interaction between SPII and a class-action lawsuit. It is observed that SPII data breaches are more likely to have class-action lawsuits, which can significantly increase the total cost. Therefore, this model has four independent variables that are PII, SPII, class-action lawsuit, SPII * class-action lawsuit. There is an interaction between SPII and class-action variable, which means that SPII data breaches and class-action lawsuit explains the change in the cost and strongly correlated.

According to the current dataset and independent variables, the highest adjusted R-squared value occurred in the last model. However, the difference between adjusted and predicted R-squared looks large enough, which may be a sign of overfitting. Although the SPII and class-action lawsuit variables have high p-values, the interaction is statistically significant. Due to the hierarchical principle, if we include the interaction in the model, we should include

the main effects, which are SPII and class-action lawsuits. Since the rate of R-squared and adjusted R-squared is reasonably high, the following figures will examine if the residual distribution is normal to meet one of the multiple regression assumptions.

The figure below represents the residuals vs. fitted values. Hannaford, Equifax, and Anthem seem to have the highest errors. In this figure, it is better to have a horizontal line, which means there is homoscedasticity. Residuals are expected to appear equally variable across the range of the predicted values.

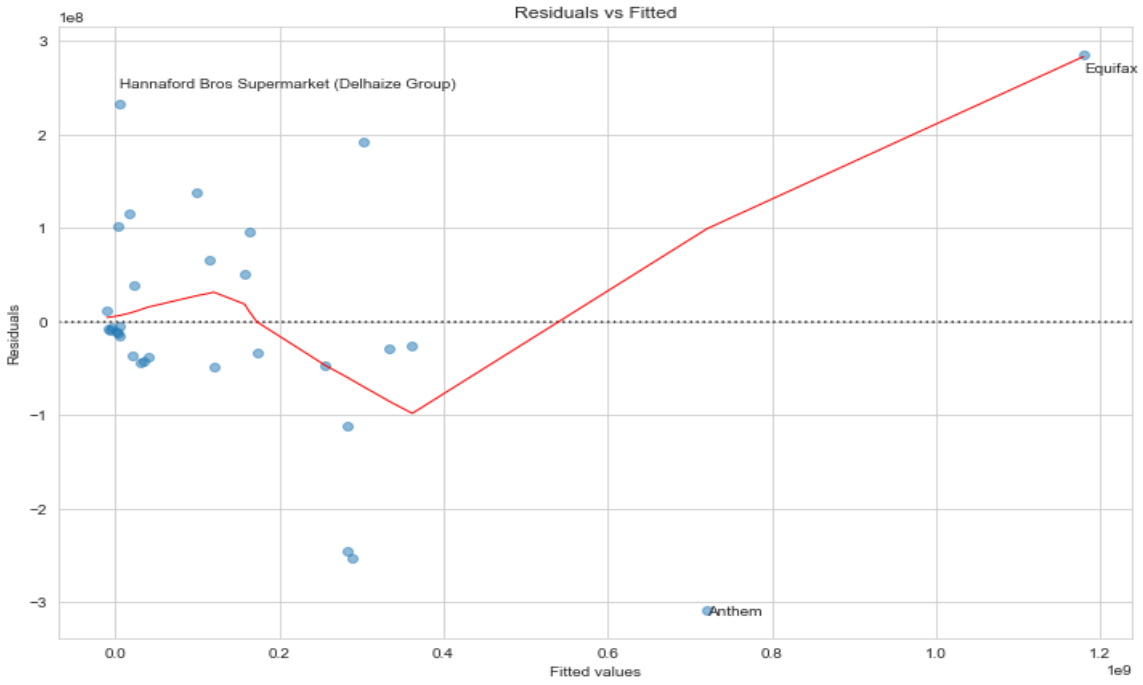

Figure 16. Residual vs. Fitted Values for Model 4

The figure below represents a standardized residuals on Q-Q plot. The residuals are expected to be on the line. However, cases in both tails violate the normality of the residuals. The

standardized residual shows how significant the data are to the chi-square value. For normally distributed residuals, all the residuals are expected to be (+- 2) standard deviation of the mean, which is zero. Also, those points should draw a linear line on a QQ-plot. However, there are three data points beyond two standard deviations, which means that residuals are not normally distributed.

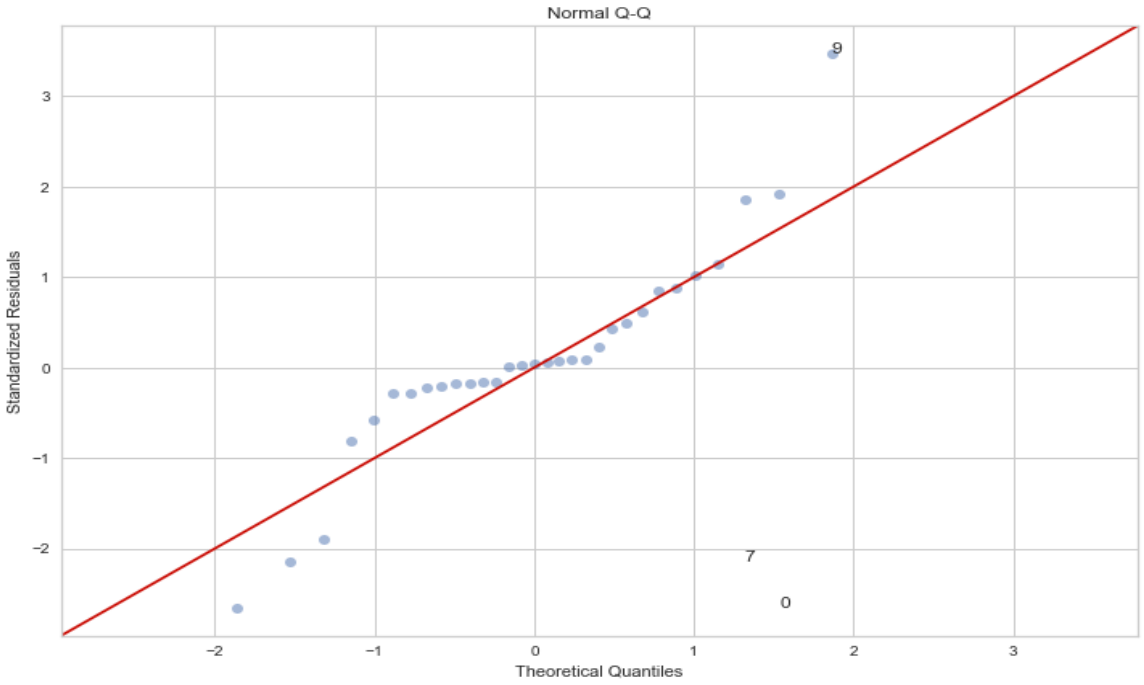

Figure 17. Normal Q-Q Plot for Model 4

The figure below shows the residuals vs. leverage plot. The plot helps to determine any influential cases in the model. There should not be any data point in the upper right corner of the plot. If there is a data point in the dashed lines, Cook's distance, it might have an influential impact on the regression model. Therefore, removing those would change the regression results. In the table, there is not an outlier or influential data point that needs to be removed from the dataset according to Cook's distance.

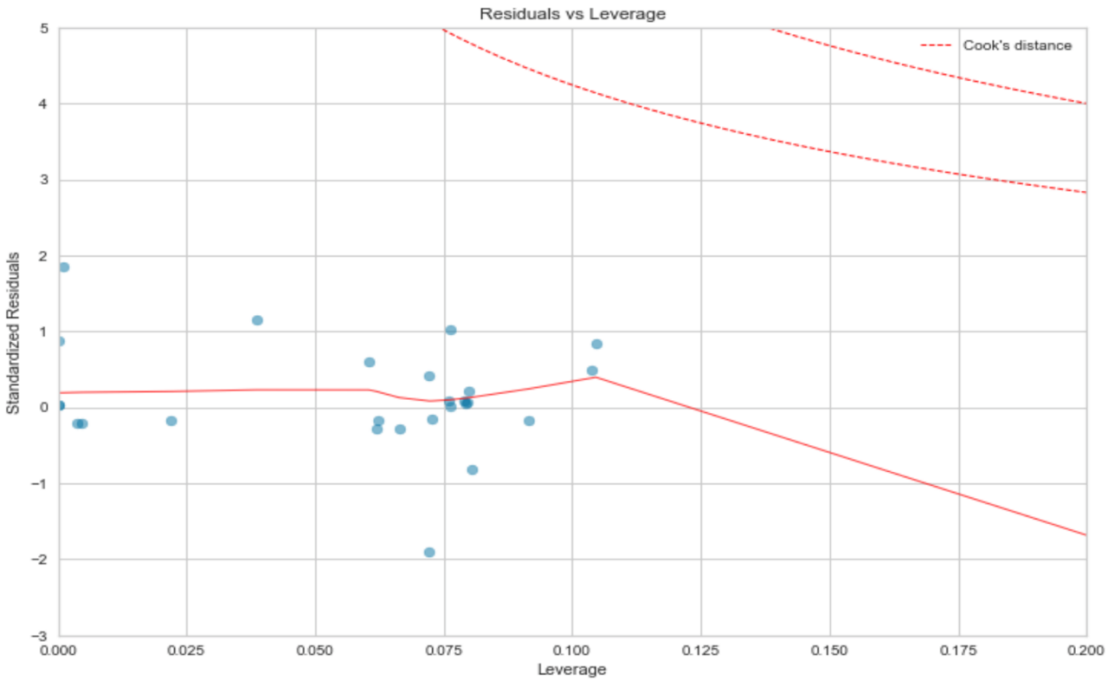

Figure 18. Residuals vs. Leverage Plot for Model 4

Figure 19 represents the scale-location plot. It shows the residuals are spread equally along with the ranges of explanatory variables. This helps to check the assumption of homoscedasticity. If there is an equal variance, there should be a horizontal line with randomly

spread points. However, in the figure, residuals do not form a horizontal line; instead, they begin to spread wider as the fitted value increases. Therefore, it shows that there is a violation of homoscedasticity.

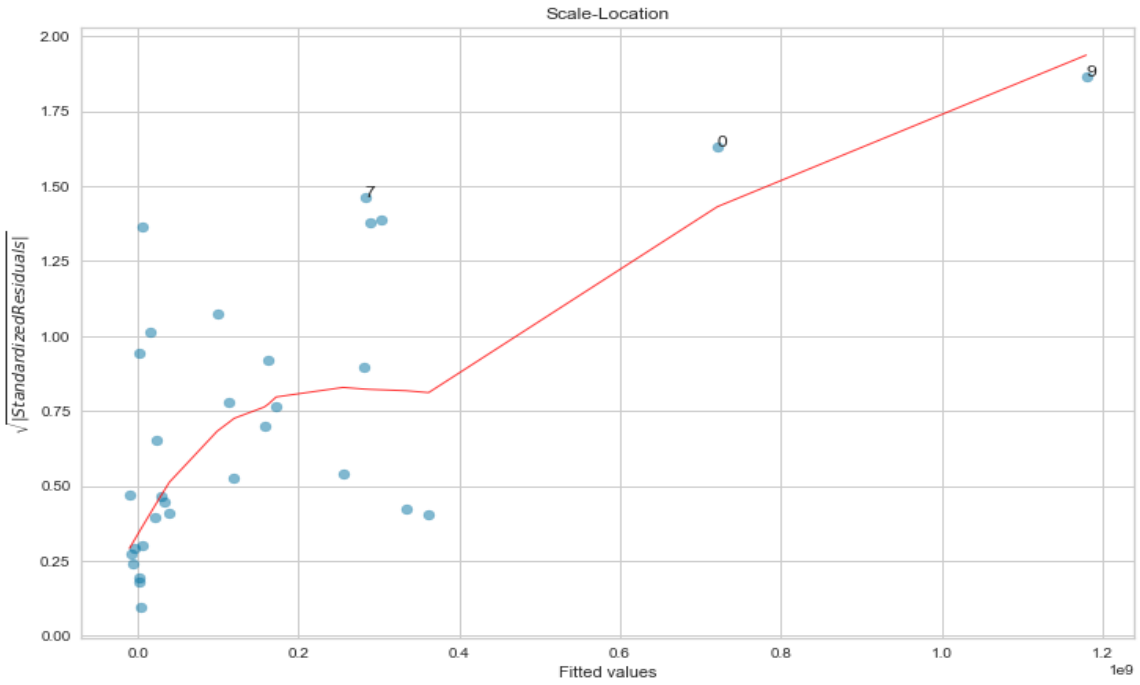

Figure 19. Scale vs. Location Plot for Model 4

The figure below shows the distribution of the residuals on a histogram. The residuals are not normally distributed.

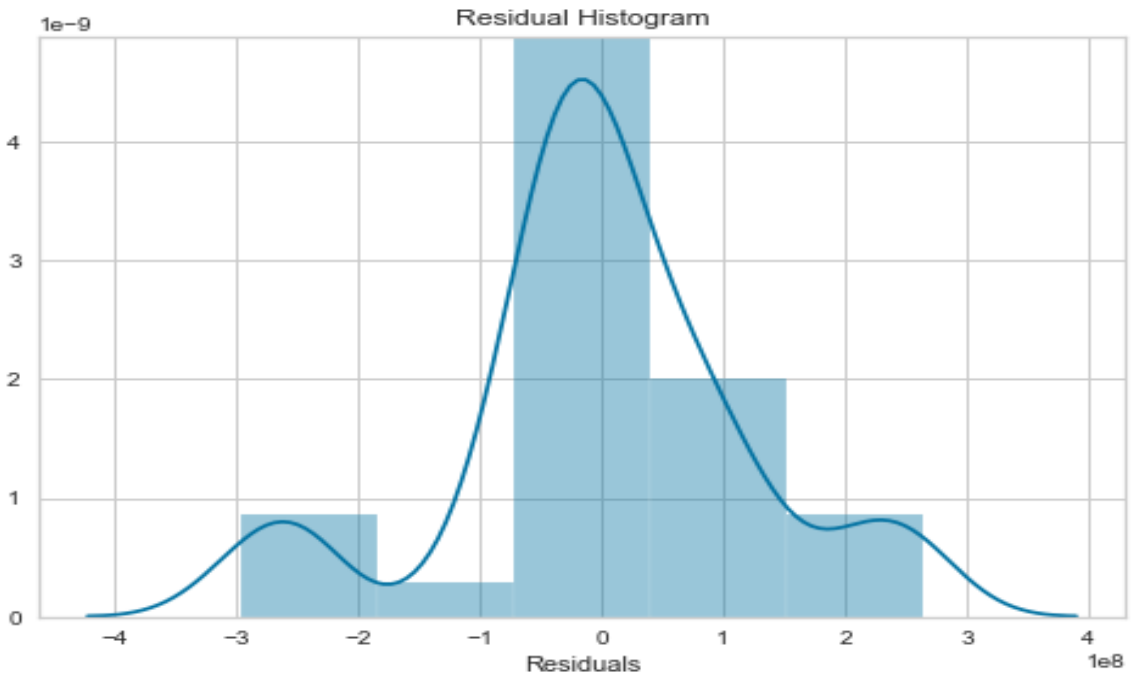

Figure 20. Histogram of the Residuals for Model 4

Breusch-Pagan test

A Breusch-Pagan test is carried out in Python by "statsmodels." According to the Breusch-Pagan test, there is a violation of homoscedasticity, which means that residuals are not normally distributed.

H_0: There is homoscedasticity

H_A: there is heteroscedasticity

In the Breusch-Pagan test, the p-value is much smaller than 0.05. Therefore, the null hypothesis is rejected. The residuals do not have homoscedasticity.

Removing the Outliers

In the next model, the outliers – Equifax and Anthem- are removed from the dataset. As a result, the sample size is 29. The backward elimination is employed.

Table 17. Model 5 Outputs

R-squared	0.586	Sample size	29	Jarque-Bera	2.585	
Adj. R-squared	0.518	Df Residuals	24	Prob(JB)	0.275	
Pred. R-squared	0.089	Df Model	4	Omnibus	3.715	
AIC	1150	Skew	0.213	Prob(Omnibus)	0.156	
BIC	1157	Kurtosis	4.399	Log-likelihood	-570	
Mallow Cp	5	Durbin-Watson		F-statistic	8.5	
PRESS	4.42522E+17					
	coefficient	std err	t	p-value	0.025	0.975
intercept	7,158,000	32,480,000	0.221	0.827	-59,750,000	74,067,000
revenue	0.0019	0.001	2.944	0.007	0.001	0.003
pii	0.2938	0.072	4.072	0	0.145	0.443
spii	1.3969	0.633	2.207	0.037	0.091	2.703
Class-action	32,753,900	35,173,000	0.931	0.361	-39,839,000	105,347,000

The p-value of the intercept is very high in addition to the “class-action” variable. Also, the “spii” variable is close to the alpha level. R-squared and adjusted R-squared are lower than the previous models. In the next model, the intercept is removed.

Table 18. Model 6 Outputs

R-squared	0.788	Sample size	29	Jarque-Bera	2.826	
Adj. R-squared	0.743	Df Residuals	25	Prob(JB)	0.243	
Pred. R-squared	0.57	Df Model	4	Omnibus	3.884	
AIC	1148	Skew	0.209	Prob(Omnibus)	0.143	
BIC	1154	Kurtosis	4.3471	Log-likelihood	-570	
Mallow Cp	4	Durbin-Watson	2.214	F-statistic	21.96	
PRESS	3.90180e+17					
	coefficient	std err	t	p-value	0.025	0.975
revenue	0.002	0.001	3.088	0.005	0.001	0.003
pii	0.3007	0.64	4.71	0	0.169	0.432
spii	1.4442	0.584	2.473	0.021	0.241	2.647
Class-action	38,006,900	25,410,000	1.496	0.147	-14,326,000	90,340,000

The R-squared, predicted, and adjusted R-squared are increased. However, the "class-action" variable is still more than the alpha level, which is 0.05. The "spii" variable now has a lower p-value. In the next model, the "class-action" variable is taken out due to a higher p-value.

