5-2022

An Examination of the Statistics and Risk Management Concepts Behind the Patient Protection and Affordable Care Act (PPACA) of 2010

Scott Sinclair

An Examination of the Statistics and Risk Management Concepts Behind the Patient Protection and Affordable Care Act (PPACA) of 2010

Scott Sinclair

April 4, 2022

An Undergraduate Thesis

Presented to The Honors Program

of Butler University

Supervised by Mohammad Shaha A. Patwary, Ph.D.

In Partial Fulfillment

of the Requirements for Graduation Honors

ABSTRACT

An Examination of the Statistics and Risk Management Concepts Behind the Patient Protection and Affordable Care Act (PPACA) of 2010

Scott Sinclair

April 4, 2022

The Patient Protection and Affordable Care Act (PPACA) is the overarching federal law that has impacted the intricacies of the health insurance market for more than a decade. Using the supervised learning method of multiple linear regression, the relationship between the medical loss ratio rebates and predictor variables such as the state, health insurance market, and the number of insurance companies owing rebates will be analyzed, along with the actuarial value of metal tiers and geographic rating area factors in terms of their relationship to the insurance premium for a standard family of four, defined as a forty-year-old couple with two children. Moreover, cluster analysis will be used to analyze any and all phenomena discovered by looking at which data points are assigned to specific clusters based on shared attributes. All datasets are from the years 2014 through 2020, as 2014 was when the PPACA went into effect and 2020 is the latest year for which data is available.

Contents

Chapter 1

Introduction

Following the completion of my eight-month long internship at Anthem in 2021, I recognized the need to know more about the political and economic forces that make a clear and consistent impact on the nature of the health insurance market, as new state and federal laws are perpetually reshaping the health insurance market and industry. In short, this thesis is intended to grow out of my previous internship experience as well as my previous academic work in terms of specific terminology and data analysis methods in statistics. Rating area factors, medical loss ratios (MLRs) and benefit tiers were all major topics that were incorporated into multiple projects during my time at Anthem, all of which were/are significantly impacted by the Affordable Care Act's passage and implementation. While there is not sufficient time nor would my thesis be genuinely focused if the entire scope of the Affordable Care Act was covered in this thesis, I can and will focus on how the three specific provisions listed above that directly impacted the work I completed at Anthem impacts other key provisions in the act that are more well-known across the country.

1.1 Research Questions

There are two research questions that I intend to answer in undertaking this thesis. First, "How have the following three provisions of the Affordable Care Act (ACA) impacted the health insurance market from a statistics and risk management standpoint: the introduction of rating

area factors for adjustment of premiums, medical loss ratio (MLR) rebates, and the creation of benefit tiers/categories (i.e. bronze, silver, gold, platinum)?" Second, "How do these specific provisions of the Affordable Care Act impact the more well-known provisions of the legislation from a statistics and risk management perspective?" That is, in asking the second question, I not only seek to relate the three specific provisions listed in the first question to the key parts of the act (e.g. lowering of costs for those with pre-existing conditions), but seek to explain why the provisions in the first question are important both within themselves and in the larger scope of the act.

1.2 Background on the Affordable Care Act (ACA)

What is known by both scholars and myself about the Affordable Care Act (ACA) is that it intends to serve three important purposes in improving the health insurance market and industry: to expand health insurance coverage, lower health care costs, and improve the delivery of quality health care. Specifically, among other provisions, the legislation sought to expand coverage for health insurance for people with pre-existing conditions as well as individuals in other high-risk groups through the creation of temporary high-risk pools, control the cost of health care through various regulations and subsidies, and improve the delivery of the health care system by eliminating fraud, waste, and abuse by the maximum extent possible. The perceived success in meeting these objectives depends in-part on one's political views, though it is widely confirmed that the act expanded health insurance coverage to thirty million individuals with pre-existing conditions who previously did not have access to it. Moreover, portions of the original legislation have been modified or eliminated, such as the zeroing out of the individual mandate in 2017. As such, while the three specific provisions that will be discussed in this thesis have

been largely unaffected by these changes, these changes will be accounted for in my discussion about how the provisions have affected the remainder of the legislation and the health insurance market as a whole.

