Predicting Medical Expenses Using Multiple Linear Regression

Santiago Quinonez

2023-04-30

Research Question

In this study, I will be conducting an analysis of the 'insurance.csv' file which contains 1,338 cases of individuals and their respective yearly medical expenses. This analysis will be focused on using multiple linear regression to predict the yearly medical expenses of each individual policyholder.

The purpose of this analysis is to answer the following research question: "How can a start-up medical insurance company use the available information collected on each individual policyholder to estimate how much it will have to to pay out in claims on a yearly basis?"

With the medical insurance company being the decision maker in this scenario, this analysis is designed to inform an insurance company how they could use the generic data collected on their policyholders to estimate the amount that they should charge each client in yearly premiums to cover the expenses incurred from paying out claims over this time period. Since a profitable insurance company needs to retain more in collected premiums than they pay out in yearly claims, this analysis will show the company the minimum value they should charge in premiums in order to match the estimated medical expenses that it will have. Therefore, it is crucial for insurance companies to conduct this type of analysis on their policyholders as doing so will give the company's management team a guideline for what the company needs to be profitable.

Data Set

The 'insurance.csv' file is simulated data set containing hypothetical medical expenses for patients, which was created using demographic statistics from the US Census Bureau to approximately reflect real-world conditions. This data set was specifically created for the book "Machine Learning with R" by Brett Lanz, and is avaialable for download at the following link: https://github.com/PacktPublishing/Machine-Learning-with-R-Third-Edition/tree/master/Chapter06 (https://github.com/PacktPublishing/Machine-Learning-with-R-Third-Edition/tree/master/Chapter06). While it would be optimal to use real observations containing real observations of medical insurance information, insurance companies do not publicly disclose this information for data protection and medical record privacy reasons.

This data set contains only one year of data, meaning that it is a cross-sectional dataset due to all of its observations occurring at a single point in time (during one fiscal year). Due to medical health information regulations, it is very difficult for medical insurance companies to exchange medical records and information, as this would violate federal Health Insurance Probability and Accountability (HIPPA) laws and other local medical record privacy laws put in place in each state. As such, using one year of data would be realistic for a start-up medical insurance company who only has a year's worth of proprietary data.

This data set contains the following variables: the age of the policyholder, their sex, their body mass index (bmi), the number of dependants that they have on the policy, their smoking habits, their region of inhabitance, and the yearly medical expenses. All of these variables contain descriptive attributes of each policyholder, excluding the yearly medical expenses which should the value of the yearly claims in medical expenses that the policyholder had. In this analysis, the yearly medical expenses will be the dependent variable while the descriptive attributes will be the independent variables.

This data set is aggregated on an individual level, as each observation contains information about a single policyholder. However, one of the variables in this dataset contains the region of inhabitance of that individual, meaning that this variable can be used to manipulate the data to be aggregated on a regional level as well.

Data Exploration

The first step of this analysis was be to thoroughly explore this data set. In doing so, I was be able to gain a better understanding of each variable within this data set, allowing me to discover patterns within the data set, draw conclusions about the data set's reliability, identify potential outliers, and examine the normality of continuous variables.

Describing The Relevant Variables

The following is a table describes each variable in the data set:

Exhibit 1: Descriptive Table of Varaibles

Variable Name	Name in Dataset	Variable Type	Description
Age	age	continuous	The age of the policy holder
Sex	sex	categorical	Whether the policy holder is a male or a female
Body Mass Index (bmi)	bmi	continuous	Measure of body fat based on weight and height
No. of dependants	dependants	integer	Number of dependants on the policy
Smoker	smoker	categorical	Whether the policyholder smokes ("yes") or does not ("no")
Region	region	categorical	Regions in the US: southeast, southwest, northeast, or northwest
Medical Expenses	expenses	continuous	Yearly medical expenses charged to insurance plan (in dollars)

Generating Summary Statistics (Including Missing Values)

Next, I generated a table of summary statistics that allowed me to examine the reliability of the dataset by examining the number of missing observations, while also displaying the minimum, median, mean, maximum, and standard deviation values for each variable.

Exhibit 2: Descriptive Statistics of Continous Variables

	count	no. missing observations	min	median	mean	max	st. deviation
age	1338	0	18.00	39.00	39.21	64.00	14.05
bmi	1338	0	16.00	30.40	30.67	53.10	6.10
dependants	1338	0	0.00	1.00	1.09	5.00	1.21
expenses	1338	0	1121.87	9382.03	13270.42	63770.43	12110.01

I was able to draw several conclusions from Exhibit 2. First, when examining the reliability of the data, I could see that each variable has 1,338 observations, meaning that there are no missing values for any of these variables. Therefore, I concluded that the data for these variables was reliable since it contained an observation for each individual in this data set.

When examining the data for normality, I could see that the mean was relatively close to the median for every variable excluding yearly medical expenses, where the mean is significantly higher than the median. This made logical sense, as I would expect that in a random set of medical cases the majority of cases would be relatively normal in severity and therefore relatively inexpensive, while a few complicated and rare cases would lead to unusually high medical expenses due to rare conditions which require extensive care. These severe and expensive cases would push the mean value of medical expenses to be much higher than the median value, which is what is displayed in Exhibit 2. From this, I thought that it would be likely that the 'expenses' variable would likely be skewed to the right instead of having normal distribution, while the other variables would likely have a curve similar to a normal distribution. Of course, this is hard to conclude by only looking at summary statistics and therefore decided to generate histograms on these variables, which are displayed below in Exhibits 3-6.