Table 19. Model 7 Outputs

R-squared	0.759	Sample size	29	Jarque-Bera	4.079	
Adj. R-squared	0.731	Df Residuals	26	Prob(JB)	0.13	
Pred.R-squared	0.515	Df Model	3	Omnibus	4.652	
AIC	1149	Skew	-0.192	Prob(Omnibus)	0.098	
BIC	1153	Kurtosis	4.797	Log-likelihood	-170	
Mallow Cp	4.2	Durbin-Watson	2.159	F-statistic	27.23	
PRESS	4.400069e+17					
	coefficient	std err	t	p-value	0.025	0.975
revenue	0.0022	0.001	3.574	0.001	0.001	0.004
pii	0.3156	0.065	4.891	0.000	0.183	0.448
spii	1.686	0.574	2.935	0.007	0.505	2.867

Now, all the variables have a lower p-value and statistically significant, although R-squared, predicted, and adjusted R-squared decrease. The interaction effect between the “spii” and the “class-action” variables is not observed. The following figures show how the residuals are distributed.

The following figure shows how the residuals are distributed on a histogram. The skewness value is -0.192, which means the is slightly concentrated on the left side of the mean.

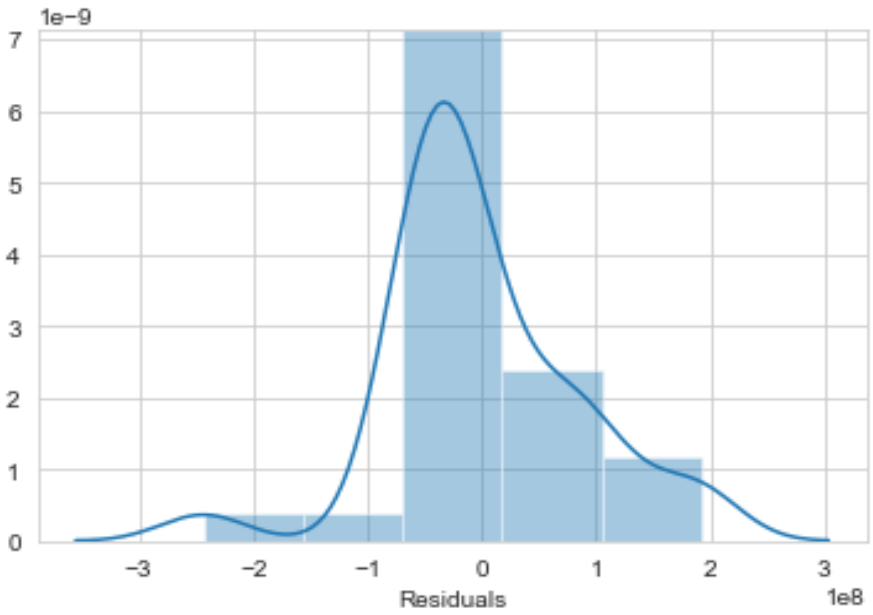

Figure 21. Histogram of the Residuals for Model 7

Figure 22 is a Normal QQ plot of the residuals. There are residuals more than two and even three, which means that a data point beyond three standard deviations.

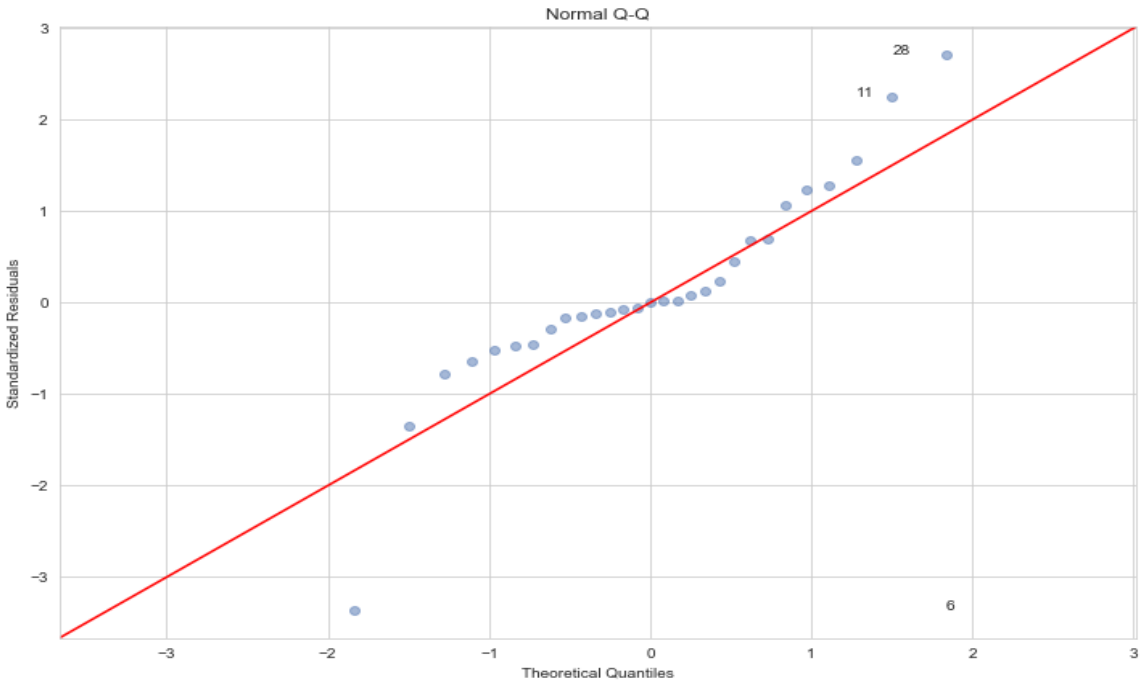

Figure 22. Normal Q-Q Plot for Model 7

Figure 23 is residuals vs. fitted values. The residuals should follow a horizontal line. The errors have been increasing as the fitted values increase. The model fails to predict the high-cost values.

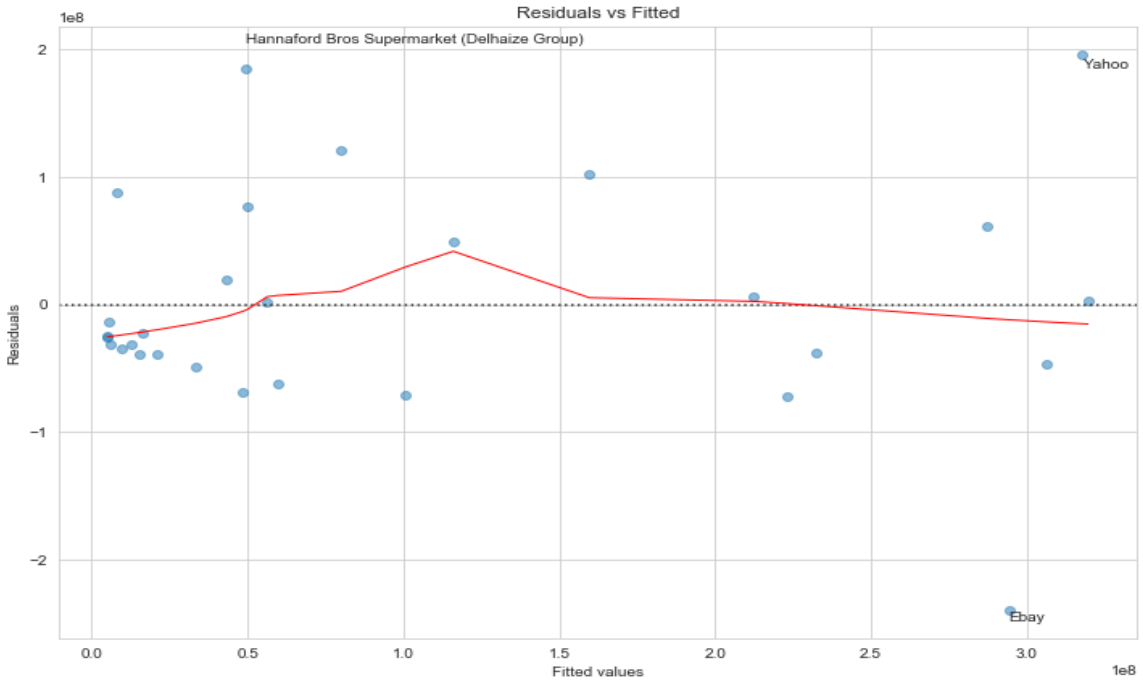

Figure 23. Residuals vs. Fitted Values Plot for Model 7

The figure below illustrates residuals vs. leverage and Cook's distance. There seems to be no data beyond the dashed line in the model.

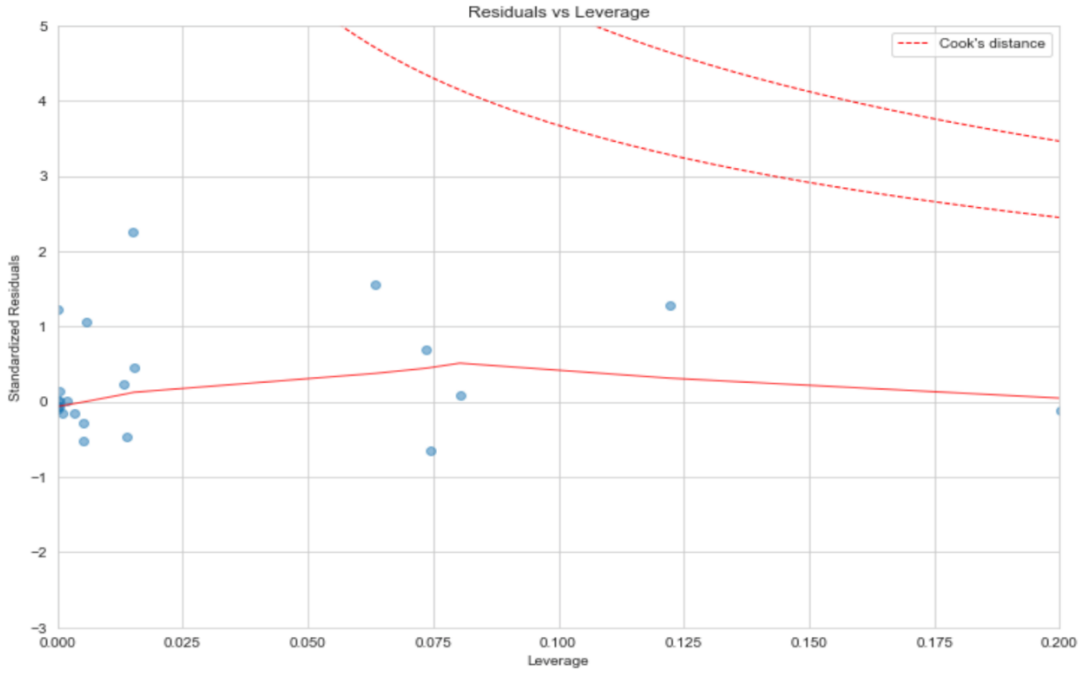

Figure 24. Residuals vs. Leverage plot for Model 7

Figure 25 shows if residuals are spread equally along with the ranges of explanatory variables. This figure helps us to check the assumption of equal variance. We expect to see if there is a horizontal line with randomly spread points.

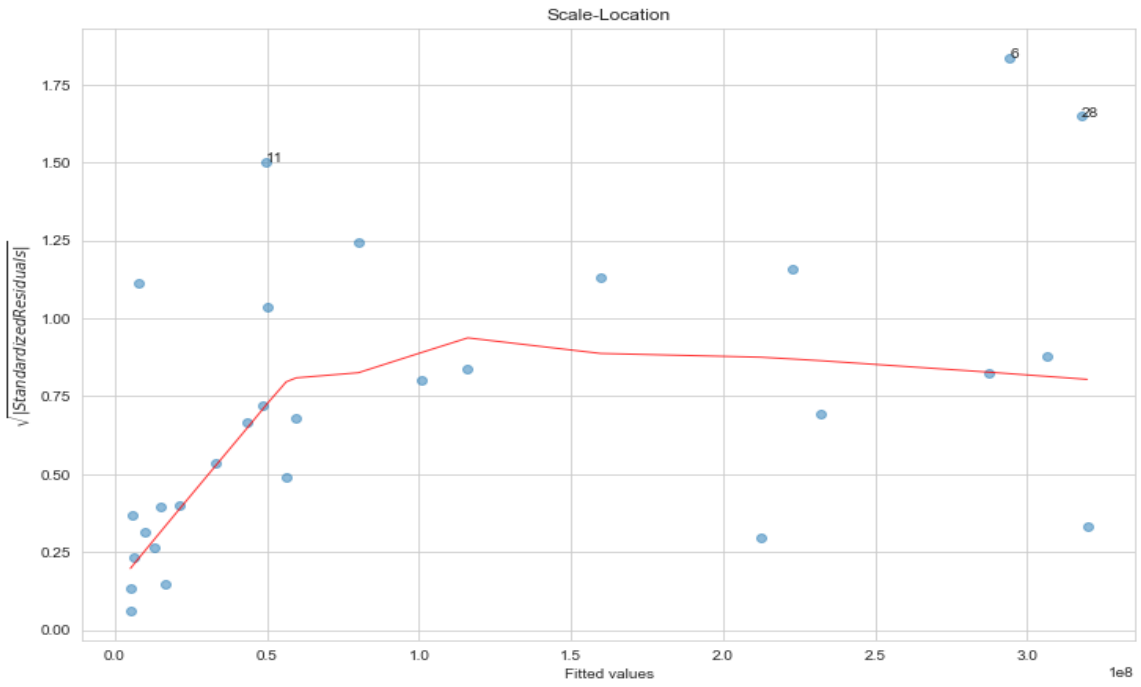

Figure 25. Scale vs Location Plot for Model 7

Breusch-Pagan Test

The test states that residuals are not normally distributed due to the very small p-value. Hence, there will be a transformation of the dependent variable.

Square Root Transformation of the Dependent Variable

Since all dependent variables are greater than zero, square root transformation is applied to meet the multiple regression assumptions. The distribution of the dependent variable after the square root transformation is shown below.

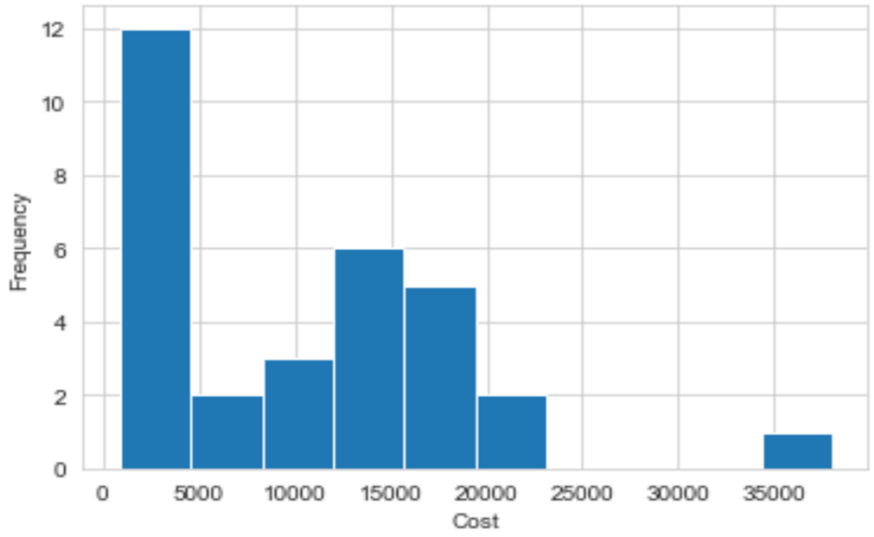

Figure 26. Distribution of the Cost After the Square-root Transformation

Model 8 runs with all variables included. Next backward elimination will be implemented.

Table 20. Model 8 Outputs

R-squared	0.677	Sample size	31	Jarque-Bera	2.038	
Adj. R-squared	0.627	Df Residuals	26	Prob(JB)	0.834	
Pred. R-squared	0.366	Df Model	4	Omnibus	0.768	
AIC	621	Skew	0.289	Prob(Omnibus)	0.681	
BIC	629	Kurtosis	2.442	Log-likelihood	-305	
Mallow Cp	5	Durbin-Watson	2.038	F-statistic	13.59	
PRESS	1,323,695,117					
	coefficient	std err	t	p-value	0.025	0.975
intercept	3672	1,664	2.207	0.036	252	7,093
revenue	6.536e-08	3.41e-08	1.919	0.066	-4.64e-09	1.35e-07
pii	0.000012	3.63e-06	3.414	0.002	4.93e-06	1.98e-05
spii	0.0001	2.43e-05	4.934	0.000	6.98e-05	0
Class-action	1,034	1,918	0.539	0.594	-2.908	4,976

There are two statistically, not significant variables. Also, the difference between adjusted and predicted R^2 looks large.