1.3 Concise Definitions of Key Terms

A medical loss ratio is the ratio of insurance claims to insurance premiums, and insurance companies are mandated to pay medical loss ratio rebates by the Affordable Care Act (ACA) when the medical loss ratio exceeds the government-defined threshold. In simple terms, the geographic rating area factor is a risk adjustment factor applied in the health insurance pricing process to account for differences in morbidity in different geographical regions within a specific state. The term “actuarial value” indicates the percentage of benefits that are covered by a health insurance plan of a specific metal tier in relation to the total cost of the plan.

Chapter 2

Data and Methodology

2.1 Data

2.1.1 Data for Provision One

Number of observations: 153 (fifty-one states, with the District of Columbia being treated as a state, with three insurance markets for each state)

Response Variable: Average annual medical loss ratio rebate

Structure of Response Variable: integer data type

Predictor Variables: State, market, and number of company rebates

Structure of Predictor Variables: The 'state' variable is categorical with fifty-one levels (i.e. categories), including the base category. The 'market' variable is a factor variable (initially a character variable) with three levels (i.e. categories), including the base category. The 'company rebates' variable is numerical, indicating that this variable can take on a whole number or a decimal number.

Model: Average Rebate ~ Intercept + State + Market + Company Rebates

$$Y = X\beta + D\delta + \varepsilon$$

where Y is the response variable, X is the numeric predictors' design matrix,
D is the categorical predictors' design matrix, and epsilon is the errors matrix

2.1.2 Data for Provisions Two and Three

Number of observations: 78,379

Response Variable: Average annual premium for a forty-year-old couple with two children

Structure of Response Variable: floating point (decimal) data type

Predictor Variables: geo_feddefault_v (default geographic rating area), geo_numrtarea_n (number of territories), geototal_div_n (number of rating areas), geo_divide_v (how rating areas are divided), geo_reins_v (whether rating area is following standard reinsurance practices), geo_riskadjinf_v (risk adjustment factor), geo_mlr_v (medical loss ratio), metal_level, plan_type (PPO, HMO, EPO), Rating.Area.Factor, couple2_children_age_40 (premium), medical_maximum_out_of_pocket__individual__standard, drug_maximum_out_of_pocket__individual__standard

Structure of Predictor Variables: The latter seven variables except for "plan type" are numerical variables, whereas all others are factor (i.e. categorical) variables

Model: couple2_children_age_40 (premium) ~ geo_feddefault_v (default geographic rating area) + geo_numrtarea_n (number of territories) + geototal_div_n (number of rating areas) + geo_divide_v (how rating areas are divided) + geo_reins_v (whether rating area is following standard reinsurance practices) + geo_riskadjinf_v (risk adjustment factor) + geo_mlr_v (medical loss ratio) + metal_level + plan_type (PPO + HMO + EPO) + Rating.Area.Factor + medical_maximum_out_of_pocket__individual__standard + drug_maximum_out_of_pocket__individual__standard

$$Y = X\beta + D\delta + \varepsilon$$

where Y is the response variable, X is the numeric predictors' design matrix,D is the categorical predictors' design matrix, and epsilon is the errors matrix

2.2 Preprocessing

Data preprocessing, in which the data is pulled from a source, cleaned, and then formatted to be user-friendly and convenient, is one of the most challenging steps in data-driven analysis due to the laborious and typically tedious nature of the process. This was certainly the case with preprocessing the data for this provision, and it was not as ideal and convenient as I laid out in my thesis proposal or as I imagined. That is, I expected to be able to pull the structured data from the webpage from which it was published using what I know about dataset scrapping in RStudio, the statistical-friendly software program used to pull, run, and analyze the data. However, after recognizing that while the data itself was formatted as a table, the webpage from which the data was pulled did not recognize the data as being located inside a table, but rather as a collection of individual data points that so happened to be residing inside the said table. As a result, I had to manually copy and paste snippets of the structured data to a Microsoft Excel file. Since there were seven years of data being analyzed, this was a seven-fold iterative process. Moreover, because Microsoft Excel did not recognize each observation (represented by each line containing a total of four values since there were a total of three predictor variables along with the response variable) as containing the value for each of the predictor variables and the response variable, the dataset read as if each line contained only one value, with spaces separating the embedded values within them. I rectified this issue by utilizing my Excel skills to clean the data so that each cell in the software program contained only one value for each variable in each observation. Following this, I then formatted each of the values to its corresponding data type, such that a quantitative variable that was intended to be structured as a numerical variable is in fact formatted as numerical and a categorical variable that was intended to be structured as a factor is in fact a factor and not a string of characters.