Generating Histograms of Each Variable to Examine Distributions

Exhibit 3: Histogram of Policy Holder Age

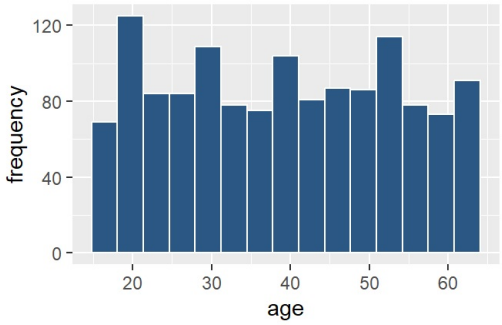


Exhibit 4: Histogram of Body Mass Index

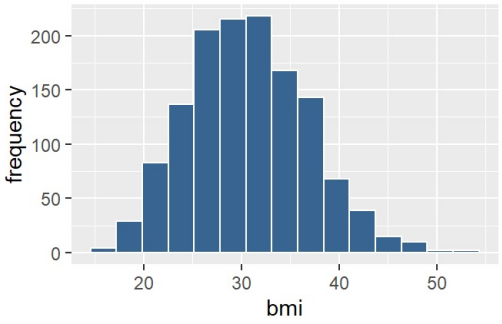


Exhibit 5: Histogram of Dependants on Policy

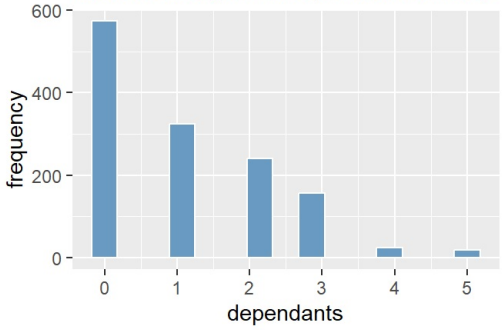


Exhibit 6: Histogram of Yearly Medical Expense

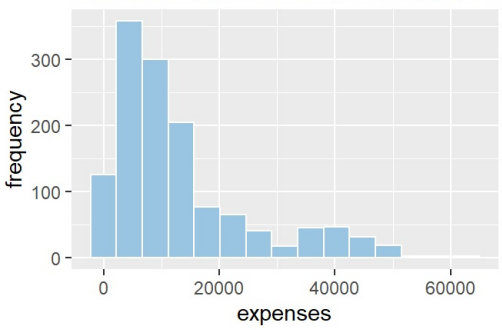


Exhibit 6 proves the previous assumption that I made regarding the skew in the distribution of the yearly medical expenses variable, as the majority of observations having medical expenses being around $10,000 (close to the median value), yet there are a significant number of cases being significantly higher at more than $30,000. On the other hand, Exhibit 4 shows that the age of participants appears to have a uniform distribution, Exhibit 5 shows that the body mass index appears to be normally distributed, and Exhibit 6 shows that there the number of dependants on each policy has a distribution that is skewed to the left, as the majority of individuals have no dependants on their policy and the number of observations decrease with each increase in reported dependants.

Generating Boxplots to Examine Potential Outliers

Next, I created boxplots of these variables to look for potential outliers. These boxplots are displayed below in Exhibits 7-10.

Exhibit 7: Boxplot of Policy Holder Age

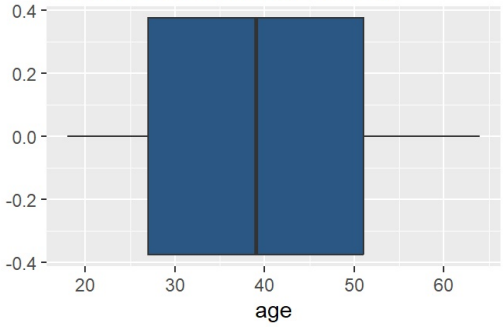


Exhibit 8: Boxplot of Body Mass Index

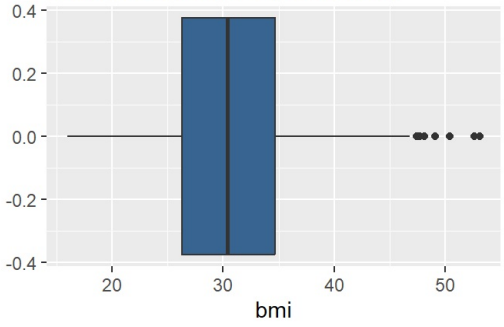


Exhibit 9: Boxplot of Dependants on Policy

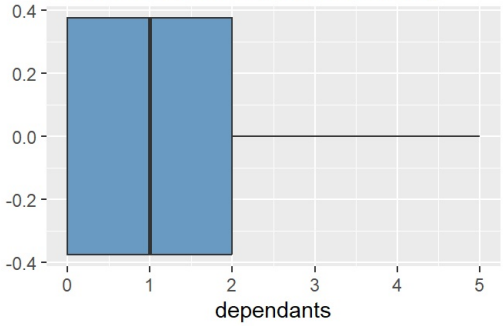


Exhibit 10: Boxplot of Yealy Medical Expenses

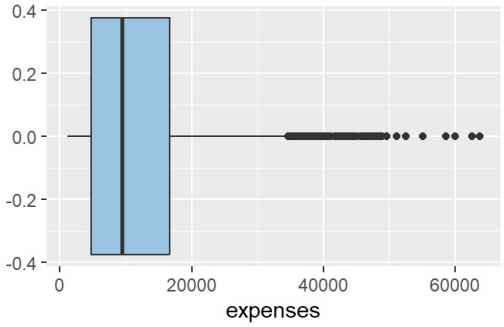


Exhibit 8 shows that there are a small amount of outliers in the observations of body mass index on the right side of the boxplot. Logically, this means that the majority of the individuals in the data set have a normal weight profile, while a few individuals are significanlty overweight and therefore have body mass index that is considered outside the normal range. Exhibit 10 confirms my previous thoughts on the yearly medical expenses of individuals in this data set, showing that the majority of cases are in a range between $500 and 17,000. Furthermore, Exhibit 10 also shows that there is a large number of cases outside the a normal range, with outliers beginning at a value of about $35,000. Logically, the outlier cases indicate that there a substantial number of cases in which medical expenses far exceeded the normal range, meaning that the policyholder received dramatic and extensive care due to their condition. Lastly, Exhibits 7 and 9 show that the age of the policy holder and the number of dependants on the policy are not likely to have outliers within their data. Since the age of the policyholders contains observations that are normally distributed, and the number of dependants on each policy do not exceed a number outside the normal range of individuals in a household (the data set contained no observations were the number of dependents was above 6), it makes sense that these two variables are unlikely to have any outliers within the data set.

Of course, I cannot remove outliers of yearly medical expenses from this analysis since it is important for a medical insurance company to take these cases into account when making predictions about how much they will have to pay out in claims over a year. As such, it is imperative that these outlier cases have an effect on the predictions from the linear regression model that will be created later on, as removing these outlier cases would lead to an underestimation in medical expenses, which would likely mean that the insurance company would pay out more in claims than it received in premiums.

Removing outliers for body mass index could also be problematic, as people who are overweight are naturally more likely to have health problems. However, doctors around the world use the BMI value of 30 as an indicator of whether a person is normal or obese. Logically, this means that body mass index could possibly turned into an indicator variable where any value below 30 was normal and any value above 30 was obese to observe how these two differnt categories affect the yearly medical expenses for policyholders. This will be tested later on in my analysis.

The only reason that I would remove these outlier cases would be if the company had a value of yearly medical expenses at which they refused to insure individuals at. However, in doing so, the company would miss out on sales and potential profits from not insuring individuals. Of course, this is a risk-reward scenario, as individuals with high expected yearly medical expenses could be riskier to insure. However, because of their risk profile is much higher, it is likely that the company could charge very high premiums to the individual for coverage, thus adding a possibility for higher profits. Since the basis for profitability for insurance companies comes from accurately estimating the claims they will have to pay from medical expenses, I will not remove the outlier cases as they will be a kay part of evaluating the acccuracy of the linear model.

Generating Frequency Bar Graphs and Relative Frequency Tables for Categorical Variables

After examining outliers, I analyzed the categorical variables by plotting the frequency of each category for each variable on a bar chart and then examining their relative frequency through a proportion table. At the same time, I was able to confirm that none of these variables contained missing observations as their frequencies added up to 1,338 observations for each variable.

The first variable that I looked into was the sex of each policyholder, where I was able to see that there was a relatively even split between males and females, with males making up around 51% of observations in the dataset, while females made up 49%. This is shown in Exhibits 12 and 13.