Model 9 is run after removing the class-action lawsuit variable. The results:

Table 21. Model 9 Outputs

R-squared	0.673	Sample size	31	Jarque-Bera	0.735	
Adj. R-squared	0.637	Df Residuals	27	Prob(JB)	0.693	
Pred. R-squared	0.40	Df Model	3	Omnibus	0.669	
AIC	620	Skew	0.187	Prob(Omnibus)	0.716	
BIC	625	Kurtosis	2.345	Log-likelihood	-306	
Mallow Cp	3.3	Durbin-Watson	2.059	F-statistic	18.51	
PRESS	1,249,528,573					
	coefficient	std err	t	p-value	0.025	0.975
intercept	4,265	1,232	3.462	0.002	1,738	6,793
revenue	6.726e-08	3.34e-08	2.012	0.054	-1.32e-09	1.36e-07
pii	0.000012	3.56e-06	3.421	0.002	4.87e-06	1.98e-05
spii	0.00012	2.37e-05	5.131	0.000	7.29e-05	0

There is still a statistically not significant variable. Also, adjusted and predicted R^2 does not improve.

Model 10 is run after removing the revenue variable. The results:

Table 22. Model 10 Outputs

R-sq	0.624	Sample size	31	Jarque-Bera	1.008	
Adj. R-sq	0.597	Df Residuals	28	Prob(JB)	0.604	
Pred. R-sq	0.43	Df Model	2	Omnibus	1.537	
AIC	622	Skew	-0.027	Prob(Omnibus)	0.464	
BIC	626	Kurtosis	2.118	Log-likelihood	-94	
Mallow Cp	5.2	Durbin-Watson	2.220	F-statistic	23.22	
PRESS	1,177,745,822					
	coefficient	std err	t	p-value	0.025	0.975
intercept	5,083	1,225	4.151	0.000	2,575	7,592
pii	0.000014	3.65e-06	3.756	0.001	6.24e-06	2.12e-05
spii	0.000127	2.48e-05	5.119	0.000	7.61e-05	0

Now all variables are significant; however, adjusted R^2 decreased, unlike predicted R^2. The difference between adjusted and predicted R^2 is considerable; therefore, it is a sign of overfitting.

Interaction between variables is performed. However, it is observed that there is not any significant interaction among variables after the transformation of the dependent variable.

In the next model, the interception is removed, and the model is run with all variables.

Table 23. Model 11 Outputs

R-sq	0.85	Sample size	31	Jarque-Bera	0.796	
Adj. R-sq	0.828	Df Residuals	27	Prob(JB)	0.672	
Pred. R sq	0.775	Df Model	4	Omnibus	1.861	
AIC	625	Skew	0.289	Prob(Omnibus)	0.681	
BIC	630	Kurtosis	3.531	Log-likelihood	-308	
Mallow Cp	4	Durbin-Watson	1.825	F-statistic	38.27	
PRESS	1,206,968,500					
	coefficient	std err	t	p-value	0.025	0.975
revenue	7.872e-08	3.58e-08	2.197	0.037	5.2e-09	1.52e-07
pii	0.000015	3.68e-06	4.056	0.000	7.37e-06	2.25e-05
spii	0.000128	2.56e-05	4.987	0.000	7.52e-05	0.000
Class-action	3,832	1,538	2.491	0.019	675	6989

All the variables in model 11 are statistically significant, and SPII, PII, and class-action have strong positive correlations with the cost. Also, revenue has a positive correlation with data breach cost but not as strong as the other variables. Adjusted and predicted R^2 values are similar. The following figure illustrates the distribution of the residuals. The histogram shows an acceptable normal distribution.

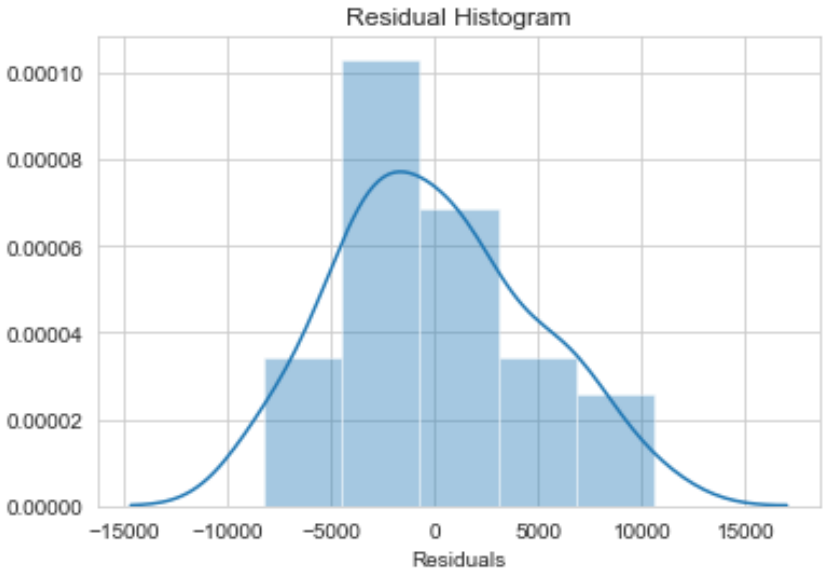

Figure 27. Histogram of the Residuals for Model 11

The figure below shows the residual vs. fitted values. Global Payments company case has the highest error. The line draws close to a horizontal line considering the sample size.

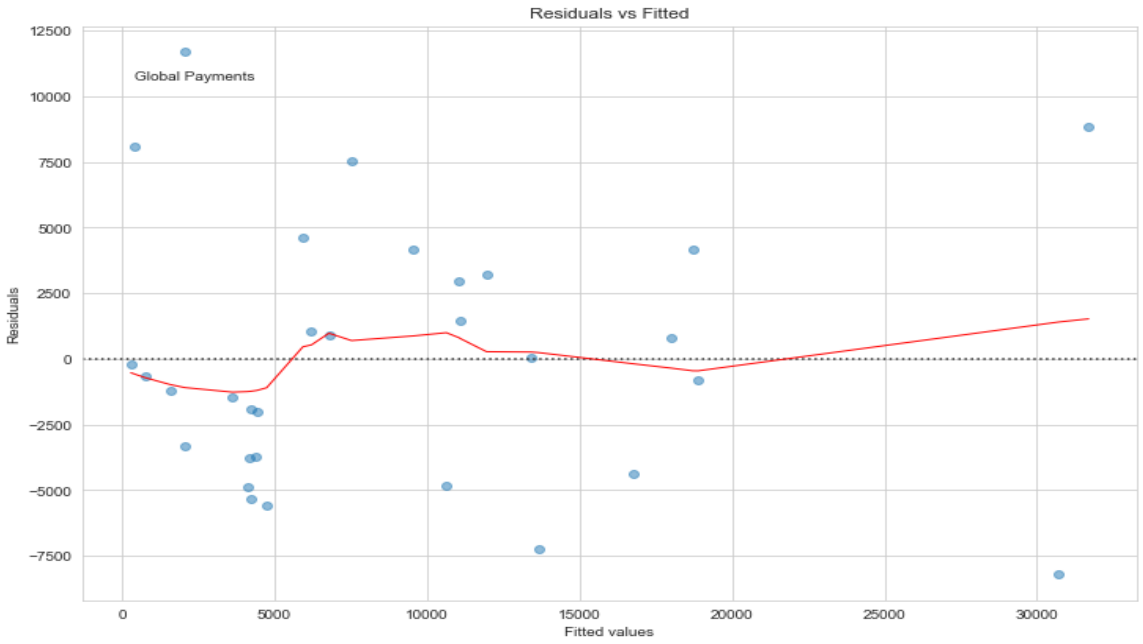

Figure 28. Residuals vs Fitted Values for Model 11

The next figure shows the Normal Q-Q Plot. The points almost draw a straight line. Only two cases- Anthem, and Hannaford-seem to be beyond (+-) standard deviation.

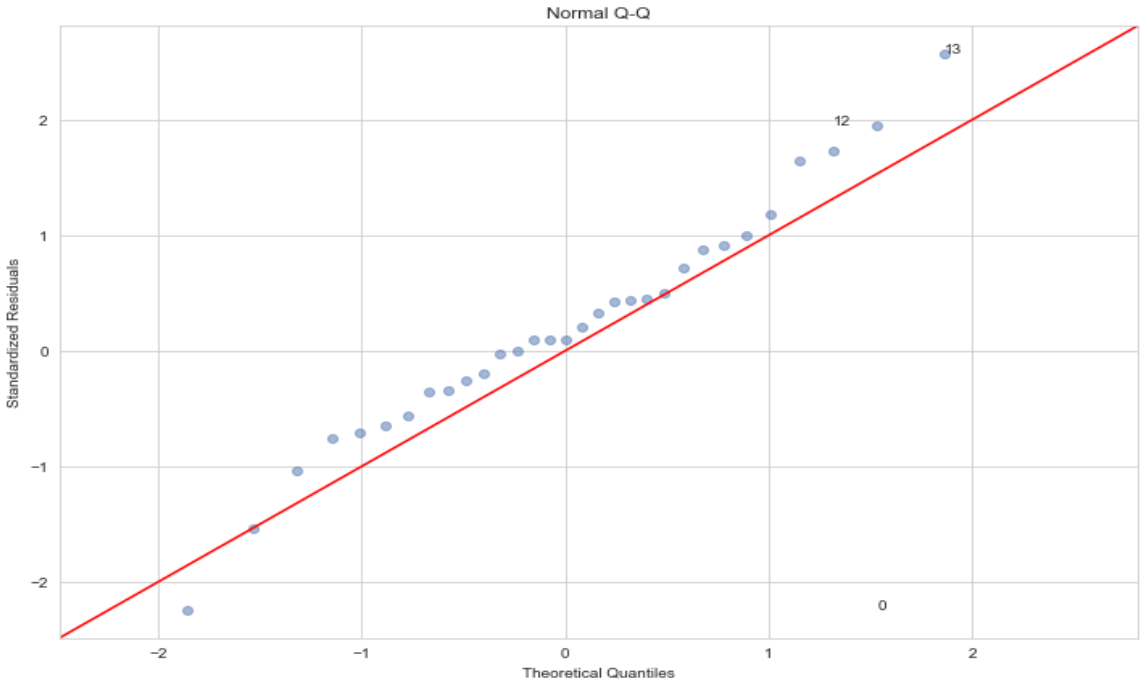

Figure 29. Normal Q-Q Plot for Model 11

The next figure is residual vs. leverage. It shows if there is an influential data point that changes the regression model. There is no data point beyond Cook's distance, which means we do not have an outlier in the model, according to Cook's D.

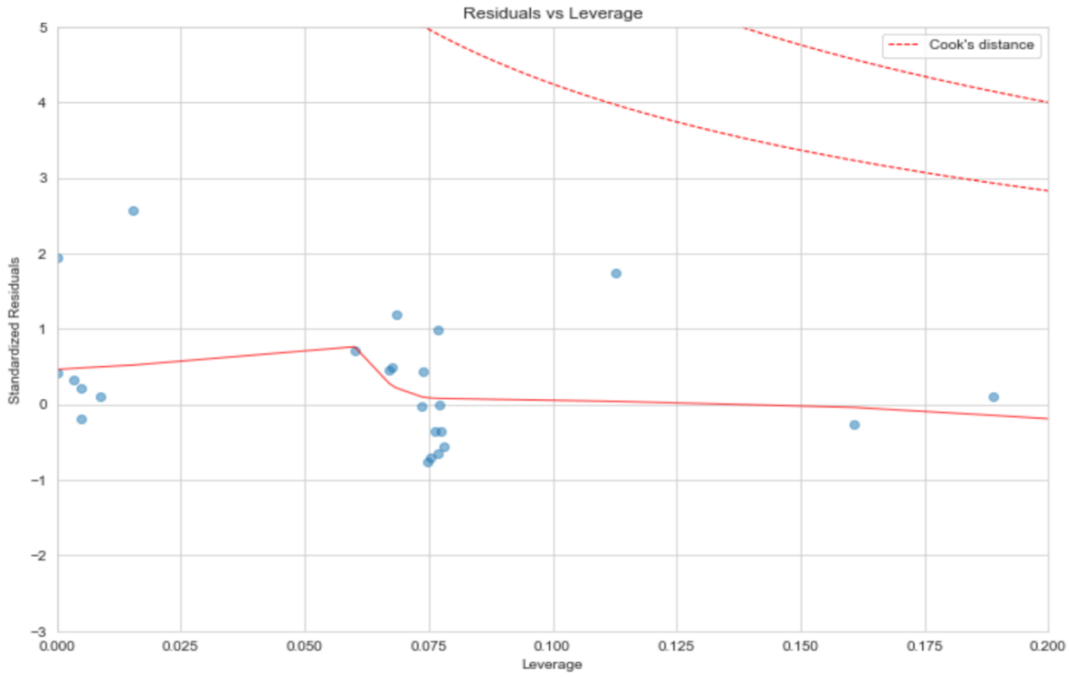

Figure 30. Residual vs. Leverage Plot for Model 11

The next figure tells about the Scale vs. Location. This plot shows that if residuals are spread equally. The output gives the three cases; Anthem, Hannaford, and Global Payments.

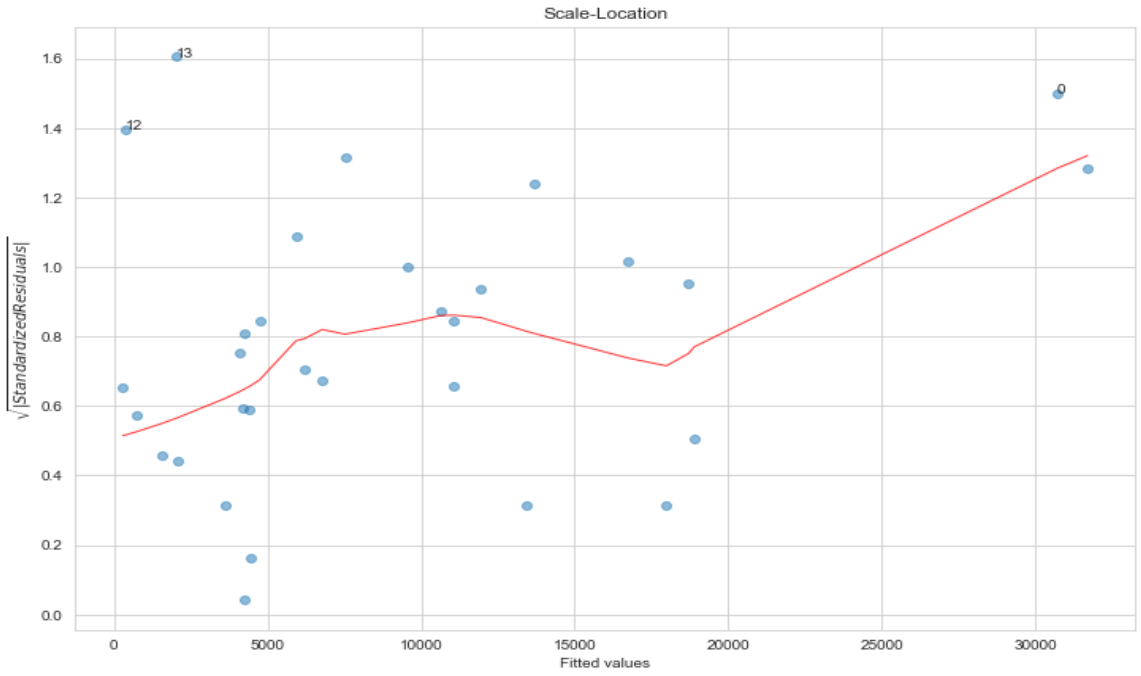

Figure 31. Scale- Location Plot for Model 11

According to the Breusch-Pagan test, the p-value is 0.07, which means residuals are acceptably spread normally. However, outliers (Anthem and Hannaford incidents) are deleted for the next two models.

Table 24. Model 12 Outputs

R-sq	0.77	Sample size	29	Jarque-Bera	0.056	
Adj. R-sq	0.731	Df Residuals	24	Prob(JB)	0.972	
Pred. R sq	0.55	Df Model	4	Omnibus	0.544	
AIC	572	Skew	0.08	Prob(Omnibus)	0.762	
BIC	579	Kurtosis	3.145	Log-likelihood	-281	
Mallow Cp	5	Durbin-Watson	2.672	F-statistic	20.04	
PRESS	875,863,086					
	coefficient	std err	t	p-value	0.025	0.975
intercept	1,535	1,540	0.997	0.329	-1,634	4,714
revenue	8.277e-08	3.06e-08	2.709	0.012	1.97e-08	1.46e-07
pii	1.596e-05	3.28e-06	4.865	0	9.19e-06	2.27e-05
spii	0.0001	2.1e-05	6.258	0	8.8e-05	0.0000
Class-action	2,603	1,692	1,539	0.137	-888	6,095

There are two statistically non-significant variables in the model. In the next, regression model, the highest one, interception, is taken out.

Table 25. Model 13 Outputs

R-sq	0.90	Sample size	29	Jarque-Bera	0.495	
Adj. R-sq	0.88	Df Residuals	25	Prob(JB)	0.781	
Pred. R-sq	0.84	Df Model	4	Omnibus	1.469	
AIC	571	Skew	0.212	Prob(Omnibus)	0.480	
BIC	577	Kurtosis	3.48	Log-likelihood	-281.473	
Mallow Cp	4	Durbin-Watson	2.713	F-statistic	56.4	
PRESS	748,459,963					
	coefficient	std err	t	p-value	0.025	0.975
revenue	8.951e-08	2.98e-08	3.004	0.006	2.81e-08	1.51e-07
pii	0.000017	3.01e-06	5.738	0	1.11e-05	2.35e-05
spii	0.000136	2.06e-05	6.595	0	0.000093	0.000178
Class-action	3,765	1,225	3.073	0.005	1,242	6,289

In the last model, R^2, adjusted $R^{2,}$ and predicted R^2 is significantly increased to 0.9 and 0.88, 0.84, respectively. All the variables are statistically significant. Therefore, the following figures check the assumptions of homoscedasticity and normality of the residuals.