2.3 Methodology Description

After the data went through its preprocessing phase, I checked and analyzed the validity of the three hypotheses described in the following chapter. This was accomplished using multiple linear regression with maximum likelihood estimation (MLE). In this technique, the response variable, the average annual medical loss ratio rebate, was plotted against the three predictor variables in the model, such that the assumed relationship between the response variable and each of these predictor variables was a linear (straight line) one. This type of regression model is used most frequently in statistical analysis not only because of its feasibility and high interpretability, but also because it accurately explains the linear relationship between the variables without suffering from "overfitting" and displays summary statistics such as the residual standard error and adjusted coefficient of determination that can be accurately explained to audiences of any level of statistical background.

Lastly, I used the unsupervised statistical learning method of cluster analysis in provision one to make sense of unique patterns, trends, and phenomena in the data. Although every attempt was made to complete cluster analysis for provision two, the algorithm crashed in RStudio, given that there were more than 78000 data points, each containing fourteen variables. The term "unsupervised" indicates that there was not a dependent variable (a.k.a. response variable) measured against independent variables (a.k.a. predictor/explanatory variables) for the purpose of finding a relationship between the variables. Rather, an unsupervised learning algorithm simply looks for meaningful patterns in the data without the need for human intervention in the form of assigning dependent and independent variables. As such, cluster analysis, as the name implies, assigns each data point in the dataset to a specific cluster, with all data points in a specific cluster having unique features in common with one another, whereas a data point in a

different cluster than the data point being analyzed suggests that it possesses a statistical feature that distinguishes it from the said data point in a statistically significant way. The number of clusters is determined using a complex machine learning algorithm called agglomerative hierarchical clustering, whereby the numerical distance between two or more observations is examined and the diagram generated is that of a tree with a hierarchy, such that the number of clusters is equivalent to the number of branches at the location on the tree where the intercluster dissimilarity is the largest.

2.4 Imputation for Dataset with Provisions Two and Three

In addition to pulling, cleaning, and formatting data so that it can be efficiently and accurately be processed and analyzed, data preprocessing also involves having a process for handling missing data points from a dataset. Such a process is referred to as *imputation,* as we are imputing values into a dataset based on a defined procedure or formula based on information that we already know about the data points that exist. Although missing observations are not an occurrence with every dataset, as demonstrated by the fact that there were no missing data points in the prior provision, it is not uncommon in larger datasets, whereby certain observations could be missing for a number of reasons, including simple machine errors, inadequate technology, and data collection errors. Since the size of the dataset used for the statistical analysis for these two combined provisions is greater than 78000, this phenomenon was not necessarily astonishing. I considered two methods in approaching this issue, the first of which was the most straightforward, yet was technologically unfeasible, and the second of which was less convenient from a precision and statistical interpretation point of view, but that was nevertheless reasonable in its imputation method as well as technologically feasible.

The first method of data imputation that I considered was multiple imputation, which is most well-suited for a dataset in which multiple variables have missing values. There are a variety of statistical packages in the statistical software program RStudio that assist with this process, and each of these packages do so in different ways. For instance, the "MICE" package, which was one of the options that I considered for addressing this issue, does so by creating a distribution for each of the missing data points by fitting a regression model based on the other variables (Alice 4). That is, if a specific observation is missing a value for one or more variables, the corresponding function in the "MICE" package will analyze the values for the remaining variables for that particular observation and then based on the fitted model, will assign a value to the missing data point(s). This package also assumes that the data points that are missing are "Missing Completely at Random", meaning that the observations that are missing are not missing on a sequential basis or intentionally. However, because only one variable, namely the response variable, had missing values, a multiple imputation method was not appropriate for imputing missing values in the dataset. Therefore, the second option that I considered was a single imputation method, which means assigning each missing value in a single variable a statistic, which could be the arithmetic mean, geometric mean, median, or any similar numerical representation of the distribution. While not the most precise method of imputation, it is the most feasible as well as the most appropriate for small amounts of missing data. The statistic I used in this case was the median, since the distribution of the response variable was not normally distributed and instead was right skewed, the median is the best representation of the approximate "center" of the distribution.