Exhibit 11: Bar Chart of Sex Variable

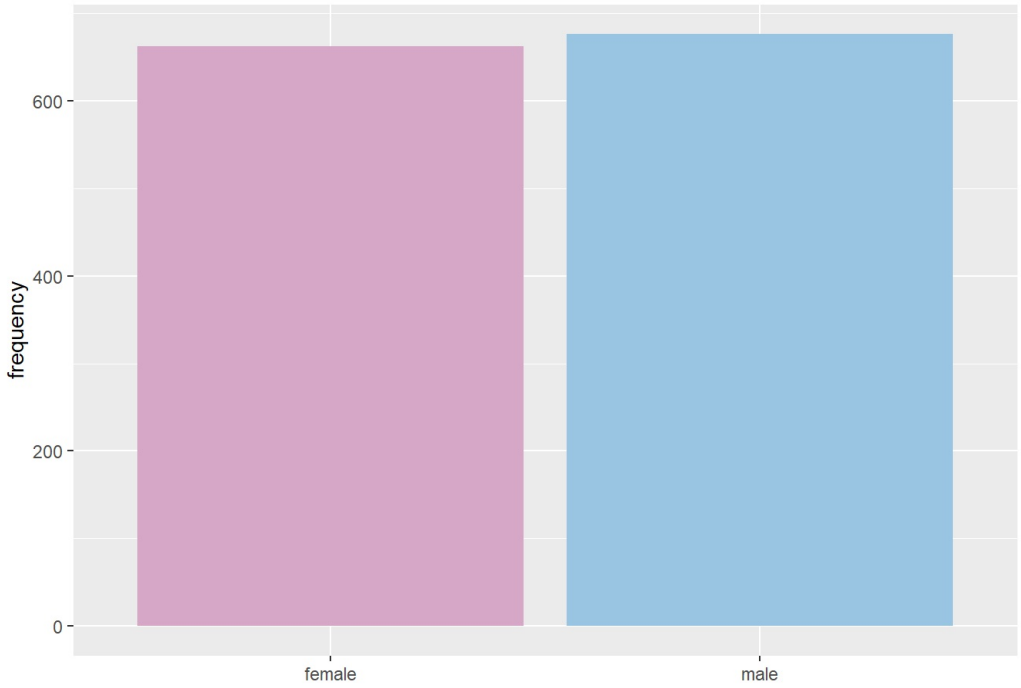


Exhibit 12: Frequency and Proportion Table of Sex Variable

	frequency	proportion
female	662	0.49
male	676	0.51

Next, I took a look at the smoker variable, where I was able to see that roughly 80% of the individuals in this data set were not smokers while roughly 20% smoked. This is shown in Exhibits 14 and 15 below.

Exhibit 13: Bar Chart of Smoker Variable

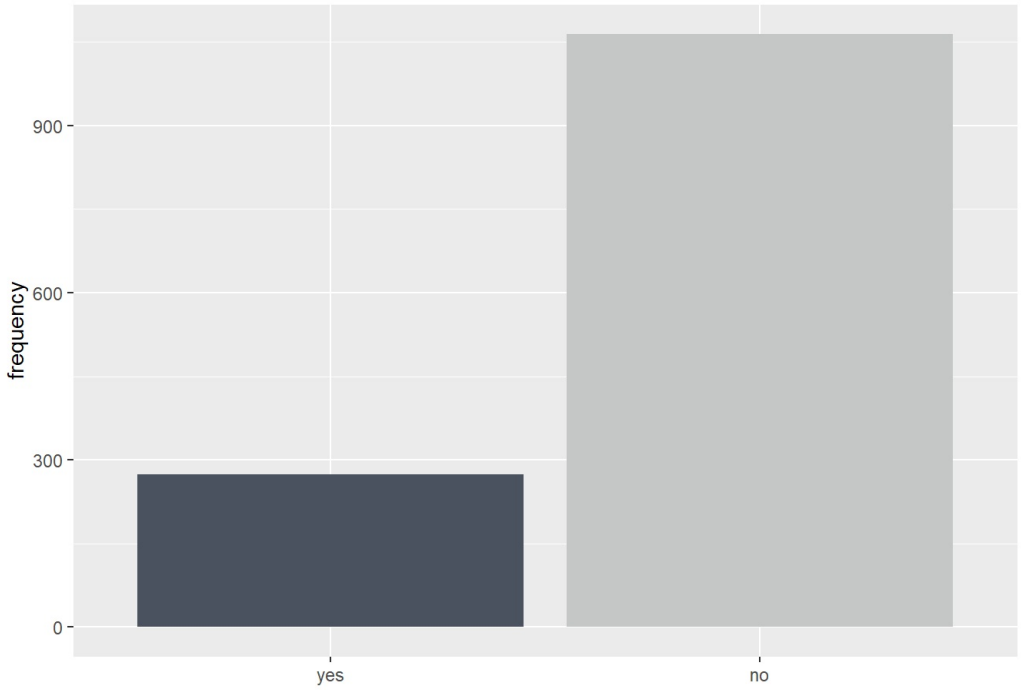


Exhibit 14: Frequency and Proportion Table of Smoker Variable

	frequency	proportion
yes	274	0.2
no	1064	0.8

For the last categorical variable, I was able to see that each region made up roughly 24% of the observations in the datasets, with the southeast region being the only exception, making up 27%. This is shown in Exhibits 16 and 17 below.