The next figure shows the histogram of the residuals of the last model. The distribution seems to be reasonably normal.

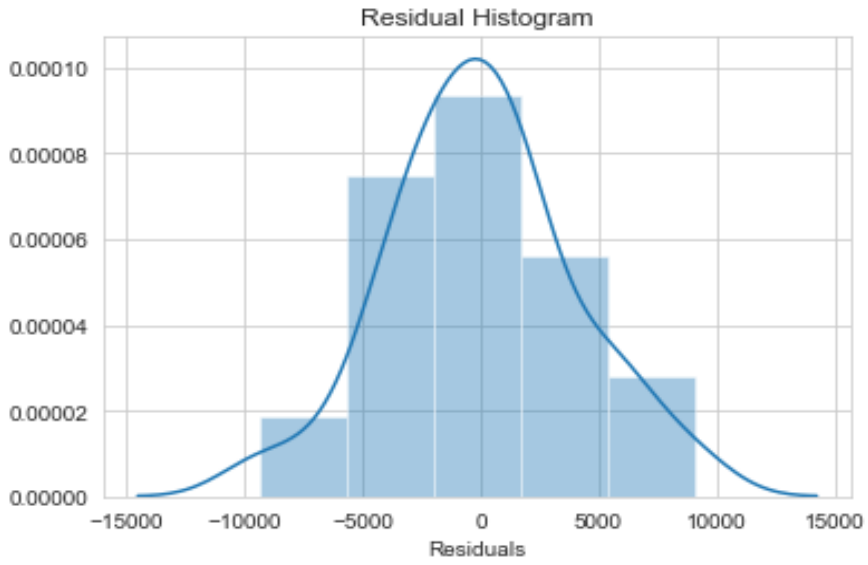

Figure 32. Histogram of the Residuals for Model 13

The next figure illustrates the residual vs. fitted values plot. The residuals should follow a horizontal line. Average of the residuals should be zero, also, they should appear to be equally variable across the entire range of fitted values.

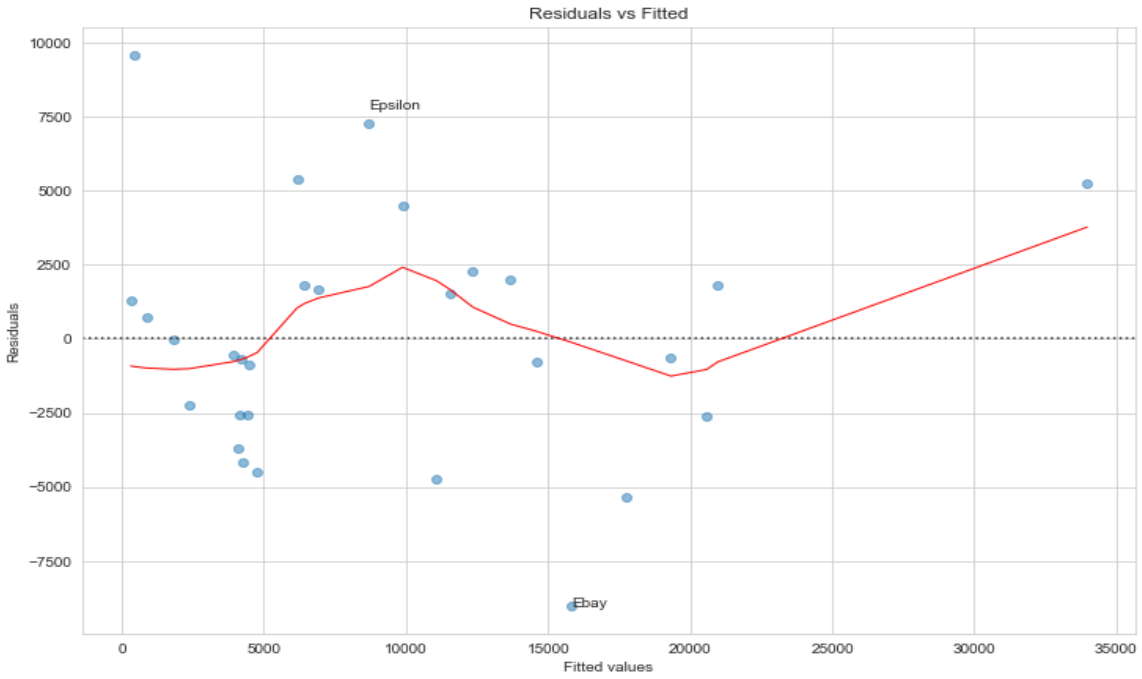

Figure 33. Residuals vs Fitted Values Plot for Model 13

The next figure shows the normal Q-Q Plot. There are only two data points that are slightly beyond (+,-) 2 standard deviation. Nevertheless, the points mostly draw a linear line.

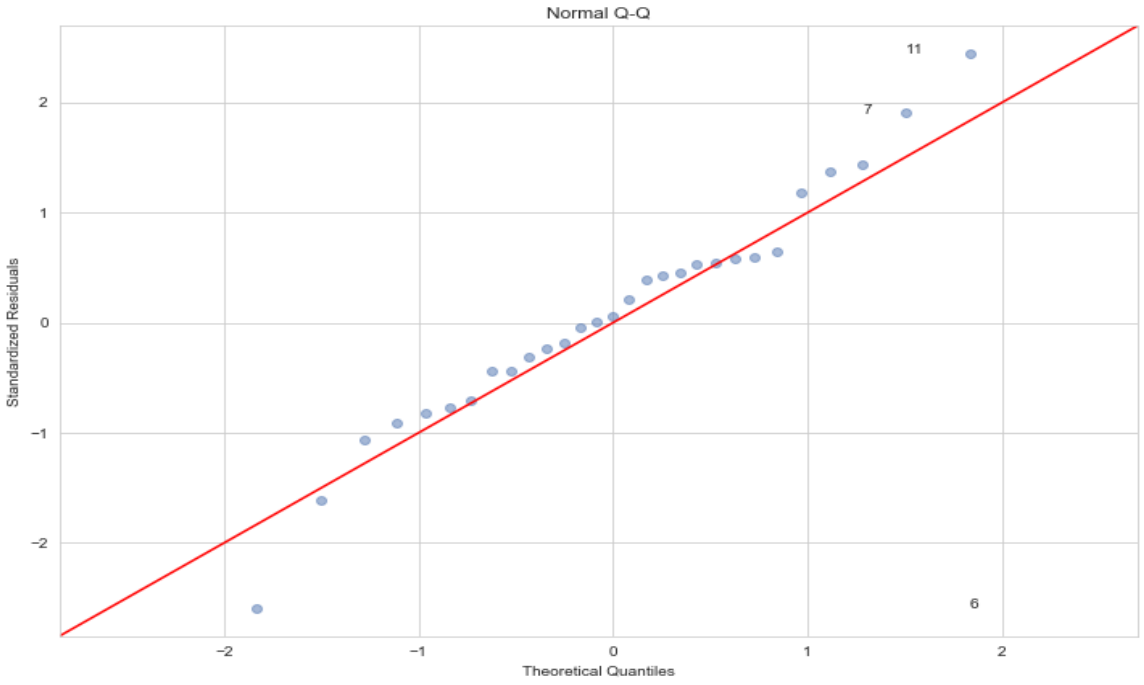

Figure 34. Normal Q-Q Plot for Model 13

The next figure is the residuals vs. leverage plot. According to Cook's D, there is no outlier that appears beyond the dashed line. However, there is one data point that have a high influence on the regression model.

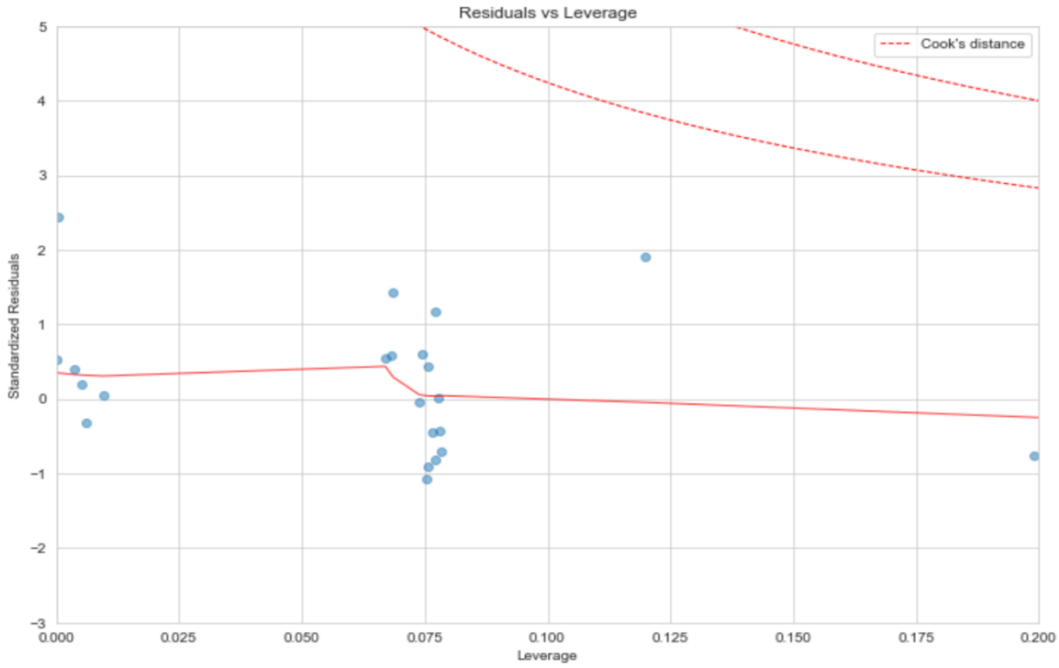

Figure 35. Residuals vs Leverage Plot for Model 13

The next figure is the Scale-Location plot. The plot shows if residuals are spread equally along with the ranges of fitted values.

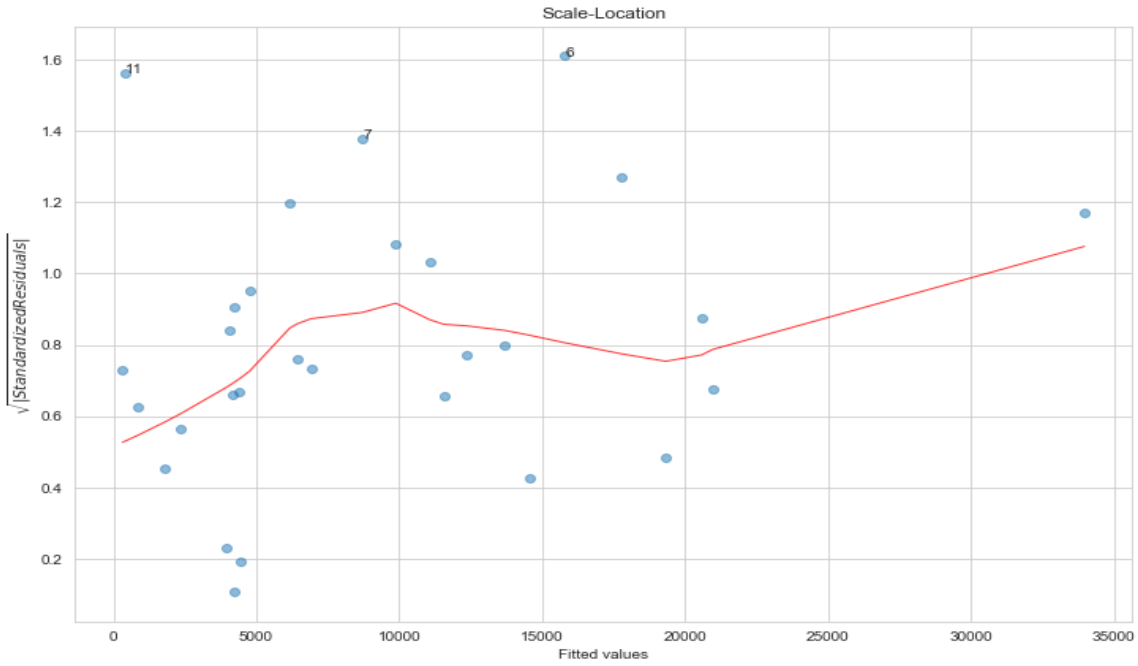

Figure 36. Scale- Location Plot for Model 13

According to the Breusch-Pagan test, the p-value is larger than 0,05. Therefore, it states that residuals are normally distributed. All of the assumptions are acceptably met in this model.

Box-Cox Transformation

The normality of the residuals is one of the assumptions of multiple regression. When the residuals do not show a normal distribution, Box-Cox transformation on the response variable is an option to meet the requirement. Box-Cox transformation can be applied if the response variable is positive.

All values of λ are regarded, and the optimal value for the dataset is assigned. The optimal value provides the best normal distribution curve. The transformation of the response variable follows (Box & Cox, 1964):

$Y(\lambda) = (y^{\lambda} - 1) / \lambda \qquad \text{if } \lambda \neq 0$

$Y(\lambda) = \log y \qquad \text{if } \lambda = 0$

The dependent variable, total cost, is transformed with Box-Cox by "SciPy" library in Python to have a more normal distribution. The optimal lambda value is determined as 0.1534 by the SciPy library, which provides the best approximation of a normal distribution curve. The histogram of the cost after Box-Cox transformation is illustrated in the figure below.

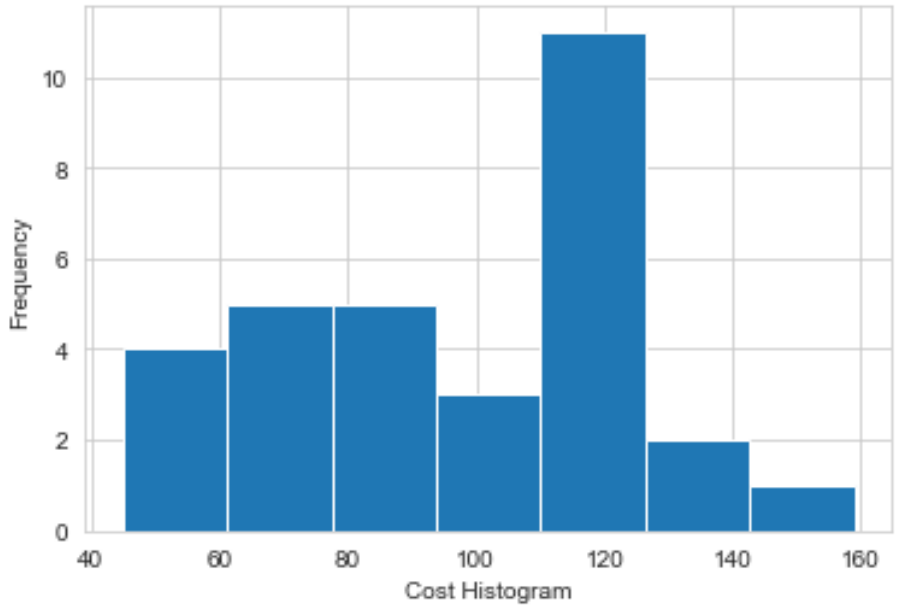

Figure 37. Histogram of the Cost After Box-Cox Transformation

All predictor variables are included in the first model after the Box-Cox transformation

Table 26. Model 14 Outputs

R-sq	0.572	Sample size	31	Jarque-Bera	1.05	
Adj. R-sq	0.506	Df Residuals	26	Prob(JB)	0.592	
Pred. R-sq	0.39	Df Model	4	Omnibus	1.645	
AIC	282	Skew	0.059	Prob(Omnibus)	.439	
BIC	289	Kurtosis	2.106	Log-likelihood	-136	
Mallow Cp	5	Durbin-Watson	2.017	F-statistic	8.684	
PRESS	16,481					
	coefficient	std err	t	p-value	0.025	0.975
intercept	75	6.916	10.827	0.00	60.667	89.1
revenue	3.296 e-10	1.42e-10	2.328	0.028	3.86e-11	6.2e-10
pii	3.912e-08	1.51e-08	2.596	0.015	8.15e-09	7.01e-08
spii	3.585e-07	1.01e-07	3.556	0.001	1.51e-07	5.66e-07
Class-action	0.1247	7.971	0.016	0.988	-16.26	16.509

The class-action lawsuit has a very high p-value. Therefore, in the next model, it is taken out.

Table 27. Model 15 Outputs

R-sq	0.572	Sample size	31	Jarque-Bera	1.06	
Adj. R-sq	0.524	Df Residuals	27	Prob(JB)	0.592	
Pred. R-sq	0.44	Df Model	3	Omnibus	1.681	
AIC	280	Skew	0.058	Prob(Omnibus)	.432	
BIC	285	Kurtosis	2.101	Log-likelihood	-136	
Mallow Cp	5.4	Durbin-Watson	2.018	F-statistic	12.02	
PRESS	15,173					
	coefficient	std err	t	p-value	0.025	0.975
intercept	74.95	5.092	14.72	0.000	64.5	85.4
revenue	3.297e-10	1.38e-10	2.387	0.024	4.63e-11	6.13e-10
pii	3.909e-08	1.47e-08	2.66	0.013	8.94e-09	6.92e-08
spii	3.587e-07	9.79e-08	3.664	0.001	1.58e-07	5.6e-07

R^2 has not changed, but adjusted R^2 slightly has increased. Although R^2 and adjusted and predicted R^2 is slightly increased, all independent variables seem statistically significant. The following figures tell about the residual distribution. The histogram shows a normal distribution.