Chapter 3

Hypotheses

3.1 Hypotheses for Provision One

Statement of Hypothesis One: *There is no significant linear relationship between the average annual medical loss ratio (MLR) rebate and the state or district being examined. Put another way, the average annual medical loss ratio (MLR) rebate for each state and district across all fifty-one states and districts does not change from state to state/district; equivalently, the 'State' variable is not a statistically significant variable. This hypothesis holds true in both the aggregate case and the individual annual year case (non-aggregate case), where the aggregate case refers to the data being analyzed over a seven-year period from calendar year 2014 to calendar year 2020. The non-aggregate case refers to the data being analyzed over each individual year in the aforementioned range, for a total of seven years of analysis results.*

Statement of Hypothesis Two: *There is no significant linear relationship between the average annual medical loss ratio (MLR) rebate and the insurance market category (individual, small group, and large group) being examined from one insurance market to another, with the 'individual' market constituting the base case. Put another way, the average annual medical loss ratio (MLR) rebate for each state and district across all fifty-one states and districts does not change base category. Equivalently, the 'Market' variable is not statistically significant. This*

hypothesis holds true in both the aggregate case and the individual annual year case (non-aggregate case).

Statement of Hypothesis Three: *There is no significant linear relationship between the average annual medical loss ratio (MLR) rebate and the number of insurance companies who owe rebates in a particular state or district. Put another way, the average annual medical loss ratio (MLR) rebate for each state and district across all fifty-one states and districts does not change based on company rebate frequency. Equivalently, the 'Company.Rebates' variable is not statistically significant. This hypothesis holds true in both the aggregate case and the individual annual year case (non-aggregate case).*

3.2 Hypothesis for Provision Two

Statement of Hypothesis One: *There is no significant linear relationship between the average premium for a forty-year-old couple with two children and the corresponding geographic rating area factor.*

3.3 Hypothesis for Provision Three

Statement of Hypothesis Two: *There is no significant linear relationship between the average premium for a forty-year-old couple with two children and the corresponding metal tier (Bronze, Silver, Gold, Platinum, Catastrophic).*

Chapter 4

Statistical Analysis Results

4.1 Statistical Analysis Results for Provision One

The initial model, with the parameter estimates included, was as follows, with "Alaska" being the base category for the factor variable "state" and the "individual" market being the base category for the factor variable "market":

$$Y = X\beta + D\delta + \varepsilon$$

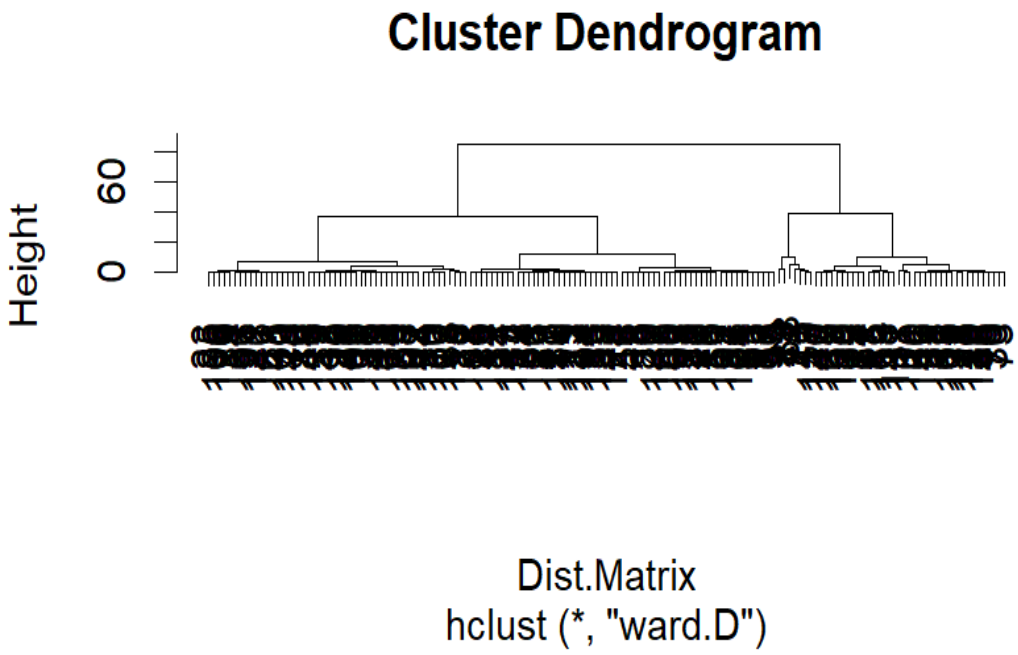
where Y is the response variable, X is the numeric predictors' design matrix,
D is the categorical predictors' design matrix, and epsilon is the errors matrix.