Exhibit 15: Bar Chart of Region Variable

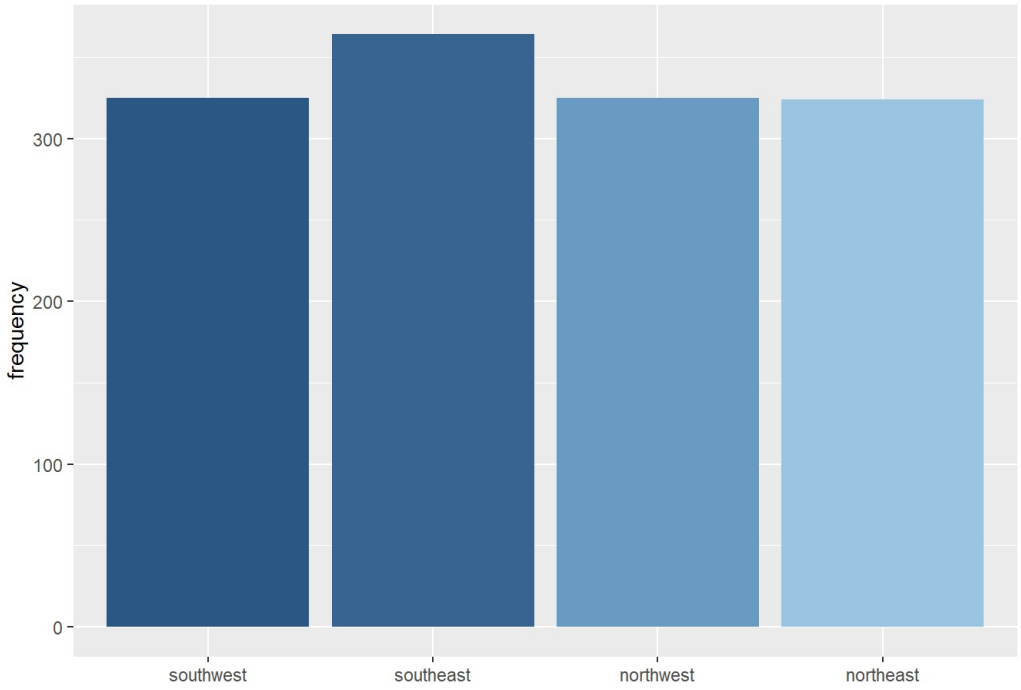


Exhibit 16: Frequency and Proportion Table of Region Variable

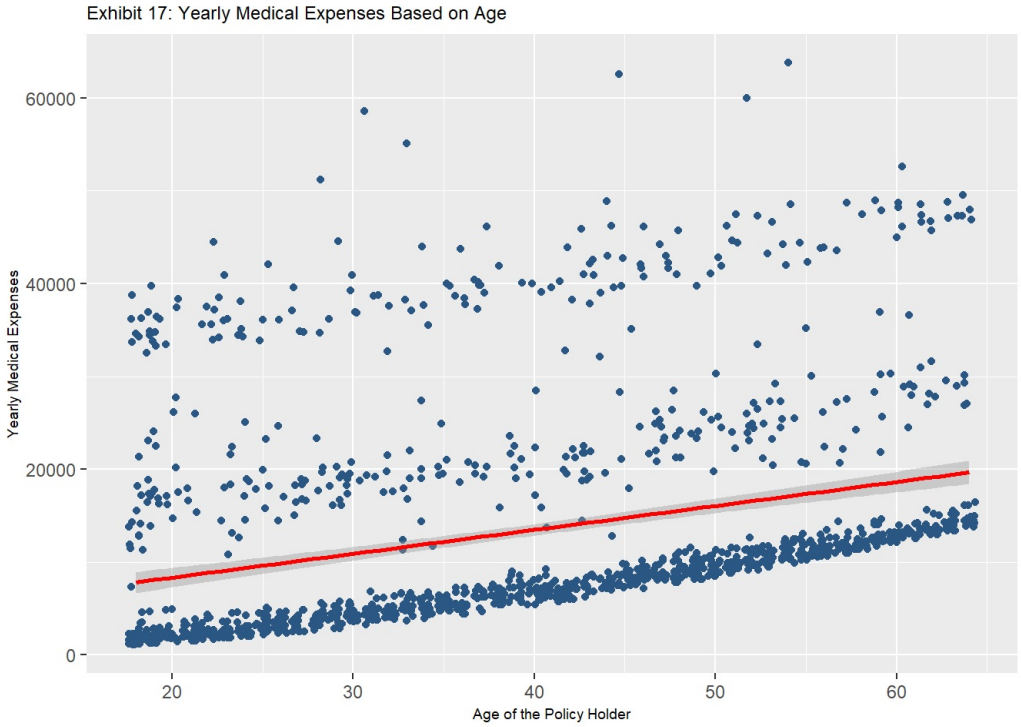
	frequency	proportion
southwest	325	0.24
southeast	364	0.27
northwest	325	0.24
northeast	324	0.24

Generating Univariate Plots Between Independent Variables and Dependent Variable

After examining the relative frequencies of each categorical variable, I decided to generate univariate plots to examine the relationship between each independent variable and the dependent variable. I did this by generating scatterplots between the continuous independent variables and yearly medical expenses, and by generating boxplots between the categorical independent variables and the dependent variable.

First, I generated a scatterplot between yearly medical expenses and the age of the policyholder, which is displayed below in Exhibit 17.

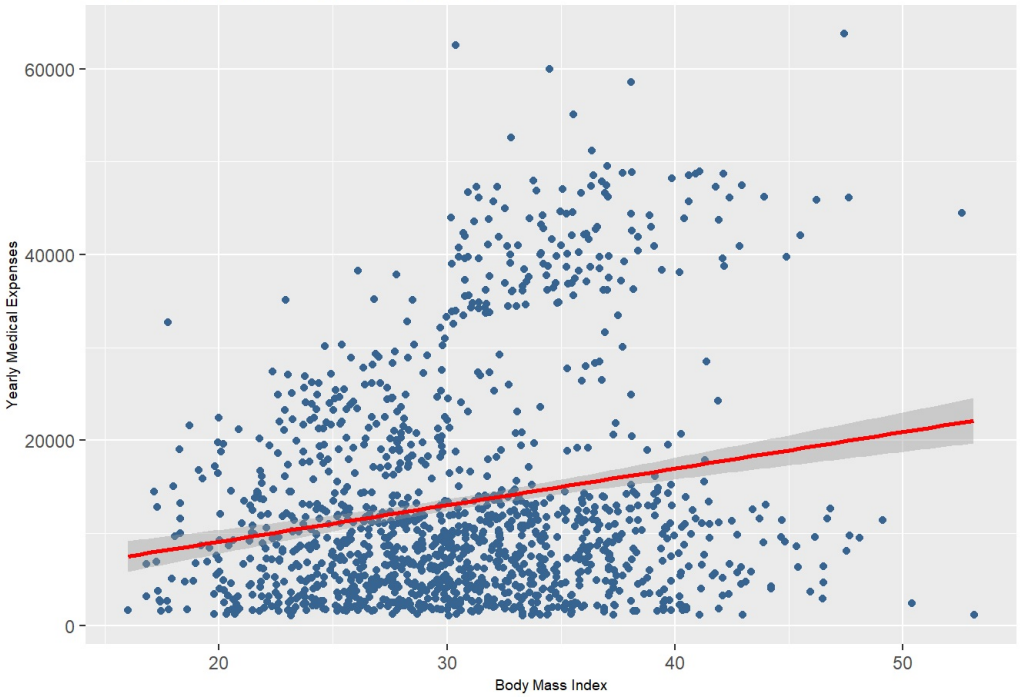
Exhibit 17: Yearly Medical Expenses Based on Age



The scatterplot displayed in Exhibit 20 is very interesting, as there appears to be a linear relationship between yearly medical expenses and the age of the policyholder, but in three separate cohorts. This is because there are three bands of observations which tend to follow an upward trend between age and yearly medical expenses, with each band starting at a different y intercept. From my perspective, this means that there are three separate cohorts of the population that have expenses that increase as age increases. The first cohort (with the lowest y-intercpet) would be minor medical expenses which are relatively low in cost and do not require hospitalization or specialized treatments. These minor medical expenses could include over-the-counter-medications, routine check ups, and minor injuries or illnesses that can be treated at home or in an outpatient clinic. The next cohort would be moderate medical expenses, which are expenses which are moderately costly and may require more specialized treatment or hospitalization. These kinds of expenses could include surgeries, emergency room visits, and diagnostic tests such as MRIs or CAT scans. The last cohort (with the highest y-intercept) would be major medical expenses, which are expenses that are significantly costly and often require extensive long-term treatment or hospitalization. These medical expenses could include expenses like chronic illnesses such as cancer, organ transplants, and prolonged hospital stays. While these three categories are not displayed in a categorical variable in this data set, I realized from this plot that there could be a potential interaction between this variable and another which could explain this relationship further, like BMI or whether the individual is a smoker or not. This is because having a higher BMI and/or being a smoker combined with being older could increase the chances of having moderate or major medical expenses.