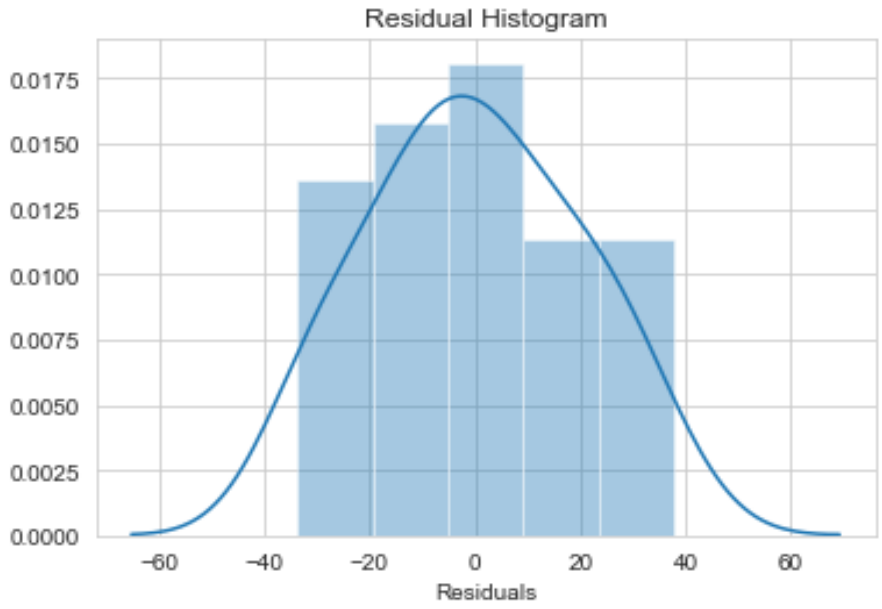

Figure 38. Histogram of the Residuals for Model 15

The Normal Q-Q plot is shown in the figure below. All of the standardized residuals are within -+2. Breusch-Pagan Test is performed for the model. According to the test, the residuals are normally distributed.

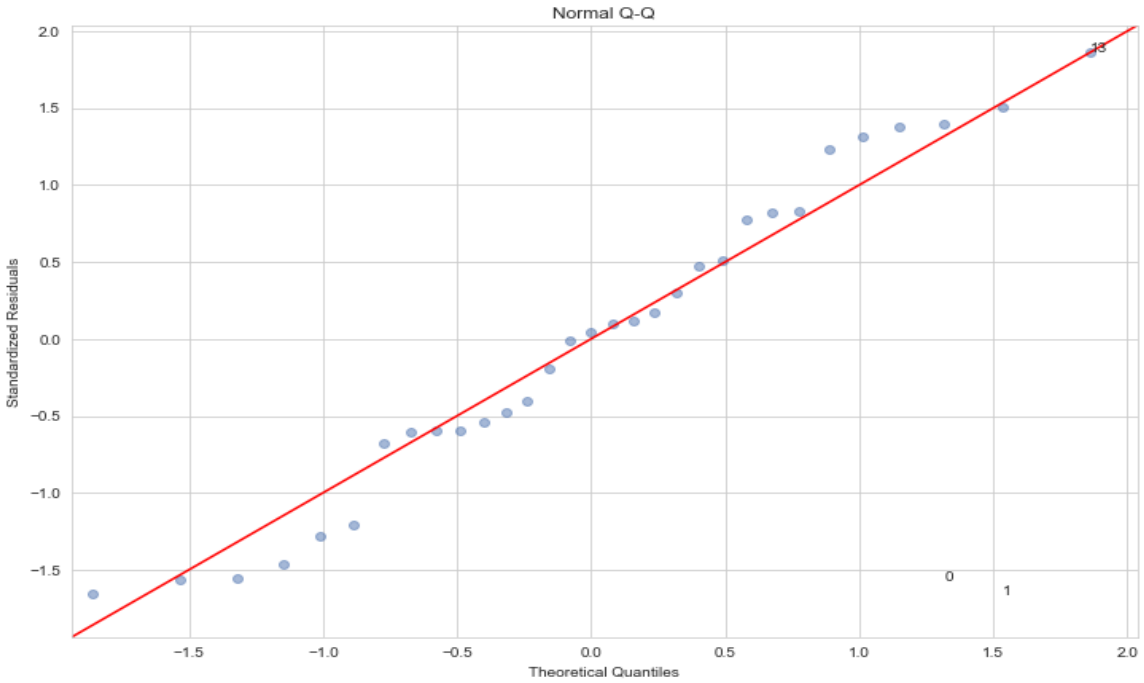

Figure 39. Normal Q-Q Plot for Model 15

The next figure illustrates the Residual vs. Leverage plot. It is observed that there is not a data point beyond the dashed lines.

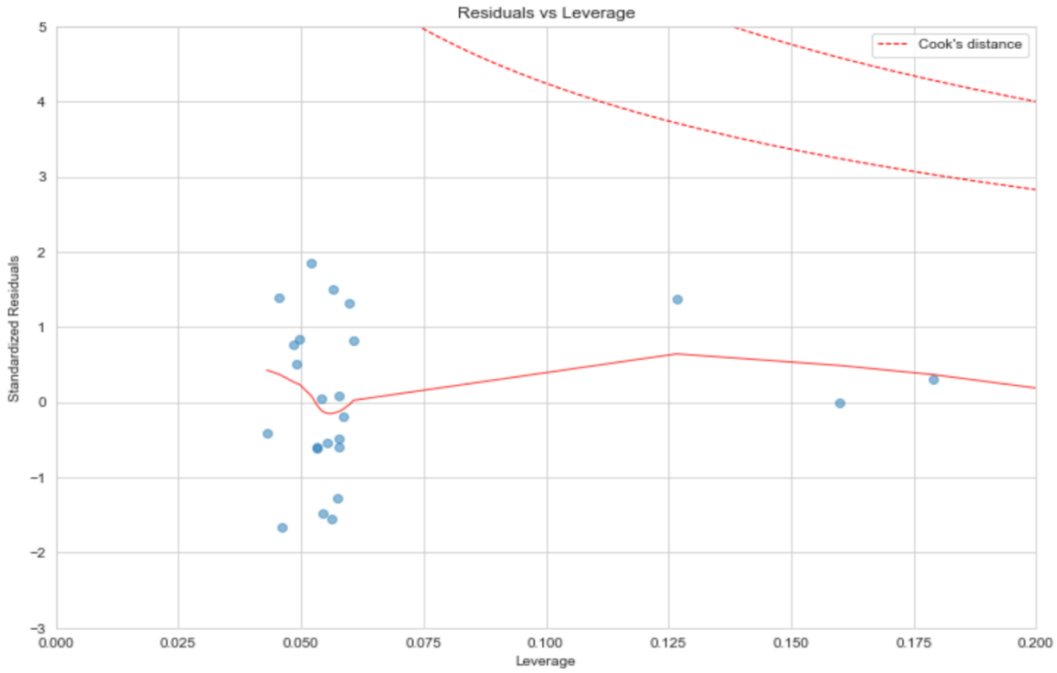

Figure 40. Residuals vs. Leverage Plot for Model 15

The next figure depicts the Residuals vs. Fitted values plot. The Hannaford case has the highest residual.

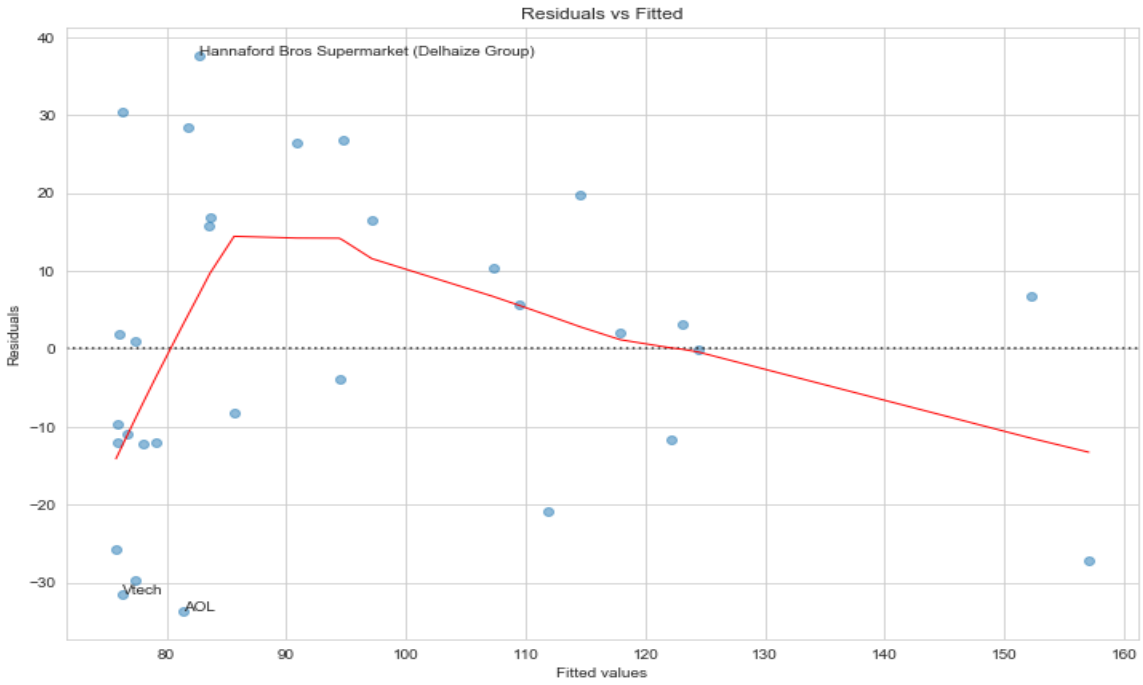

Figure 41. Residuals vs Fitted Values Plot for Model 15

The figure below shows how residuals are spread along with the ranges of predictors.

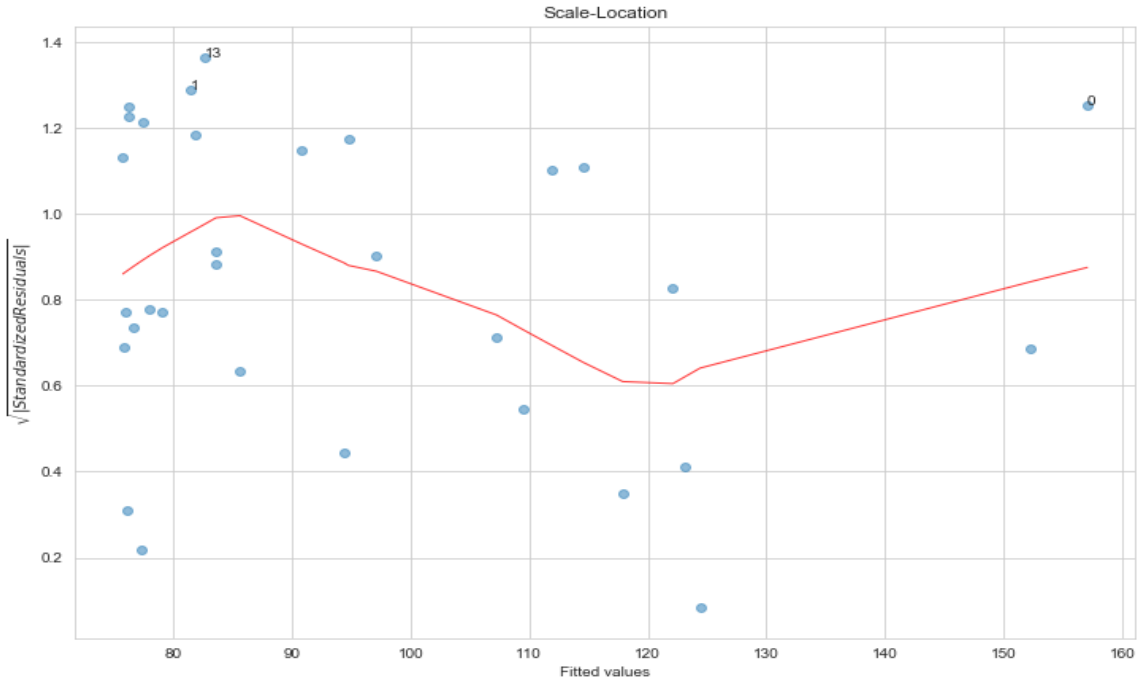

Figure 42. Scale-Location Plot for Model 15

In the next model, the intercept is removed.

Table 28. Model 16 Outputs

R-sq	0.79	Sample size	31	Jarque-Bera	0.096	
Adj. R-sq	0.76	Df Residuals	27	Prob(JB)	0.953	
Pred. R-sq	0.74	Df Model	4	Omnibus	0.725	
AIC	332	Skew	-0.019	Prob(Omnibus)	0.696	
BIC	338	Kurtosis	3.27	Log-likelihood	-162	
Mallow Cp	4	Durbin-Watson	1.364	F-statistic	28.84	
PRESS	79,389					
	coefficient	std err	t	p-value	0.025	0.975
revenue	6.021e-10	3.21e-10	1.877	0.07	-5.61e-11	1.26e-09
pii	9.079e-08	3.29e-08	2.758	0.01	2.32e-08	1.58e-07
spii	5.249e-07	2.29e-07	2.287	0.03	5.41e-08	9.96e-08
Class-action	57.1830	13.773	4.152	0.000	28.922	85.444

We have an increased adjusted and predicted R^2. The difference between adjusted and predicted R^2 is small. Only the revenue variable is slightly larger than the alpha value. Therefore, two data points will be removed from the dataset for the following models.

Removing the outliers (Hannaford and Global Payments case)

Table 29. Model 17 Outputs

R-sq.	0.67	Sample size	29	Jarque-Bera	0.694	
Adj. R-sq.	0.62	Df Residuals	24	Prob(JB)	0.707	
Pred. R.sq	0.52	Df Model	4	Omnibus	0.604	
AIC	208	Skew	-0.197	Prob(Omnibus)	0.739	
BIC	215	Kurtosis	2.35	Log-likelihood	-99	
Mallow Cp	5	Durbin-Watson	2.236	F-statistic	11.3	
PRESS	12,576					
	coefficient	std err	t	p-value	0.025	0.975
intercept	35.20	2.965	11.87	0.000	29.085	41.324
revenue	1.353e-10	5.46e-11	2.478	0.02	2.26 e-11	2.48e-10
pii	1.91e-08	5.9e-09	3.3	0.003	6.93e-09	3.13e-08
spii	1.539e-08	3.88 e-08	3.96	0.001	7.38e-08	2.34e-07
Class-action	2.92	3.253	0.9	0.37	-3.786	9.64

Since the class-action lawsuit variable is statistically not significant, it is removed for the next model.

Table 30. Model 18 Outputs

R-sq	0.64	Sample size	29	Jarque-Bera	0.835	
Adj. R-sq	0.60	Df Residuals	25	Prob(JB)	0.66	
Pred. R sq	0.53	Df Model	3	Omnibus	1.039	
AIC	207	Skew	-0.07	Prob(Omnibus)	0.6	
BIC	212	Kurtosis	2.18	Log-likelihood	-99	
Mallow Cp	4	Durbin-Watson	2.13	F-statistic	15	
PRESS	12,378					
	coefficient	std err	t	p-value	0.025	0.975
intercept	37.09	2.089	17.265	0.000	32.788	41.394
revenue	1.404e-10	5.41e-11	2.596	0.016	2.9e-11	2.52 e-10
pii	2.822e-08	5.79e-09	3.145	0.004	6.3e-09	3.02e-08
spii	1.578e-07	3.84e-08	4.106	0.001	7.87e-07	2.37e-07

Although all variables have a p-value lower than 0.05, R^2, predicted, and adjusted R^2 is still low. Therefore, in the next model, the intercept is taken out.

Table 31. Model 19 Outputs

R-sq	0.84	Sample size	29	Jarque-Bera	0.775	
Adj. R-sq	0.81	Df Residuals	25	Prob(JB)	0.679	
Pred. R sq	0.79	Df Model	4	Omnibus	1.259	
AIC	262	Skew	-0.4	Prob(Omnibus)	0.533	
BIC	267	Kurtosis	2.97	Log-likelihood	-127	
Mallow Cp	4	Durbin-Watson	1.79	F-statistic	32	
PRESS	57,625					
	coefficient	std err	t	p-value	0.025	0.975
revenue	2.338e-10	1.39e-10	1.687	0.10	-5.16e-11	5.19e-10
PII	4.443e-08	1.41e-08	3.146	0.004	1.53e-08	7.35e-08
SPII	2.307e-07	9.84e-07	2.345	0.027	2.81e-08	4.33e-07
Class-action	30.22	5.911	5.11	0.000	18.05	42.40

Although we have a good adjusted and predicted R^2 that explains the variance in the response variable, we still have a statistically not significant variable. Since the revenue variable is statistically not significant, it will be removed from the next model.