Average Rebate = −32201 – 132189*StateAL – 207404*StateAR – 405066*StateAZ + 696404*StateCA + ... – 161662*StateWY – 11443*Market.L – 366431*Market.Q + 173104*CompanyRebates

where the −32201 is the intercept, the two digit code in front of "State" indicates the state itself, the ".L" in front of "Market" indicates the large group market, and the ".Q" in front of "Market" indicates the small group market. The value −32201 indicates that when the state being analyzed is Alaska (the base "State" category), the market being analyzed is the individual market (the

base “Market” category), and there are zero companies owing rebates in the corresponding state and market, the average rebate is $0.00, because a negative value for a monetary amount is not possible. That is, all of the insurance companies in the state and market being analyzed cannot mathematically owe $−32201.00, because that would imply that the companies somehow managed to recoup $32201, which is not possible in an insurance context because the rebates are paid by the insurance company to policyholders for an insufficient medical loss ratio, thus indicating a one-way transaction. On the other hand, if, for instance, all the variable values assumed above were held to be true except the number of company rebates for the state and market being analyzed was one instead of zero, the average rebate for the state and market at hand would be $140903.00 (173104 – 32201). Each of the parameter estimates indicate that for each one-unit increase in the value for the corresponding factor or quantitative variable, the average rebate increases or decreases by the value of the parameter estimate, depending on whether the value of the parameter estimate is positive or negative. If the value of the parameter estimate is a positive number the average rebate increases and if the value of the parameter estimate is a negative number, the average rebate decreases.

Following analysis of the dataset using multiple linear regression with maximum likelihood estimation, I utilized the unsupervised learning method of cluster analysis to recognize patterns in the data that would not be recognizable based on the supervised method of linear regression alone. Specifically, the hierarchical clustering algorithm, which compares the intercluster dissimilarity among the different observations as described earlier, generated a dendrogram, showing that the optimal number of clusters is two. This was determined by drawing a horizontal line across the diagram at the point where the height between branches of the “tree” below was at a maximum:



This means that the algorithm determined that the 153 data points in the dataset can be classified into two clusters, or categories of observations. However, identifying the characteristics that each of the two clusters possesses within themselves that make the observations in that cluster unique from those in the other cluster requires not only determining which specific data points are located within each cluster, but also analyzing those data points for shared attributes. After completing this analysis, it was determined that all of the observations in cluster one shared the feature of having company rebates of at most 2.57. In other words, on average, the number of insurance companies in a particular state owing medical loss ratio rebates based on the preestablished threshold set by the Affordable Care Act is no larger than 2.57 across seven years, 2014 to 2020. As such, all observations with company rebates above 2.57 were grouped into cluster two.

Lastly, an ANOVA partial F-test was completed to check whether the “State” variable should be removed from the model or not, given that only two states in the model had an

individual t-test p-value of less than five percent. In the presence of two variables, market and company.rebates, the state variable was found not to be significant by the partial F-test. As a result, the “State” variable was removed from the final model.

4.2 Statistical Analysis Results for Provisions Two and Three

The initial model, with the parameter estimates included, was as follows:

Average Premium = 296.5 – 60.42*geo_feddefault_v (if state defaults to federal geographic rating area) + 0.4466*geo_numrtarea_n (number of geographic rating areas in the state) – 1.538*geototal_div_n (total number of geographical divisions in the state) + 74.38*geo_divide_v (how the state is dividing its geographic rating areas) + 11.86*geo_reinsfac_v (if geographic rating area factor is considered in setting reinsurance premium) - 187.6*geo_reins_v (if the state is managing its own reinsurance) + 65.89*geo_riskadjinf_v (if geographic rating areas influence a state’s risk adjustment) – 89.53*geo_mlr_v (if the state is managing its MLRs itself) – 143.6*metal_level (Bronze, Silver, Gold, Platinum, Catastrophic) + 238.7*plan_type (PPO, HMO, EPO, POS) + 325.0*Rating.Area.Factor – 0.0539*medical_maximum_out_of_pocket__individual__standard – 0.01094*drug_maximum_out_of_pocket__individual__standard

In this dataset, there were a total of thirteen predictor variables regressed against the response variable, which is the average premium for a 40-year-old couple with two children. Of the thirteen explanatory variables, five were quantitative: geo_numrtarea_n, geototal_div_n,

Rating.Area.Factor, medical_maximum_out_of_pocket__individual__standard, and drug_maximum_out_of_pocket__individual__standard. The first and last of these predictors were not statistically significant in the model given above, based on the p-value given in the individual t-tests that were outputted in RStudio when running the multiple regression model. However, the remaining three quantitative variables were statistically significant, including the rating area factor variable. As such, the conclusion that can be drawn from hypothesis two is that rating area factor does have a significant impact on the health insurance premium for a standard family of four.