Next, I generated a scatterplot between yearly medical expenses and the BMI of the policyholder, which is displayed below in Exhibit 18.

Exhibit 18: Yearly Medical Expenses Based on BMI



The scatterplot displayed in Exhibit 18 also shows a likely linear relationship between yearly medical expenses and body mass index, where the medical expenses rise as body mass increases. However, there is a cluster of observations that appear to have significantly higher yearly medical expenses as the body mass index gets to 30. This means that 30 is likely to be a significant number when looking at BMI index, as previously stated when eximining the data for outliers. As such, I decided to create an indicator variable from the BMI variable called bmi30, where observations which contained a body mass index would receive a 0, and ones with a body mass index that was higher than 30 would be represented by a 1. I then used this indicator variable to examine how the effect of BMI being over or under 30 would impact yearly medical expenses through a boxplot which is displayed below in Exhibit 19.

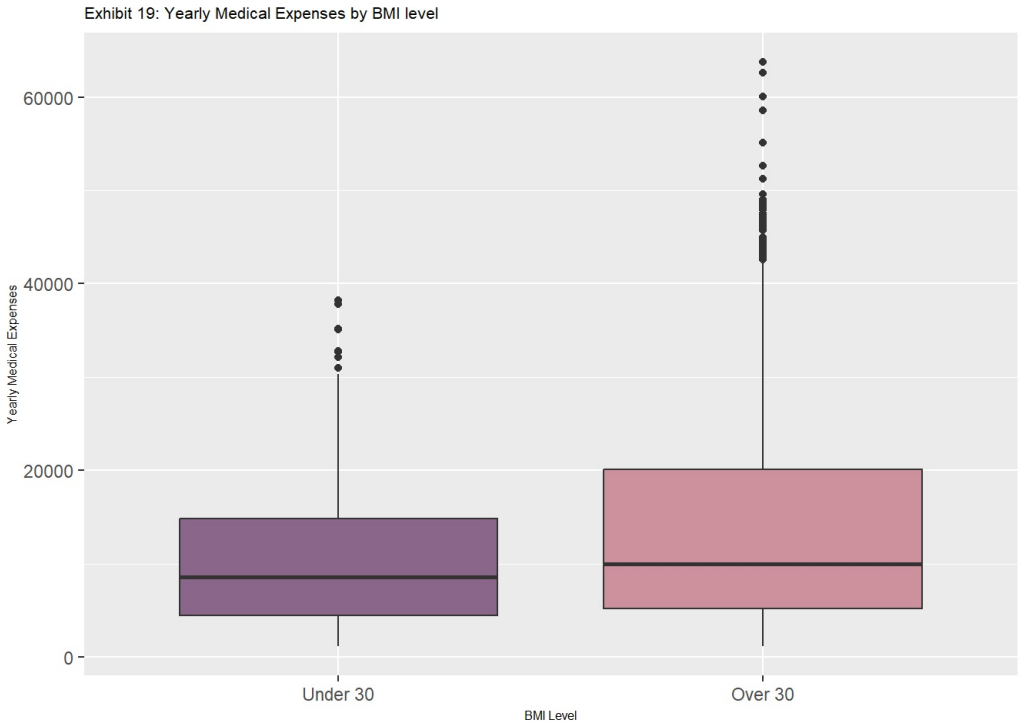


Exhibit 19 shows that when the BMI is over 30, there are significantly more obervations which contain higher expenses. This is represented by observations where the BMI is higher than 30 having a wider interquartile range, but also having the 3rd quartile be substantially higher on the y-axis (representing expenses) than the 3rd quartile than observations where the BMI lower than 30. At the same time, observations which have a BMI over 30 have outlier cases between roughly 40,000 - 60,000 dollars in yearly medical expenses, while observations where the BMI is under 30 have outlier cases between 30,000 - 40,000 dollars. This is another substantial difference.

I also wanted to see what proportion of the observations had a BMI over 30 in this data set, and therefore I created a proportion table and displayed it in Exhibit 20.

Exhibit 20: Proportion Table of bmi30 Variable in Minor Medical Expense Cases

	Frequency	Proportion
Under 30	631	0.47
Over 30	707	0.53

Exhibit 20 shows that roughly 53% of the policyholders in this dataset have a BMI over 30, meaning that this data set contains a substantial amount of people that doctors would consider to have obesity problems.

Next, I plotted a scatterplot between the number of dependants on the policy and the yearly medical expenses for each policyholder. In this plot, I expected there to be a clear distinction linear relationship between yearly medical expenses and the number of dependants on the policy, as it would make sense that more people on the policy would lead to higher medical expenses. However, this was not exactly the case, as shown below in Exhibit 21.

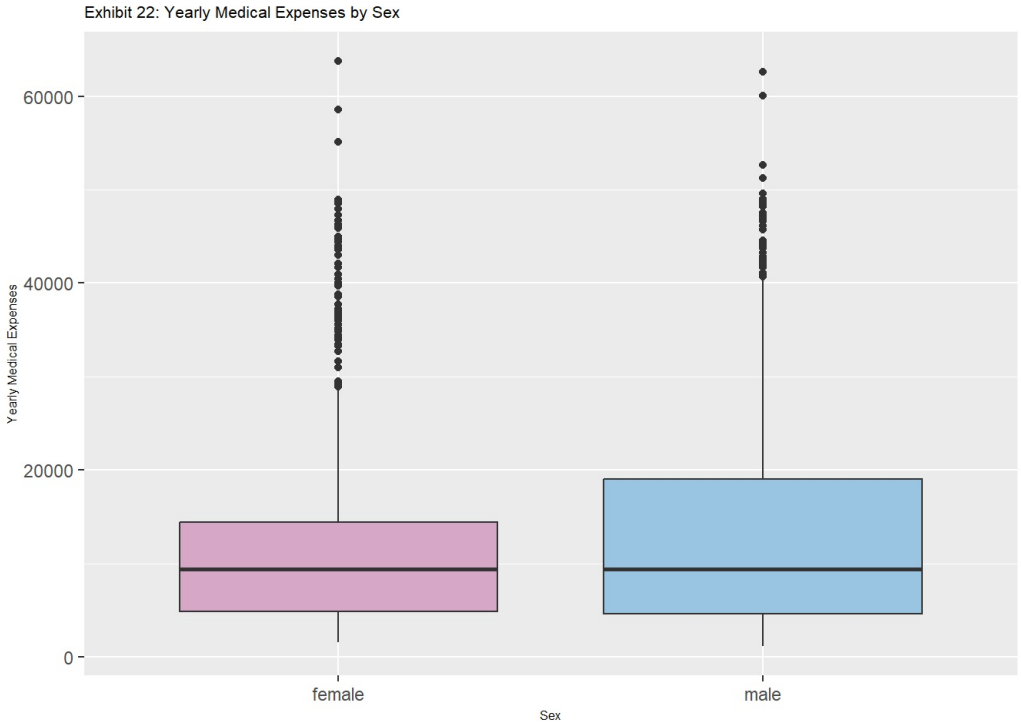
Exhibit 21: Yearly Medical Expenses Based on dependants on Policy

Exhibit 21 shows that there is a slightly positive linear relationship between yearly medical expenses and the number of children on the policy. However, this is not very easy to conclude visually as there are lower amounts of observations as the number of dependants increase, meaning that it is not easy to see what the overall impact of the number of dependants has on the yearly expenses. From this, I became skeptical of how impactful this predictor variable would be in my linear model due to the phenomenon that I was able to observe from this Exhibit.