Table 32. Model 20 Outputs

R-sq	0.82	Sample size	29	Jarque-Bera	0.614	
Adj. R-sq	0.80	Df Residuals	26	Prob(JB)	0.679	
Pred. R sq	0.78	Df Model	3	Omnibus	0.736	
AIC	263	Skew	-0.205	Prob(Omnibus)	0.786	
BIC	267	Kurtosis	2.417	Log-likelihood	-128	
Mallow Cp	4	Durbin-Watson	1.91	F-statistic	39.6	
PRESS	61,305					
	coefficient	std err	t	p-value	0.025	0.975
pii	5.173e-08	1.39e-08	3.718	0.001	2.31e-08	8.03e-08
spii	2.514e-07	1.01e-07	2.489	0.020	4.37e-08	4.59e-07
Class-action	33.224	5.83	5.702	0.000	21.26	45.22

Now, all variables are statistically significant, and we have a good adjusted and predicted R^2 number. Therefore, the next figures will explore if the model meets the homoscedasticity assumption.

Breusch-Pagan test is employed to test if the residuals' distribution is normal. The result states that the distribution of the residuals is not normal.

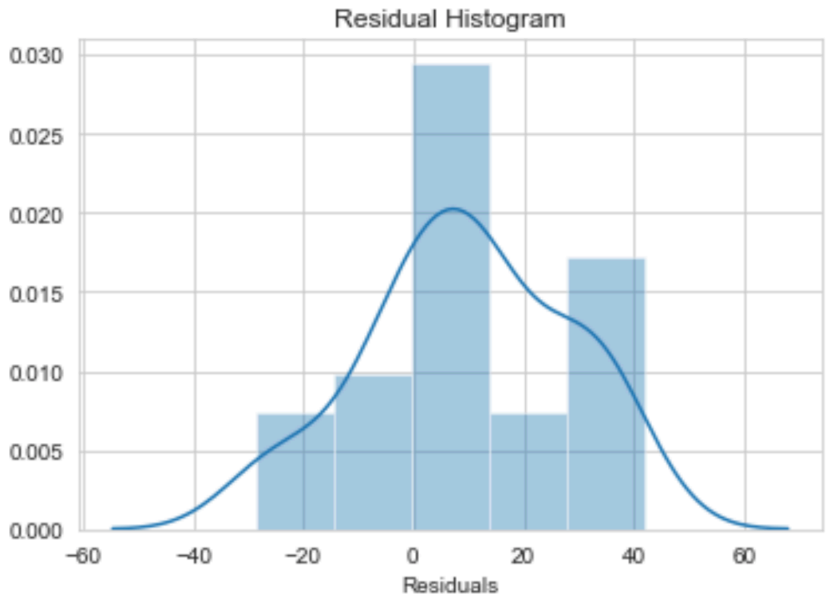

Figure 43. Histogram of the Residuals for Model 20

The next figure is a Q-Q plot. The residuals are expected to be on the line. For a normally distributed residuals, all the residuals are expected to be within (+- 2) standard deviation of the mean. All residuals are within (+-2) standard deviation; however, they do not draw a straight line.

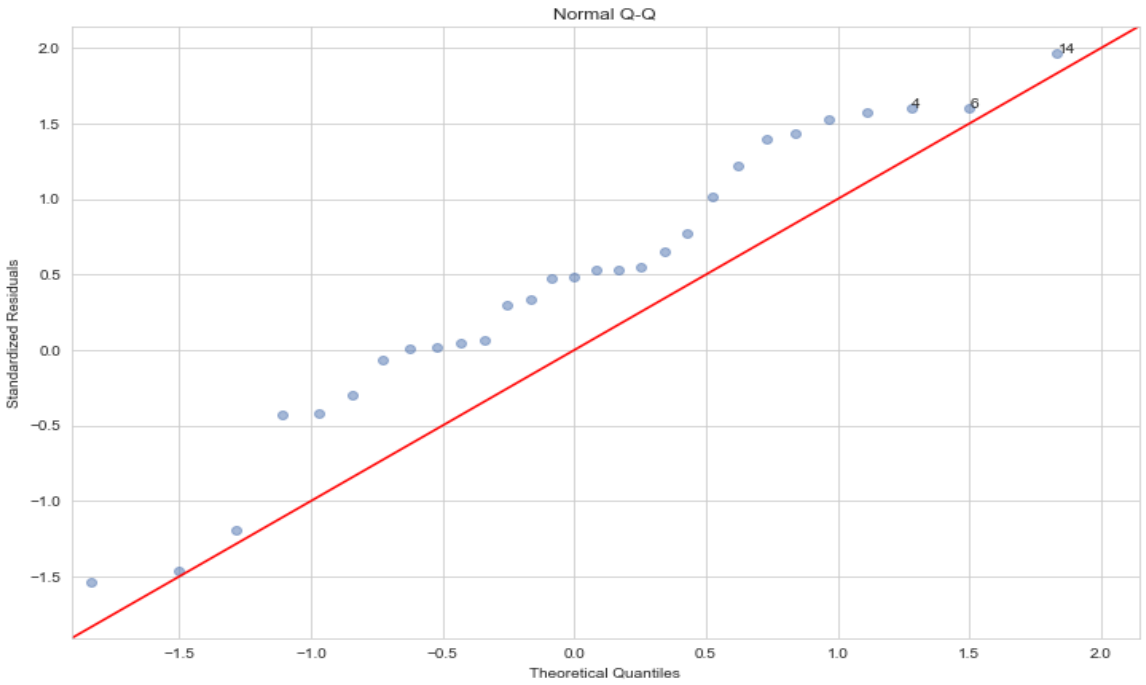

Figure 44. Normal Q-Q Plot for Model 20

The next figure shows the residuals vs. fitted values. In this plot, there should be a horizontal line to satisfy the homoscedasticity condition of the multiple regression assumptions.

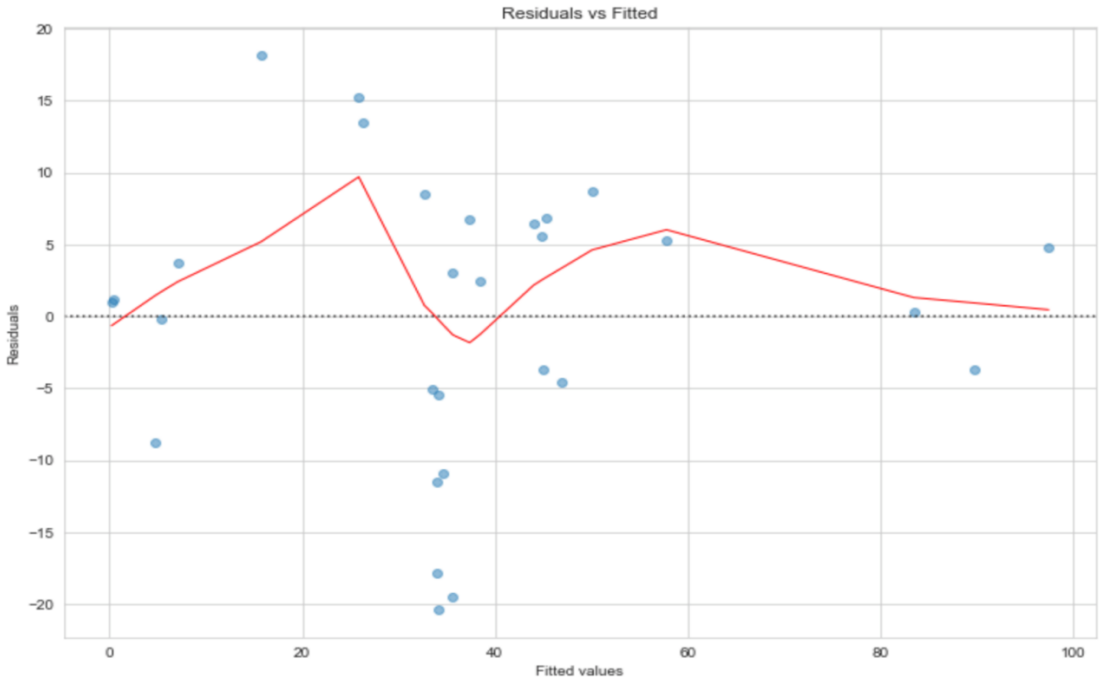

Figure 45. Residuals vs Fitted Values Plot for Model 20

The next plot shows residuals vs. leverage. Here, there should not be any dot beyond the dashed line.

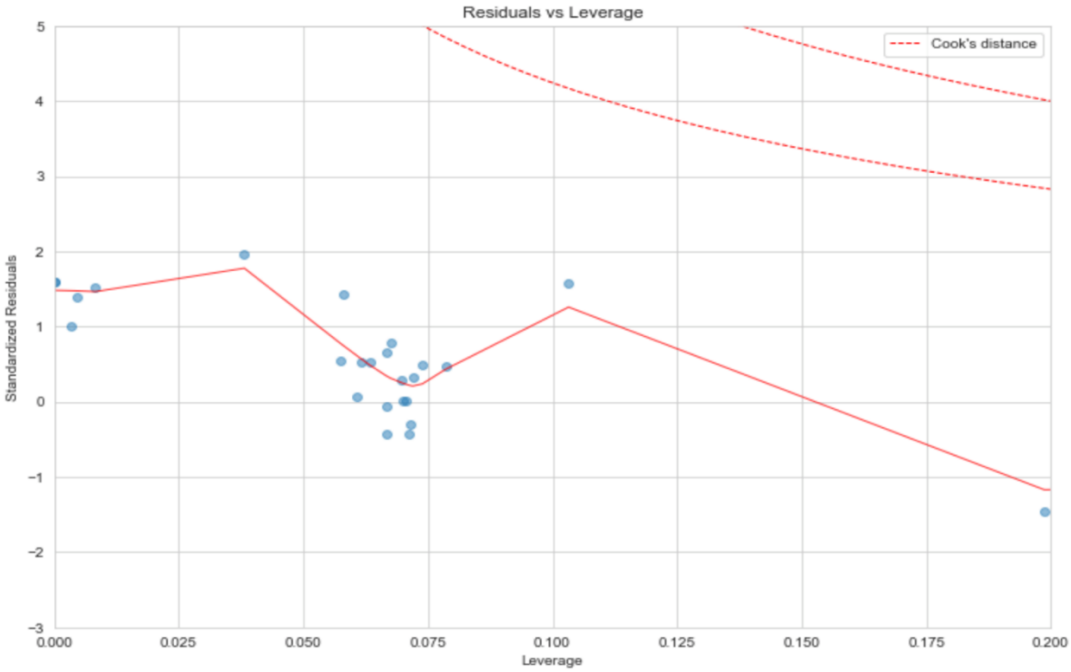

Figure 46. Residuals vs Leverage Plot for Model 20

The figure below shows the scale-location plot to illustrate if the residuals are spread equally along with the range of predictors. This plot helps to check the assumption of equal variance. It is good if there is a horizontal line.

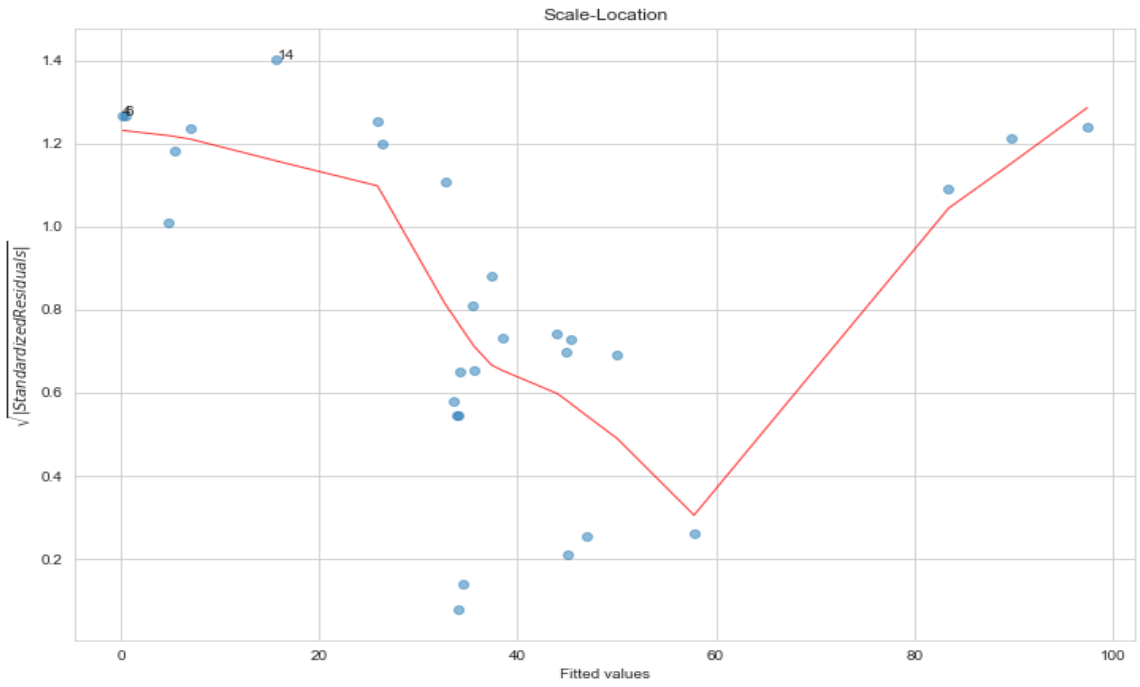

Figure 47. Scale- Location Plot for Model 20

CHAPTER 5

RESULTS

5.1 Introduction

In this chapter, the results of the regression models are analyzed, and the developed approaches to support the study are detailed. The following section examines and compares the results of the multiple regression models. Then the models that explain the correlation between dependent and independent variables are discussed. The chapter ends with discussing the predictive potential of the models.

5.2 Comparison of the Models

The primary aim of the study is to test the correlation of the independent variables with the dependent variables. The second goal is to develop a predictive model to estimate the cost of massive data breaches. In this study, there are twenty multiple regression models developed and tested to reach the goals. The comparison of the models is provided below grouped by the dataset and transformation.

"√" states that the variables are statistically significant. "X" shows that the variables are statistically not significant. Also, if the cell is blank, that means the variable is not included in the model.

Table 33. Comparison of the Models - Group 1

Model	R^2	Adjusted R^2	Predicted R^2	F stat	Mallow Cp	intercept	Revenue	PII	SPII	Class-action	
1	0.65	0.60	0	12.5	5	X	X	√	√	X	
2	0.65	0.62	0.13	17.3	3	X		√	√	X	
3	0.75	0.72	0.43	28.8	2			√	√	X	
4	0.84	0.82	0.56	37.2	3			√	X	X	Spii-class-action interaction (√)

Models after outliers are removed:

Table 34. Comparison of the Models – Group 2

Model	R^2	Adjusted R^2	Predicted R^2	F stat	Mallow Cp	intercept	Revenue	PII	SPII	Class-action
5	0.58	0.51	0.08	8.5	5	X	√	√	√	X
6	0.78	0.74	0.57	21.9	4		√	√	√	X
7	0.75	0.73	0.51	27.2	4.2		√	√	√	

Square Root Transformation of the Dependent Variable

Table 35. Comparison of the Models - Group 3

Model	R^2	Adjusted R^2	Predicted R^2	F stat	Mallow Cp	intercept	Revenue	PII	SPII	Class-action
8	0.67	0.62	0.36	13.5	5	√	X	√	√	X
9	0.67	0.63	0.40	18.5	3.3	√	X	√	√	
10	0.62	0.59	0.43	23.2	5.2	√		√	√	
11	0.85	0.82	0.77	38.2	4		√	√	√	√

Models after outliers are removed:

Table 36. Comparison of the Models - Group 4

Model	R^2	Adjusted R^2	Predicted R^2	F stat	Mallow Cp	intercept	Revenue	PII	SPII	Class-action
12	0.77	0.73	0.55	20	5	X	√	√	√	X
13	0.90	0.88	0.84	56.4	4		√	√	√	√

Box-Cox Transformation of the Dependent Variable

Table 37. Comparison of the Models - Group 5

Model	R^2	Adjusted R^2	Predicted R^2	F stat	Mallow Cp	intercept	Revenue	PII	SPII	Class-action
14	0.57	0.50	0.39	8.6	5	√	√	√	√	X
15	0.57	0.52	0.44	12	5.4	√	√	√	√	
16	0.79	0.76	0.74	28.8	4		X	√	√	√

Models after outliers are removed:

Table 38. Comparison of the Models - Group 6

Model	R^2	Adjusted R^2	Predicted R^2	F stat	Mallow Cp	intercept	Revenue	PII	SPII	Class-action
17	0.67	0.62	0.53	11.3	5	√	√	√	√	X
18	0.64	0.60	0.44	15	5.4	√	√	√	√	
19	0.84	0.81	0.79	32	4		X	√	√	√
20	0.82	0.80	0.78	39.6	4			√	√	√

The interaction effect is only seen in model 4. It means that the SPII data breaches may trigger class-action lawsuits that can considerably increase the data breach cost. PII and SPII variables are found statistically significant in all models. All independent variables except the intercept are found statistically significant in models 11 and 13 and have a positive correlation. Although model 16 and 19 have good values for adjusted and predicted R-squared and F statistics, the revenue variable is the only one that's p-value is slightly larger than 0.05.

5.3 Models with Correlation

This study only targets the data breaches, where the number of affected people is at least one million. The goal here is to determine the correlation of the independent variables with the dependent variables to identify the most relevant variables to forecast the massive data breaches. Since the condition of the model is that number of affected people must be one million; therefore, the X value at least is one million, even if the revenue variable is zero. Thus, in this study, the intercept is not necessary. After the backward elimination is applied, specific models include all variables except intercept to find out the correlation of the proposed variables.