With regards to the eight categorical/factor variables in the model, seven of them were statistically significant in at least one of the factors. That is, among the factor variables, at least one of the categories had a statistically significant p-value (less than 0.05) when changing from the base category to the category at-hand, with several such factor variables having statistically significant p-values in all of its associated categories. For example, for the metal tier variable, all four non-base categories (silver, gold, platinum, catastrophic) produce statistically significant results in comparison to the base category (bronze). Thus, the conclusion that can be drawn from the third hypothesis is that the actuarial value of the different metal tiers have a significant impact on the health insurance premium for a standard family of four. On the other hand, the "plan_type" variable, for instance, has only one category (HMO) outside the base category (EPO) with a statistically significant result, meaning that as the plan type changes from an EPO to either a PPO or POS, the premium for a standard family of four is not significantly altered. However, the premium is significantly impacted by a change in the health insurance plan from an EPO to an HMO.

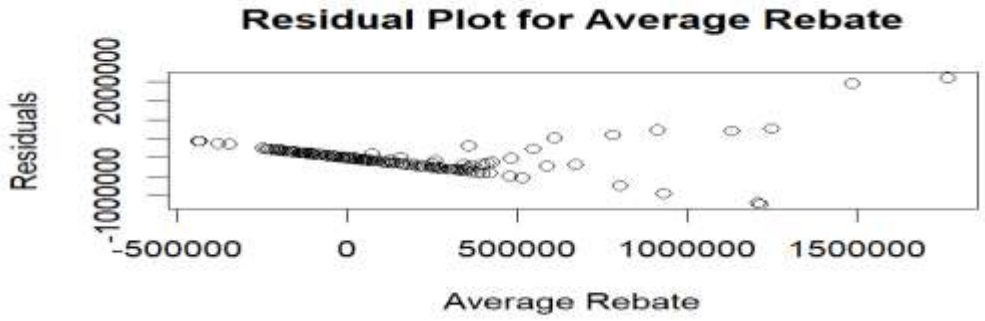
Chapter 5

Model Diagnostics

5.1 Model Diagnostics for Provision One

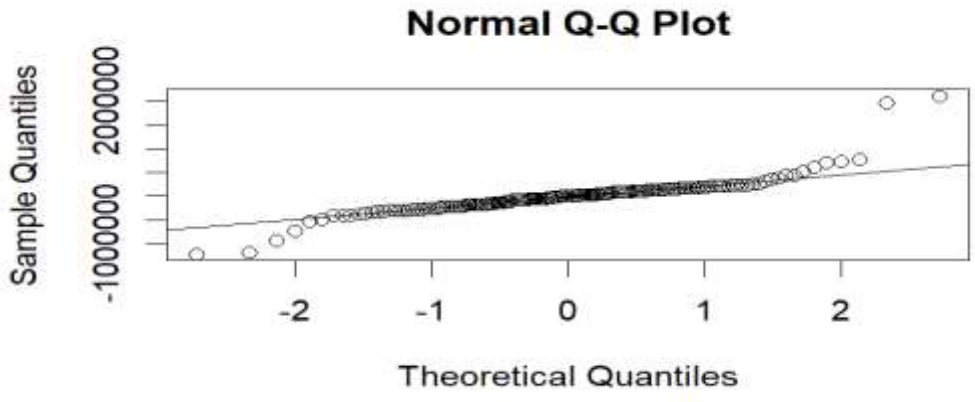
After fitting the model described above, I proceeded to check the assumptions for the fitted multiple linear regression model. There are four assumptions that are made when employing multiple linear regression, which all revolve around the validity of assuming a straight-line (linear) relationship between the response variable and each of the explanatory variables. Specifically, here were the assumptions that were made when running the model:

1. The mean of the residuals (the differences between the fitted values of the response variable as calculated by the regression model and the corresponding observed values of the response variable found in the dataset) is approximately centered at zero. That is, the mean of the residuals is zero such that no error term is included in the model. This was checked by plotting the residuals themselves against the individual values of the response variable, Average Rebate, generating a residual plot. This assumption was met.