The next step of my analysis was to generate boxplots which allowed me to see the relationship between independent categorical variables and the dependent variable.

I first started by generating a boxplot between the sex of the policy holder and the yearly medical expenses, which is displayed in Exhibit 22 below.

Exhibit 22: Yearly Medical Expenses by Sex



From Exhibit 24, I was able to see that males and females have similar median values of yearly medical expenses, yet males have greater ranges of variation in yearly medical expenses, as the interquartile range in the boxplot of males was larger than the one of females. At the same time, I can see that the largest non-outlier observation in the male boxplot is at roughly $40,000 of yearly medical expenses, while the lowest non-outlier observation of females was nearly 10,000 dollars lower. Since the proportions of males and females were roughly equivalent, this means that males are more likely to have cases which require major medical expenses even though males and females have similar median values of yearly medical expenses, .

Next, I plotted the smoker variable against yearly medical expenses, where I expected medical expenses to be substantially higher for smokers than for nonsmokers. This turned out to be correct, as displayed below in Exhibit 23.

Exhibit 23: Yearly Medical Expenses by Smoker

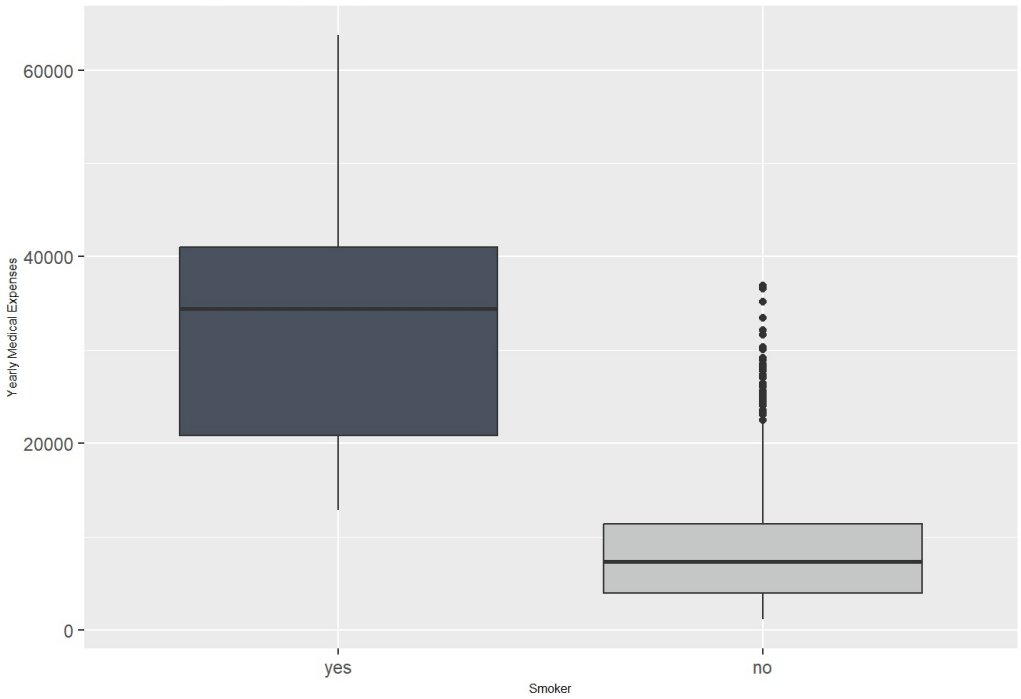


Exhibit 23 shows that there is a substantial difference in yearly expenses between smokers and nonsmokers. Smokers had a median value of approximately $37,000 in the dataset, while nonsmokers had a median value of below 10,000 dollars. This is a difference of more than 25,000, which is bigger than the entire interquartile range of the nonsmoker boxplot. Interestingly, the difference between these factors is so pronounced that the largest non-outlier value of nonsmokers was about 22,000 dollars, while the largest value for non-outlier values the maximum value in the dataset of about 63,000.

Lastly, I created a boxplot of the region variable relative to the yearly medical expenses. In this boxplot, I expected that nearly every region would have similar median values for expenses, as there is nothing about a particular region that would lead me to think that individuals in that region would have higher yearly medical expenses. This boxplot is displayed below in Exhibit 24.

Exhibit 24: Yearly Medical Expenses by Region

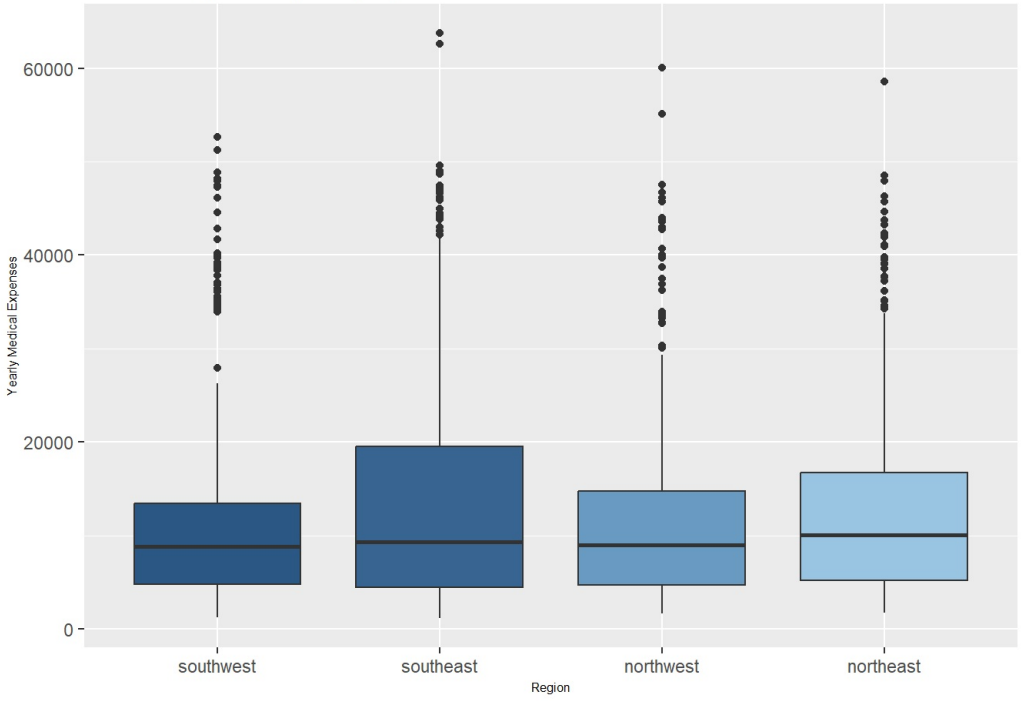


Exhibit 24 shows that my prediction was partially right, as the median value for all of these regions were approximately the same at a value of roughly 10,000 in yearly medical expenses. However, the southeast region has significantly higher non-outlier values than all of the other regions at approximately 41,000 dollars, while all the other regions have their highest non-outlier values below 35,000. However since this is only one region that is different, I became skeptical of how impactful this variable would be in predicting yearly medical expenses.