All models state that PII and SPII are positively correlated with the data breach cost. In model 5,6,7,11,12,13,14,15,17, and 18, the revenue variable is found to be positively correlated

with the cost, too. However, class-action lawsuit variables only are found positively correlated with the cost in model 11,13,16,19, and 20.

The independent variables explain the variance in the data breach cost in models 4, 11, 13, 16, 19, and 20 better than earlier models (Jacobs, 2014; Romanosky, 2016). Adjusted and predicted R-squared proves that overfitting is seen in those models; also, F statistic, and t statistic values show the correlation between the dependent and independent variables.

5.4 Models with Predictive Potential

In this section, the models are compared in terms of predictability. The sample size is small to split the data as train and test to develop a predictive data breach cost model; therefore, the study does not claim developing a predictive model. However, model 11 and 13 may have predictive potential. The models are compared within their groups and summarized in the tables below.

Model 4 has better values; however, the difference between adjusted R-squared and predicted R-squared is substantial that may be a sign of overfitting. Also, the Residuals vs. Fitted values plot shows that the residuals do not spread out normally. Therefore, in group 1, there is not any model that has a predictive potential.

Table 39. Models with Predictive Potential Comparison – Group 1

Model	Adj. R^2	Pred. R^2	AIC	BIC	Mallow Cp	PRESS
1	0.60	0	1268	1275	5	2.24464E+18
2	0.62	0.14	1266	1272	3	1.93228E+18
3	0.72	0.43	1264	1269	2	1.79373E+18
4	0.82	0.56	1263	1266	3	1.38116E+18

Outlier Removed

After the outliers are removed, the models in group 2 are not showing improvement. The difference between predicted and adjusted squared is still considerable. Also, AIC and BIC values are not much changed. Therefore, in group 2, there is not any model that has a predictive potential.

Table 40. Models with Predictive Potential Comparison – Group 2

Model	Adj. R^2	Pred. R^2	AIC	BIC	Mallow Cp	PRESS
5	0.52	0.09	1150	1157	5	4.42522E+17
6	0.74	0.57	1148	1154	4	3.90180e+17
7	0.73	0.51	1149	1153	4.2	4.4e+17

Square Root Transformation of the dependent variable

After the square root transformation of the dependent variable, four models are developed. In this group, model 11 looks promising. Although there is not a significant change in BIC or AIC values PRESS value is low, and the difference between adjusted and predicted R-squared is small, which means there is not overfitting.

Table 41. Models with Predictive Potential Comparison – Group 3

Model	Adj. R^2	Pred. R^2	AIC	BIC	Mallow Cp	PRESS
8	0.63	0.37	621	629	5	1,323,695,117
9	0.64	0.40	620	625	3.3	1,249,528,573
10	0.60	0.43	622	626	5.2	1,177,745,822
11	0.83	0.78	625	630	4	1,206,968,500

Outliers removed

After removing the outliers, the model is improved on adjusted and predicted R-squared. Also, in model 13, AIC and BIC are slightly improved compared to model 12 besides Mallow Cp and PRESS value. Therefore, model 13 may have predictive potential.

Table 42. Models with Predictive Potential Comparison – Group 4

Model	Adj. R^2	Pred. R^2	AIC	BIC	Mallow Cp	PRESS
12	0.73	0.55	572	579	5	875,863,086
13	0.88	0.84	571	577	4	748,459,963

Box-Cox Transformation

The final transformation of the dependent variable is the Box-Cox transformation. Although model 16 has better values for adjusted and predicted R-squared; however, AIC, BIC, and PRESS values become worsen.

Table 43. Models with Predictive Potential Comparison – Group 5

Model	Adj. R^2	Pred. R^2	AIC	BIC	Mallow Cp	PRESS
14	0.51	0.39	282	289	5	16,481
15	0.52	0.44	280	285	5.4	15,173
16	0.76	0.74	332	338	4	79,389

Removing the Outlier

After removing the outlier, the models are improved, considering the adjusted and predicted R-squared. However, AIC, BIC, and PRESS values are considerably increased. As a result, there is not any model that has a predictive potential.

Table 44. Models with Predictive Potential Comparison – Group 6

Model	Adj. R^2	Pred. R^2	AIC	BIC	Mallow Cp	PRESS
17	0.62	0.52	208	215	5	12,576
18	0.64	0.53	207	212	4	12,378
19	0.81	0.79	262	267	4	57,625
20	0.80	0.78	263	267	4	61,305

It may be inferred that models 11 and 13 may have predictive potential regarding the adjusted and predicted R-squared, AIC, BIC, and PRESS values. The difference between adjusted and predicted R-squared tells about overfitting. Predicted R-squared is showing how well a model estimates response for new observations. It helps conclude when the model fits the original data; however, less successful in estimating for new observations. Especially, model 13 has a very high adjusted and predicted R-squared that can explain the variance in the response variable. The figure of residuals vs. fitted values, the probability of Omnibus values will be compared for the models 11 and 13 to satisfy the homoscedasticity assumption of the multiple regression. Ideal conditions for the normality of residuals and homoscedasticity are:

- The difference between adjusted and predicted R-squared is small
- Probability of Omnibus is close to 1
- Breusch-Pagan test should give p-value is larger than 0.05
- Residual vs. fitted should draw a horizontal line

Table 45. Predictive Model Comparison

Model	Adjusted R^2	Predicted R^2	Prob. omnibus	Breusch-Pagan test p-value > 0.05	Residual vs. Fitted line
11	0.82	0.77	0.68	yes	slightly
13	0.88	0.84	0.48	yes	slightly

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

This chapter presents the conclusions of this study, including a summary of the dissertation and its main contributions. Also, several suggestions for future work are discussed.

6.2 Summary of the Study

This study presents significant results that demonstrate the correlations between revenue, PII, SPII, and class-action lawsuits, and the dependent variable, which is the total cost of the data breach. Also, specific models developed in this study are able to predict the responses for new observations. Although the model fits the original data well; however, it is less qualified for providing valid predictions for new observations, and the limited number of observations hinders generalized conclusion. This study scrutinizes the type of information that is stolen from organizations in data breach incidents; it introduces a model that explains the relation between the stolen information and incurred costs due to a massive data breach. Furthermore, it elucidates the magnitude of a massive data breach cost in monetary terms.

Types of stolen information and costs incurred after a massive data breach are illustrated in the table below.

Table 46. Types of Costs and Stolen Information

Types of stolen PII	Types of stolen SPII	Types of Cost
• Name • Address • Email • Login information • Non-sensitive medical information	• Social security number • Debit/credit card numbers • Driver's license numbers • Tax ID • Passport numbers • Bank account numbers	• Remediation • Investigation • Increase in cybersecurity budget • Fines, fees • Data breach settlement

• Insurance membership number • Employment information • Date of birth • Driver's license state		• Professional services • Legal expenses • Credit/debit card re-issuance • ID theft protection • Canceled business deals • Service unavailability • Reduction in bidding

The cost of data breaches in this study change from $0.65 million to $1,445 million, with an average of $172 million. Among the possible causes for the small incurred cost is the lack of regulations and agencies, and they were mostly PII data breaches. Also, among the 31 companies, only 11 companies had cyber-insurance, and the monetary range of the policies is between $1 million and $125 million. The ratio of the insured amount to the total data breach cost is between 0.02 and 1.00. The developed regression models show that there is a positive linear correlation between dependent and independent variables. Model 13 looks promising due to little difference between adjusted and predicted $R^{2,}$ which implies that overfitting is not an issue.

6.3 Discussion of Contributions

This study introduces two new categories for personal information; these are PII and SPII. This new taxonomy accentuates the impact of sensitive information, which is more costly than not sensitive personal information. According to the models that are developed in this study, SPII can increase the cost of a data breach ten times more than the PII. Thus, data breaches that include sensitive information that may incur higher charges than non-sensitive data breaches. Organizations store sensitive information that must be more careful while managing their cybersecurity risk. They may need to invest more smartly in cybersecurity or purchase cyber insurance to reduce the financial impact of sensitive data breaches. Besides, there is an

interaction effect between SPII and class-action lawsuits. SPII data breaches may trigger more class-action lawsuits, which may beget more financial harms, poor reputation, or loss of sale.

This study focuses on the number of stolen records based on the affected people. It considers the type of stolen information and amount of it. The major contributions of the study to the earlier works (Jacobs, 2014; Romanosky, 2016) are listed as:

- Categorizing the information as PII and SPII
- Distinguishing the stolen type of information and its amount
- Including class-action lawsuits
- Trying Box-Cox transformation and square-root transformation
- Focusing on cases that the number of affected people is more than one million

Among 31 victim companies, only 11 companies had cyber-insurance ranging from $1 million to $125 million. The ratio of the insured amount to the total cost is between 0.02 and 1.00. The ratio becomes higher as the incidents become recent and cyber risk becomes more understandable. Therefore, cyber-insurance is undoubtedly helpful in reducing the financial impact of the data breach. While the cyber-insurance market is growing, the criticality of the cyber-insurance may depend on the data-owning company. This new insurance notion is not one size fit them all situation; for example, the more critical the nature of the data, the more indispensable the cyber-insurance need.

A U.S. court approved that web-scraping without permission is legal (Mehta, 2019). Besides, some of the recent verdicts of lawsuit cases indicate that a victim must have financial or other types of harm to get compensation from the data-owning companies due to a data breach (Hong, 2016). Therefore, this verdict indicates that the data breaches that involve PII or publicly

available information will result in less cost to involving companies. However, the companies that possess sensitive personal information or SPII must store, use, or transmit data by maintaining the necessary security protocols. As a result, cyber-insurance will be a means to mitigate the financial risk that is associated with data or information. This also means that companies are in this category, may be required to more cognizant while buying cyber-insurance. The models developed in this study and introduced new categorization of the information provide a more comprehensive understanding of the monetary impact of a data breach; for example, the potential impact of a data breach can be better estimated with these models that capitalize on the type and number of stored information. In addition, the insured amount can be compared with the potential data breach impact; as a result, a determination of under or over-insured can be made.

From the insurer perspective, the study may guide insurance firms while distinguishing between high and low-risk cyber-insurance customers. Companies that store PII have significantly less data breach costs compared to SPII data breaches because of the legality of web-scraping and unproven harm of PII data breaches. Furthermore, the developed models demonstrate that SPII loss increases the cost of a data breach up to ten times than PII loss. Therefore, companies own not sensitive personal information may be grouped under low-risk cyber-insurance customers.

On the other hand, companies that store sensitive information such as SSN, bank account, passport number may face much higher costs in case of an SPII data breach. As a result, they may face severe financial consequences due to class-action lawsuits, a settlement with governments, or technical costs. Therefore, cyber-insurance firms may use the categorization of

the information as PII and SPII to distinguish the clients as a low-risk customer and high-risk customer depending on the data the customers keep.

6.4 Future Research

Cyber-risk management has become more complicated, sophisticated, and multi-faceted in today's complex information ecosystem. Financial, customer relations, legal, and social aspects are becoming very important in addition to the technical aspect of cyber-risk management. Therefore, any cybersecurity failure is much more than a technical issue. This study investigates the financial impact of personal information data breaches by categorizing data as PII and SPII. The monetary impact of cyber-risk is a new field that needs to be examined. Few studies exist for data breach cost modeling. This study offers a foundation to address data breach cost forecasting. Areas for future research include the following:

- The monetary impact of availability and integrity compromise: This study only focuses on confidentiality breaches. However, integrity and availability attacks may cause a significant amount of loss. Identification of the factors to develop cost models for integrity and availability attacks is still an open area for further research.
- Likelihood of data breaches: This study addresses the impact part of data breach risk. Extending this study by addressing the uncertainty aspect, which is calculating the likelihood of data breaches, will give a more accurate data breach risk calculation.
- Models at different intervals: This study considers the cases where the number of affected people is more than one million. However, the number of affected people in the majority of the cases are less than one million. Therefore, another study would be useful to

develop a cost model for different intervals, such as a model where the number of affected people is between:

- 0-10,000
- 10,000 – 100,000
- 100,000 – 1 million

REFERENCES

Advisen. (2019). Cyber Loss Data. Retrieved July 31, 2019, from https://www.advisenltd.com/data/cyber-loss-data/

Arora, A., Hall, D., Pinto, C. A., Ramsey, D., & Telang, R. (2004). Measuring the risk-based value of IT security solutions. *IT Professional*. https://doi.org/10.1109/MITP.2004.89

Ashford, W. (2018). Economic impact of cyber crime is significant and rising. Retrieved April 5, 2018, from https://www.computerweekly.com/news/252435439/Economic-impact-of-cyber-crime-is-significant-and-rising

Aven, T., Ben-Haim, Y., Boje Andersen, H., Cox, T., Droguett, E., Greenberg, M., … Zio, E. (2015). *SRA glossary*. Retrieved from http://www.sra.org/sites/default/files/pdf/SRA-glossary-approved22june2015-x.pdf

Bernard, H. R. (2006). *Research Methods in Anthropology, 4th edition*. *Research Methods in Anthropology*. https://doi.org/10.1525/aa.2000.102.1.183

Biener, C., Eling, M., & Wirfs, J. H. (2015). Insurability of cyber risk: An empirical analysis. *Geneva Papers on Risk and Insurance: Issues and Practice*, *40*(1), 131–158. https://doi.org/10.1057/gpp.2014.19

Bolster, P., Pantalone, C. H., & Trahan, E. A. (2010). Security Breaches and Firm Value. *Journal of Business Valuation and Economic Loss Analysis*, *5*(1). https://doi.org/https://doi.org/10.2202/1932-9156.1081

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Butte.edu. (2013). Deductive, Inductive and Abductive Reasoning - TIP Sheet - Butte College.

Retrieved June 16, 2020, from http://www.butte.edu/departments/cas/tipsheets/thinking/reasoning.html

Calvert, S., & Kamp, J. (2018). More U.S. Cities Brace for 'Inevitable' Hackers. Retrieved July 25, 2019, from https://www.wsj.com/articles/more-cities-brace-for-inevitable-cyberattack-1536053401

Campbell, D., & Cook, T. (1979). *Quasi-experimentation: Design and analysis for field settings*. *Skokie, IL: Rand McNally*. Retrieved from http://scholar.google.de/scholar?q=Quasi-Experimentation%3A+Design+and+Analysis+Issues+for+Field+Settings&btnG=&hl=en&as_sdt=0%2C5#2

Campbell, D. T., & Stanley, J. C. (1967). *Experimental and quasi-experimental design for esearch*. *Handbook of Research on Teaching (1963)*. https://doi.org/10.1037/022808

Campbell, K., Gordon, L. A., Loeb, M. P., & Zhou, L. (2003). The economic cost of publicly announced information security breaches: Empirical evidence from the stock market. *Journal of Computer Security*, *11*(3), 431–448. https://doi.org/10.3233/JCS-2003-11308

Cavusoglu, H., Raghunathan, S., & Yue, W. T. (2008). Decision-Theoretic and Game-Theoretic Approaches to IT Security Investment. *Journal of Management Information Systems*, *25*(2), 281–304. https://doi.org/10.2753/MIS0742-1222250211

Cebula, J. J., & Young, L. R. (2010). A Taxonomy of Operational Cyber Security Risks. *Carnegie-Mellon Univ Pittsburgh Pa Software Engineering Inst*, (December), 1–47. https://doi.org/10.1007/978-1-4419-7133-3

Christensson, P. (2006). Cybercrime Definition. Retrieved November 1, 2018, from https://techterms.com/definition/cybercrime

Clifton, D. (2015). Cyber Security Return on Investment - Schneider Electric Blog. Retrieved

March 2, 2018, from https://blog.schneider-electric.com/cyber-security/2015/08/24/cyber-security-return-investment/

Creswell, J. (2009). *Research design: Qualitative, Quantitative, and Mixed Method Approaches*. *Sage* (3rd ed.). Sage. https://doi.org/10.2307/1523157

DHS. (2017). DHS Handbook Safeguarding Sensitive PII | Homeland Security. Retrieved May 27, 2019, from www.dhs.gov/privacy.

DoD. (2015). *DoD Program Manager ’s Guidebook for Integrating the Cybersecurity Risk Management Framework (RMF) into the System Acquisition Lifecycle*.

Douven, I. (2017). Abduction. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/sum2017/entries/abduction

DWT. (2018). Summary of U.S. State Data Breach Notification Statutes. Retrieved July 30, 2019, from https://www.dwt.com/gcp/state-data-breach-statutes

Edwards, B., Hofmeyr, S., & Forrest, S. (2016). Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity*, *2*(1), 3–14. https://doi.org/10.1093/cybsec/tyw003

Eling, M., & Loperfido, N. (2017). Data breaches: Goodness of fit, pricing, and risk measurement. *Insurance: Mathematics and Economics*, *75*, 126–136. https://doi.org/10.1016/j.insmatheco.2017.05.008

Eling, M., & Schnell, W. (2016). What do we know about cyber risk and cyber risk insurance? *Journal of Risk Finance*. https://doi.org/10.1108/JRF-09-2016-0122

Eling, M., & Wirfs, J. (2019). What are the actual costs of cyber risk events? *European Journal of Operational Research*, *272*(3), 1109–1119. https://doi.org/10.1016/j.ejor.2018.07.021

Ferran, L. (2016). Data Breaches Bigger, Worse Than You Think, Report Says - ABC News.