2. The variance (spread) of the residuals is constant over all fitted values of the response variable. In other words, it is not the case where the value of the residuals increase as the value of the response variable increases. The term for this assumption, in which the variance of the residuals is constant over all fitted values of the response variable, is homoscedasticity. I checked this assumption using the residual plot explained in the preceding assumption as well as the Breusch-Pagan Test. The Breusch-Pagan test is a quantitative hypothesis test in statistics used specifically for checking this assumption, where the level of significance is set in advance. Using a significance level of five percent ($\alpha = 0.05$), the result of the hypothesis test was that there was insignificant evidence to indicate that there was an absence of homoscedasticity in the model, with the corresponding p-value equating to approximately 0.61.

3. The residuals are normally distributed with a mean of zero and a constant variance over all values of the fitted response variable. That is, the distribution of the residuals follow a normal distribution. I verified this assumption using a normal probability plot in which a linear pattern of data points indicates a normally distributed set of residuals, where the axes of the plot are the sample quantiles and the theoretical quantiles.

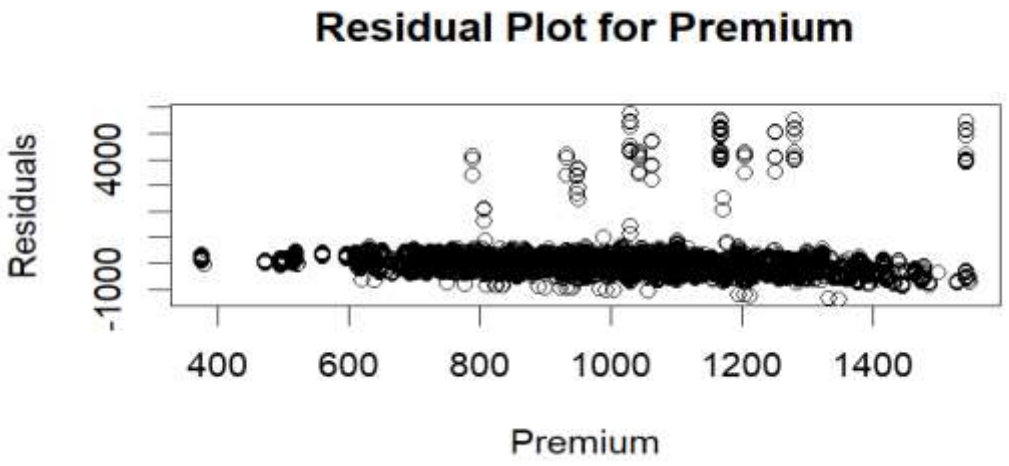


4. There is no multicollinearity in the dataset. Multicollinearity means that two or more of the predictor variables have a correlation coefficient of greater than ninety percent, indicating a significant correlation. Since only the response variable and the predictor variables are supposed to be aligned, the presence of the above phenomenon would be a major problem. To check this assumption, I analyzed the value of the variance inflation factor (VIF) between each set of predictor variables, such that if the value of any VIF was at least ten, multicollinearity was present in the model. Since this was not the case, no multicollinearity was present in the model.

5.2 Model Diagnostics for Provisions Two and Three

After fitting the model described above, I proceeded to check the assumptions for the fitted multiple linear regression model. There are four assumptions that are made when employing multiple linear regression, which all revolve around the validity of assuming a straight-line (linear) relationship between the response variable and each of the explanatory variables. Specifically, here were the assumptions that were made when running the model:

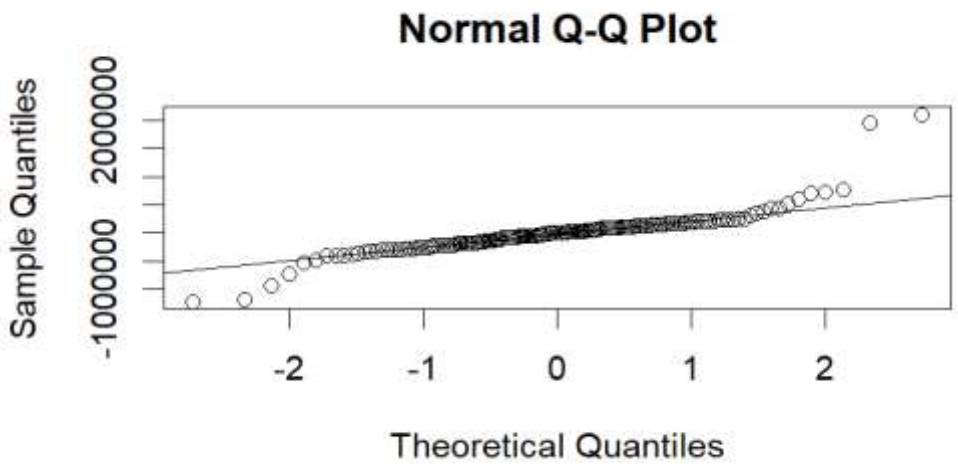
1. The mean of the residuals (the difference between the fitted values of the response variable as calculated by the regression model and the observed values of the response variable found in the dataset) is approximately centered at zero. That is, the mean of the residuals is zero such that no error term is included in the model. I checked this by plotting the residuals themselves against the individual values of the response variable, Average Rebate, generating a residual plot. This assumption was met.