From all of these univariate plots (both the scatterplots and boxplots), I was able to see that all the variables in the data set were likely to provide meaningful impact in predicting yearly medical expenses, possibly excluding the number of dependants on the policy and the region in which the policyholder lives in.

Generating a Correlation Matrix Between Continuous Variables

After generating univariate plots between the independent and dependent variables, I decided to examine the correlations between the continuous variables in this data set. In doing so, I would be able to get an understanding of the relationship that each independent continuous variable and yearly medical expenses. To achieve this, I created a correlation matrix which would show me the correlation coefficients between the independent

and dependent variable, as well as the correlation coefficients between each of the independent variables. This correlation matrix is displayed in Exhibit 24 below.

Exhibit 25: Correlation Matrix Heatmap

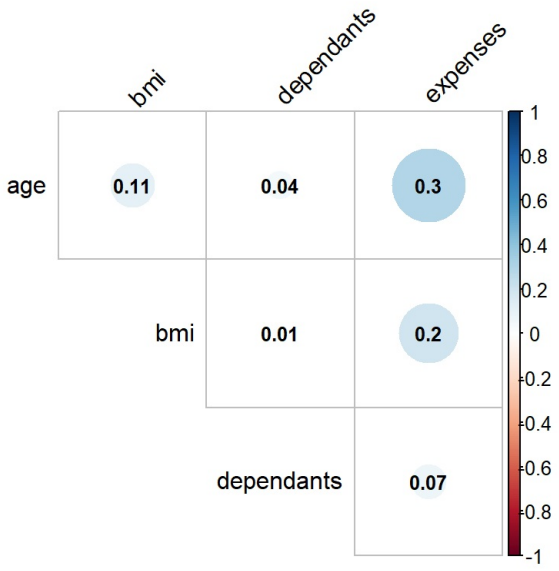


Exhibit 25 shows that none of the independent variables are highly correlated with the dependent variable (yearly medical expenses), as no relationship has a correlation coefficient that is higher than 0.30. Age and medical expenses have the strongest positive relationship, exhibiting correlation coefficient of 0.30, meaning that as age increases, medical expenses tend to slightly increase as well. BMI and yearly medical expenses also have a slightly positive relationship, as they have a correlation coefficient of 0.20, meaning that as the body mass index of an individual increases, the medical expenses also slightly increase. Lastly, the number of dependants on the policy and yearly medical expenses also has a very slightly positive relationship, having a correlation coefficient of 0.07. This means that as the number of dependents on the policy increase, the yearly medical expenses tend to marginally increase as well.

Exhibit 25 also shows that none of the independent variables are highly correlated with each other, as the correlation coefficients between each of these variables fail to exceed 0.11. This is actually an encouraging sign, as this means that it is unlikely that my model will contain multicollinearity (when independent variables are highly correlated), which can cause problems in the model as it could lead to difficulties in determining the individual effect of each of the independent variable on the dependent variable. While the correlation matrix showed encouraging signs, this will be formally tested later on in a multicollinearity analysis using variance inflation factor tests.

Creating Subsets of the Data Based on Specific Criteria

The last step in exploring this data set was to create subsets of this data based on specific criteria. In the previous steps of data exploration, I found that this data set could be split into three different groups: minor medical expenses, moderate medical expenses, and major medical expenses. This was shown very clearly in Exhibit 17, where there were three bands of linear clusters containing observations of medical expenses. The first band was from 0 - 17,000 dollars in medical expenses, the second between 17,000 and 31,000 dollars, and the last band contained observations where medical expenses were above 31,000 dollars.

Because of this, I created three subsets of data based on that criteria.

After creating these three subsets, I explored how each subset of data is different. Specifically, I wanted to see if these three subsets differed substantially in the percentage of observations that contained smokers and people with a BMI over 30. As such I created proportion tables for each of these variables within each subset of data.

First, I did this for the subset of data which contained minor expenses (less than $17,000).

Exhibit 26: Proportion Table of Smoker Variable in Minor Medical Expense Cases

	Frequency	Proportion
yes	22	0.02
no	985	0.98

Exhibit 26 shows that only 2% of people in this subset are smokers, which is very different from the full data set where 20% of the policyholders were smokers. This is very interesting, as it shows that being a smoker could be very influential in determining health problems which lead to higher yearly medical expenses.

Next, I will do a proportion table of individuals with BMI values of either under or over 30.

Exhibit 27: Proportion Table of bmi30 Variable in Minor Medical Expense Cases

	Frequency	Proportion

Under 30	490	0.49
Over 30	517	0.51

Exhibit 27 shows that about half of the individuals in this subset have a BMI over 30, which is slightly lower than the full data set's value of 53%. While this is not as substantial of a difference as seen in Exhibit 26, this could still indicate that individuals with a BMI below 30 are more likely to have lower yearly medical expenses.

Lastly for this subset of data, I will create a multi-panel facet wrap plot to examine how both of these variables combined interact together with yearly medical expenses. This plot is displayed below in Exhibit 26.