Retrieved August 12, 2019, from https://abcnews.go.com/International/data-breaches-bigger-worse-report/story?id=38340691

Fielder, A., Panaousis, E., Malacaria, P., Hankin, C., & Smeraldi, F. (2016). Decision support approaches for cyber security investment. *Decision Support Systems*, *86*, 13–23. https://doi.org/10.1016/j.dss.2016.02.012

Freund, J., & Jones, J. (2015). *Measuring and Managing Information Risk: A FAIR Approach*. *Measuring and Managing Information Risk*. Waltham: Butterworth-Heinemann.

FTC. (2016). *Protecting Personal Information: A Guide for Business*.

Garvey, P. R., Moynihan, R. A., & Servi, L. (2013). A macro method for measuring economic-benefit returns on cybersecurity investments: The table top approach. *Systems Engineering*, *16*(3), 313–328. https://doi.org/10.1002/sys.21236

Gatzlaff, K. M., & McCullough, K. A. (2010). The effect of data breaches on shareholder wealth. *Risk Management and Insurance Review*, *13*(1), 61–83. https://doi.org/10.1111/j.1540-6296.2010.01178.x

Gliner, J., Morgan, G. A., & Leech, N. (2017). *Research Methods in Applied Settings* (3rd ed.). New York: Routledge.

Goddard, W., & Melville, S. (2004). *Research methodology: An introduction*. Juta and Company Ltd.

Goel, S., & Shawky, H. A. (2009). Estimating the market impact of security breach announcements on firm values. *Information and Management*, *46*(7), 404–410. https://doi.org/10.1016/j.im.2009.06.005

Gordon, L. a., & Loeb, M. P. (2002). The economics of information security investment. *ACM Transactions on Information and System Security*, *5*(4), 438–457.

https://doi.org/10.1145/581271.581274

Gratt, L. B. (1987). Risk Analysis or Risk Assessment; A Proposal for Consistent Definitions. In V. T. Covello, L. B. Lave, A. Moghissi, & V. R. R. Uppuluri (Eds.), *Uncertainty in Risk Assessment, Risk Management, and Decision Making* (pp. 241–249). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4684-5317-1_20

Haimes, Y. Y., Kaplan, S., & Lambert, J. H. (2002). Risk filtering, ranking, and management framework using hierarchical holographic modeling. *Risk Analysis*, *22*(2), 383–397. https://doi.org/10.1111/0272-4332.00020

Hillestad, B. (2018). Risk Assessment: Qualitative vs Quantitative | SBS CyberSecurity. Retrieved July 26, 2019, from https://sbscyber.com/resources/risk-assessment-qualitative-vs-quantitative

Hinz, O., Nofer, M., Schiereck, D., & Trillig, J. (2015). The influence of data theft on the share prices and systematic risk of consumer electronics companies. *Information and Management*, *52*(3), 337–347. https://doi.org/10.1016/j.im.2014.12.006

Hong, N. (2016, June 26). For Consumers, Injury Is Hard to Prove in Data-Breach Cases. Retrieved July 2, 2020, from https://www.wsj.com/articles/for-consumers-injury-is-hard-to-prove-in-data-breach-cases-1466985988

IT-Online. (2016). How many cyber-heists go unreported? - IT-Online. Retrieved August 12, 2019, from https://it-online.co.za/2016/07/05/how-many-cyber-heists-go-unreported/

ITRC. (2019). Data Breaches. Retrieved from https://www.idtheftcenter.org/data-breaches/

Jacobs, J. (2014). Analyzing Ponemon Cost of Data Breach. Retrieved March 31, 2018, from http://datadrivensecurity.info/blog/posts/2014/Dec/ponemon/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical*

Learning with Applications in R. (G. Casella, S. Fienberg, & I. Olkin, Eds.), *Springer* (8th ed.). Springer. Retrieved from http://books.google.com/books?id=9tv0taI8l6YC

Johnson, A. (2018). Equifax breaks down just how bad last year's data breach was. Retrieved November 2, 2018, from https://www.nbcnews.com/news/us-news/equifax-breaks-down-just-how-bad-last-year-s-data-n872496

Jones, A. (2007). A framework for the management of information security risks. *BT Technology Journal*, *25*(1), 30–36. https://doi.org/10.1007/s10550-007-0005-9

Kaplan, S. (1997). The words of risk analysis. *Risk Analysis*, *17*(4), 407–417. https://doi.org/10.1111/j.1539-6924.1997.tb00881.x

Karabacak, B., & Sogukpinar, I. (2005). ISRAM: Information security risk analysis method. *Computers and Security*, *24*, 147–159. https://doi.org/10.1016/j.cose.2004.07.004

Kenton, W. (2019). Durbin Watson Statistic Definition. Retrieved October 21, 2019, from https://www.investopedia.com/terms/d/durbin-watson-statistic.asp#targetText=The Durbin Watson (DW) statistic,autocorrelation detected in the sample.

Kerlinger, F. (1986). Foundations of Behavioral Research 3rd Ed. *Foundations of Behavioral Research*.

Kissel, R. (2013). *Glossary of Key Information Security Terms Glossary of Key Information Security Terms*. *NIST* (Vol. NISTIR 729). https://doi.org/10.6028/NIST.IR.7298r2

Ko, M., & Dorantes, C. (2006). The impact of information security breaches on financial performance of the breached firms: An empirical investigation. *Journal of Information Technology Management*, *17*(2), 13–22. https://doi.org/10.4018/irmj.2012010102

Kopp, E., Kaffenberger, L., & Wilson, C. (2017). *Cyber Risk, Market Failures, and Financial Stability*. Retrieved from

http://www.imf.org/~/media/files/publications/wp/2017/wp17185.ashx

Krebs, B. (2014). The Target Breach, By the Numbers. Retrieved March 20, 2019, from https://krebsonsecurity.com/2014/05/the-target-breach-by-the-numbers/

Krebs, B. (2015). Anthem Breach May Have Started in April 2014. Retrieved September 20, 2019, from https://krebsonsecurity.com/2015/02/anthem-breach-may-have-started-in-april-2014/

Kuypers, M. (2017). *Risk in Cyber Systems*. Stanford University. Retrieved from https://searchworks.stanford.edu/view/11957204

Lam, W. M. W. (2015). Attack-prevention and damage-control investments in cybersecurity. In *WEIS*. https://doi.org/10.1016/j.infoecopol.2016.10.003

Landoll, D. (2011). *The Security Risk Assessment Handbook: A Complete Guide for Performing Security Risk Assessments, Second Edition* (2nd ed.). CRC Press, Inc.

Lawrence, D. (2014). KKR Adds Cyber-Risk Score to Its Assessment of Companies - Bloomberg. Retrieved December 10, 2018, from https://www.bloomberg.com/news/articles/2014-04-11/kkr-adds-cyber-risk-score-to-its-assessment-of-companies

Layton, R., & Watters, P. A. (2014). A methodology for estimating the tangible cost of data breaches. *Journal of Information Security and Applications*, *19*(6), 321–330. https://doi.org/10.1016/j.jisa.2014.10.012

Lee, A. S., & Baskerville, R. L. (2003). Generalizing Generalizability in Information Systems Research. *Information Systems Research*, *14*(3).

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. *Statistical Analysis with Missing Data* (2nd ed.). Wiley - Interscience.

Lodico, M. G., Spaulding, D. T., & Voegtle, K. H. (2010). *Methods in educational research: From theory to practice* (Vol. 28). John Wiley & Sons.

Matthews, D. (2018). *Cybersecurity Insurance and Identity Theft Coverage*. Retrieved from http://naic.org/documents/topic_insurance_industry_snapshots_2017_ye.pdf.

McCarty, K. (2018). Interpreting Results from Linear Regression – Is the data appropriate? Retrieved October 21, 2019, from https://www.accelebrate.com/blog/interpreting-results-from-linear-regression-is-the-data-appropriate

McNeese, B. (2016). Are the Skewness and Kurtosis Useful. Retrieved October 21, 2019, from https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics#targetText=It measures the amount of,(more in the tails).

Mehta, I. (2019). US court says scraping a site without permission isn't illegal. Retrieved November 26, 2019, from https://thenextweb.com/security/2019/09/10/us-court-says-scraping-a-site-without-permission-isnt-illegal/

Michaels, D. (2018). Yahoo's Successor to Pay $35 Million in Settlement Over Cyberbreach. Retrieved April 30, 2018, from https://www.wsj.com/articles/yahoos-successor-to-pay-35-million-in-settlement-over-cyber-breach-1524588040

Minitab. (2013). Multiple Regression Analysis: Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables. Retrieved March 28, 2020, from https://blog.minitab.com/blog/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables

Minitab. (2019). Residual plots for Fit Regression Model. Retrieved March 28, 2020, from https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/regression/how-to/fit-regression-model/interpret-the-results/all-statistics-and-

graphs/model-summary-table/

Modi, S. B., Wiles, M. A., & Mishra, S. (2015). Shareholder value implications of service failures in triads: The case of customer information security breaches. *Journal of Operations Management*, *35*, 21–39. https://doi.org/10.1016/j.jom.2014.10.003

N/A. (2018). Best Subsets Regression, Adjusted R-Sq, Mallows Cp. Retrieved August 12, 2019, from https://newonlinecourses.science.psu.edu/stat501/node/330/

NetDiligence. (2018). *Cyber Claims Study 2018*. Retrieved from https://netdiligence.com/wp-content/uploads/2018/11/2018-NetDiligence-Claims-Study_Version-1.0.pdf

NIST/SEMATECH. (2012). 1.3.5.11. Measures of Skewness and Kurtosis. In *e-Handbook of Statistical Methods* (pp. 3–6). Retrieved from https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

NIST. (2018). *Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1*. https://doi.org/10.6028/NIST.CSWP.04162018

OCR. (2003). *Summary of the HIPAA Privacy Rule*.

Owens, J. (2018). The Equifax data breach, in one chart. Retrieved November 8, 2018, from https://www.marketwatch.com/story/the-equifax-data-breach-in-one-chart-2018-09-07

Paganini, P. (2013). The Impact of Cybercrime. Retrieved October 12, 2017, from http://resources.infosecinstitute.com/2013-impact-cybercrime/#gref

Palmer, D. (2019). What is GDPR? Everything you need to know about the new general data protection regulations. Retrieved July 31, 2019, from https://www.zdnet.com/article/gdpr-an-executive-guide-to-what-you-need-to-know/

Pinto, A. (2018). Fundamentals of Cybersecurity. Retrieved June 18, 2019, from https://clark.center/details/capinto/Fundamentals of Cybersecurity

Pinto, A., & Garvey, P. (2012). *Advanced Risk Analysis in Engineering Enterprise Systems* (1st ed.). Boca Raton, FL: CRC Press. https://doi.org/10.1201/b13100

Pinto, A., & Magpili, L. (2015). *Operational Risk Management*. Momentum Press.

Pinto, C. A., McShane, M. K., & Bozkurt, I. (2012). System of systems perspective on risk: towards a unified concept. *International Journal of System of Systems Engineering*, *3*(1), 33. https://doi.org/10.1504/ijsse.2012.046558

Polit, D. F., & Beck, C. T. (2010). Generalization in quantitative and qualitative research: Myths and strategies. *International Journal of Nursing Studies*. https://doi.org/10.1016/j.ijnurstu.2010.06.004

Ponemon. (2019). *Cost of a Data Breach Report*. Retrieved from https://www.ibm.com/security/data-breach

Poyraz, O., Serttas, O., Keskin, O., Tatar, U., & Pinto, A. (2018). Impact of Cyber-attacks on Valuation of Public Companies. In *Ninth International Conference on Complex Systems*. Cambridge, MA: New England Complex Systems Institute.

PRC. (2019). Privacy Rights Clearinghouse | Data Breaches. Retrieved July 31, 2019, from https://www.privacyrights.org/data-breaches?title=&breach_type%255B%255D=285&breach_type%255B%255D=268&breach_type%255B%255D=267&breach_type%255B%255D=264&breach_type%255B%255D=265&breach_type%255B%255D=266&breach_type%255B%255D=269&breach_type%255B%255D=270&org_type%255B%255D=260&org_type%25

Rabai, L. B. A., Jouini, M., Aissa, A. Ben, & Mili, A. (2013). A cybersecurity model in cloud computing environments. *Journal of King Saud University - Computer and Information Sciences*, *25*, 63–75. https://doi.org/10.1016/j.jksuci.2012.06.002

Refsdal, A., Solhaug, B., & Stølen, K. (2015). Cyber-risk Management. In *Cyber-Risk Management* (pp. 33–47). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-23570-7_5

Romanosky, S. (2016). Examining the costs and causes of cyber incidents. *Journal of Cybersecurity*, *2*(2), 121–135. https://doi.org/10.1093/cybsec/tyw001

Ruan, K. (2017). Introducing cybernomics: A unifying economic framework for measuring cyber risk. *Computers and Security*, *65*(2017), 77–89. https://doi.org/10.1016/j.cose.2016.10.009

Sanna, N. (2016). What Is a Cyber Value-at-Risk Model? Retrieved February 25, 2018, from http://www.fairinstitute.org/blog/what-is-a-cyber-value-at-risk-model

Saunders, M., Lewis, P., & Thornhill, A. (2012). *Research methods for business students. Fitfth Edition*. *Pearson* (6th ed.). Pearson.

Schatz, D., & Bashroush, R. (2016). The impact of repeated data breach events on organisations' market value. *Information and Computer Security*, *24*(1), 73–92. https://doi.org/10.1108/ICS-03-2014-0020

Schneier, B. (2008). Security ROI - Schneier on Security. Retrieved March 2, 2018, from https://www.schneier.com/blog/archives/2008/09/security_roi_1.html

SEC. (2018). *Commission Statement and Guidance on Public Company Cybersecurity Disclosures*.

Shameli-sendi, A., Ezzati-jivan, N., Jabbarifar, M., & Dagenais, M. (2012). Intrusion Response Systems : Survey and Taxonomy. *IJCSNS International Journal of Computer Science and Network Security*, *12*(1), 1–14.

Statistics Solutions. (n.d.). Assumptions of Multiple Linear Regression. Retrieved July 23, 2019, from https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/

STIP. (2018). Personally Identifiable Information. Retrieved May 28, 2019, from https://www.osti.gov/stip/pii

Su, X., Bolzoni, D., & Van Eck, P. (2006). A business goal driven approach for understanding and specifying information security requirements. In *CEUR Workshop Proceedings* (Vol. 364, pp. 103–110). Retrieved from http://arxiv.org/abs/cs/0603129

Tavory, I., & Timmermans, S. (2014). *Abductive analysis: Theorizing qualitative research*. University of Chicago Press.

Wall, R., & Olson, P. (2019). British Airways Faces $230 Million Fine Over Data Breach as European Privacy Rules Start to Bite. Retrieved July 31, 2019, from https://www.wsj.com/articles/u-k-privacy-regulator-show-its-teeth-fines-british-airways-parent-230-million-for-data-breach-11562573218

Wavefront. (n.d.). A Brief History of Cybercrime. Retrieved March 22, 2019, from https://www.wavefrontcg.com/A_Brief_History_of_Cybercrime.html

WDPI. (n.d.). *DPI Personally Identifiable Information*. Retrieved from https://dpi.wi.gov/sites/default/files/imce/wisedash/pdf/PII list of Examples.pdf

Westfall, P. H. (2014). Kurtosis as Peakedness. *American Statisticians*, *68*(3), 191–195. https://doi.org/10.1080/00031305.2014.917055.Kurtosis

Wheatley, S., Hofmann, A., & Sornette, D. (2019). Data breaches in the catastrophe framework & beyond. *ArXiv*. Retrieved from http://arxiv.org/abs/1901.00699

Wheeler, D. J. (2011). Problems with Skewness and Kurtosis, Part One. *Quality Digest Daily*, (August 1-2). Retrieved from www.spcpress.com/pdf/DJW231.pdf1

World Economic Forum. (2016). *The Global Risks Report 2016 | World Economic Forum*. Retrieved from https://www.weforum.org/reports/the-global-risks-report-2016

Yayla, A. A., & Hu, Q. (2011). The impact of information security events on the stock value of firms: The effect of contingency factors. *Journal of Information Technology*, *26*(1), 60–77. https://doi.org/10.1057/jit.2010.4

Zdaniuk, B. (2014). Ordinary Least-Squares (OLS) Model. In A. C. Michalos (Ed.), *Encyclopedia of Quality of Life and Well-Being Research* (pp. 4515–4517). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_2008

Zhang, X., Wuwong, N., Li, H., & Zhang, X. (2010). Information Security Risk Management Framework for the Cloud Computing Environments (pp. 1328–1334). https://doi.org/10.1109/CIT.2010.501

VITA

Omer Ilker Poyraz

Department of Engineering Management and Systems Engineering
Old Dominion University, Norfolk, VA
Email: iegri001@odu.edu

EDUCATION

Old Dominion University, Norfolk, VA — 2020
Ph.D. Engineering Management and Systems Engineering

Old Dominion University, Norfolk, VA — 2015
M.B.A.

Yeditepe University, Istanbul, Turkey — 2012
B.A. Business and Administration

RESEARCH AREAS

Quantification of Cyber-risk, Machine Learning, Cyberwarfare, Complex Systems

PROFESSIONAL EXPERIENCE

Graduate Research Assistant | January 2017 – February 2020

Turkish Military Academy | August 2012 – January 2017