2. The variance (spread) of the residuals is constant over all fitted values of the response variable. In other words, it is not the case that, when the value of the residuals increases, the value of the response variable increases. The term for this assumption, in which the variance of the residuals is constant over all fitted values of the response variable, is homoscedasticity. I checked this assumption using the residual plot explained in the preceding assumption as well as the Breusch-Pagan Test. The Breusch-Pagan test is a quantitative hypothesis test in statistics used specifically for checking this assumption, where the level of significance is set in advance. Using a significance level of five percent ($\alpha = 0.05$), the result of the hypothesis test was that there was insignificant evidence to indicate that there was an absence of homoscedasticity in the model, with the corresponding p-value equating to approximately 0.49.

3. The residuals are normally distributed with a mean of zero and a constant variance over all values of the fitted response variable. That is, the distribution of the residuals themselves follow a bell-shaped (normal) distribution. To verify this assumption, I used a

normal probability plot in which a linear pattern of data points indicates a normally distributed set of residuals, where the axes of the plot are the sample quantiles and the theoretical quantiles.



4. There is no multicollinearity in the dataset. Multicollinearity means that two or more of the predictor variables have a correlation coefficient of greater than ninety percent, indicating a significant correlation. Since only the response variable and the predictor variables are supposed to be aligned, the presence of the above phenomenon would be a major problem. I checked this assumption by analyzing the value of the variance inflation factor (VIF) between each set of predictor variables, such that if the value of any VIF was at least ten, multicollinearity was present in the model. Since this was not the case, no multicollinearity was present in the model.

Chapter 6

Future Research

Although I was able to explore and analyze much in this thesis, there remains much to determine on a number of fronts regarding the nature and quantitative analysis of these three provisions. Specifically, while I have discovered much with regards to how geographic rating area factors and the actuarial value of the different metal tiers impact the premium for a standard family of four, what variables impact the aforementioned rating area factors and metal tier actuarial values, and the variables with the largest impact on medical loss ratio rebates, it will be essential to stay up-to-date with the changing nature of the trends that have been shown and that will be experienced in the future. For example, the COVID pandemic has had a major impact on not only these three provisions but on the health insurance market as a whole, and it will be immensely intriguing to see if the pattern of the number of company rebates having the largest impact on the level of medical loss ratio rebates will continue. As such, monitoring the model developed in this thesis and modifying it with the passage of time will be crucial in not only ensuring accurate prediction and representation of health insurance trends, but also in understanding what confounding variables may be impacting the results. If these three provisions were simple by nature, this thesis could not have been written.

Bibliography

Garcia, Alise, and Martha King. *The Affordable Care Act: A Brief Summary*, National Conference of State Legislatures, Mar. 2011.

PUBLIC LAW 111–148, 2010, pp. 1–906. 111th Congress.

Rosenbaum, Sara. "The Patient Protection and Affordable Care Act: Implications for Public Health Policy and Practice." *Public Health Reports*, vol. 126, no. 1, 2011, pp. 130–135.

"Summary of the Affordable Care Act - Kaiser Family Foundation." *Focus on Health Reform*, The Henry J. Kaiser Family Foundation, 25 Apr. 2013.

United States, Congress, *Public Law 111-148: The Patient Protection and Affordable Care Act Summary*. 2010, pp. 1–57.

"Medical Loss Ratio." *NAIC*, National Association of Insurance Commissioners, 28 July 2021.

"Standardized Health Plans: Four Tiers of Coverage." American Plasma Users Corporation.

Wu, Jonathan. "What Is an Insurance Ratings Area?" *ValuePenguin*, ValuePenguin, 14 Dec. 2020.