Exhibit 28: Expenses by bmi30 and smoker in Minor Medical Expense Cases

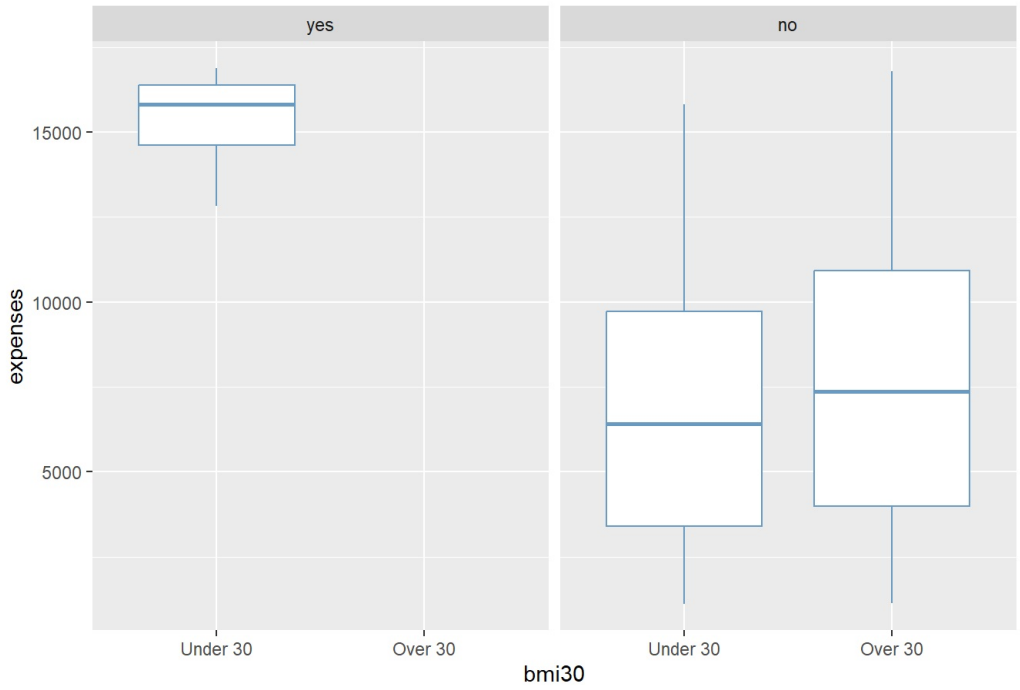


Exhibit 28 shows that there are no smokers in this subset of data with a BMI over 30. However, the smokers in this subset of data (which only account for 2% of the observations), have signigicantly higher value of median medical expenses than non-smokers. For the non-smokers in this subset of data, those with a BMI over 30 have a median value of yearly medical expenses which is slightly higher then those with a BMI less than 30.

After examining the subset of data containing minor expenses, I created proportion tables of the smoker and bmi30 variable for the subset of data containing moderate expenses and then examining these variables using a multi-panel facet wrap plot.

Exhibit 29: Proportion Table of Smoker Variable in Moderate Medical Expense Cases

	Frequency	Proportion
yes	102	0.58
no	73	0.42

Exhibit 29 shows that there is a significantly larger proportion of smokers in this subsets of data, with 58% of policyholders being smokers in this subset. This is much higher than the first susbet of data, along with the overall dataset where only 2% and 20% of their observations were smokers. Again, this could be an important indicator for health problems and therefore yearly medical expenses.

Next, I will do a proportion table for the bmi30 variable on this subset of data.

bmi30 Under 30 Over 30 133 42 bmi30 Under 30 Over 30 0.76 0.24

Exhibit 30: Proportion Table of bmi30 Variable in Moderate Medical Expense Cases

	Frequency	Proportion
Under 30	133	0.76
Over 30	42	0.24

Interestingly, Exhibit 30 shows that only 24% of the policyholders in this subset of data have BMI values over 30. This is confusing, as doctors have said that having a BMI below 30 leads to lower health problems. This does show how significant the smoker variable is in determining medical expense values, as it likely that the smoker variable is very influential in determining the moderate medical expenses for this subset of data.

Lastly for this subset of data, I will create a multi-panel facet wrap plot to examine how both of these variables combined interact together with yearly medical expenses. This plot is displayed below in Exhibit 31.

Exhibit 31: Expenses by bmi30 and smoker in Moderate Medical Expense Cases

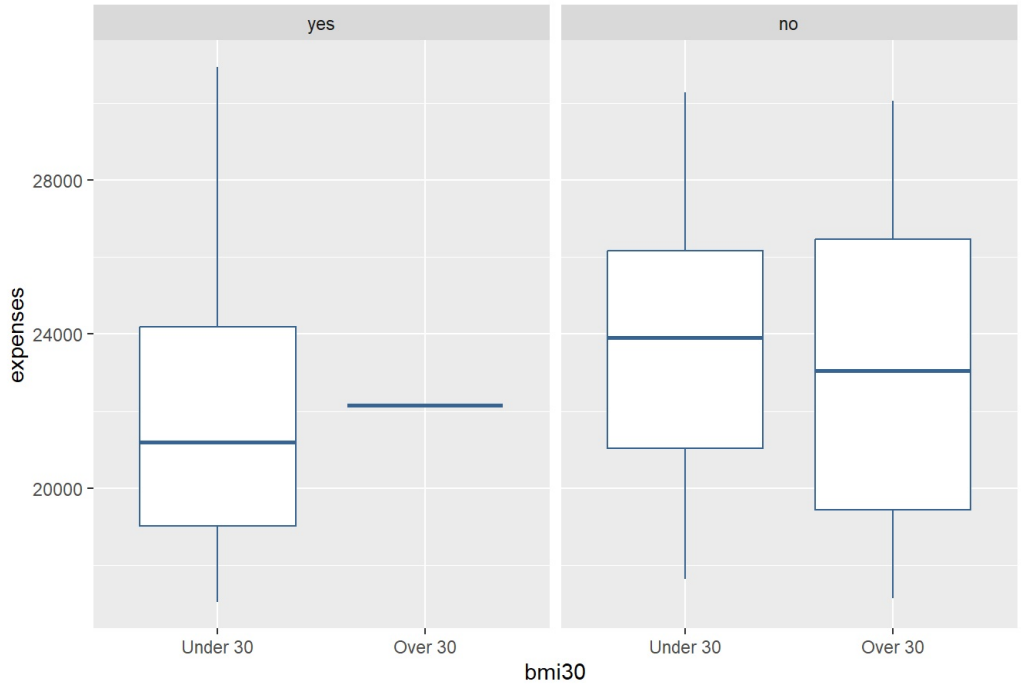


Exhibit 31 is incredible to me, as it seems highly illogical. Exhibit 31 shows that from this subset of data, medical expenses are lower for smokers than for smokers, while non-smokers with a BMI of over 30 have lower expenses than those with a BMI of under 30. While this subset of data is not indicative of the whole data set, as shown by the proportion tables in Exhibits 29 and 30, it is interesting to see that the moderate expenses susbet is very different to both the first subset and the overall data set.

Lastly, I will generate two more proportion tables for the smoker and bmi30 variables and a final multi-panel facet wrap plot for the major expenses subset of data.

First, I will create a proportion table for the smoker variable within this subset.

Exhibit 32: Proportion Table of Smoker Variable in Major Medical Expense Cases

	Frequency	Proportion
yes	150	0.96
no	6	0.04

Exhibit 32 shows that almost all of the individuals in this subset of data are smokers, with smokers accounting for 96% of the observations in this subset of data. This is obviously significantly higher than the first two susbsets of data, again showing that this is an important indicator for health problems and yearly medical expenses.

Next, I will create a proportion table for the bmi30 variable.

Exhibit 33: Proportion Table of Smoker Variable in Major Medical Expense Cases

	Frequency	Proportion
Under 30	8	0.05
Over 30	148	0.95

Exhibit 33 shows that roughly 95% of individuals in this subset have a BMI over 30, showing that in cases of very high medical expenses, obesity plays a big factor.

Lastly , I will create a multi-panel facet wrap plot to examine how both of these variables combined interact together with yearly medical expenses. This plot is displayed below in Exhibit 34.

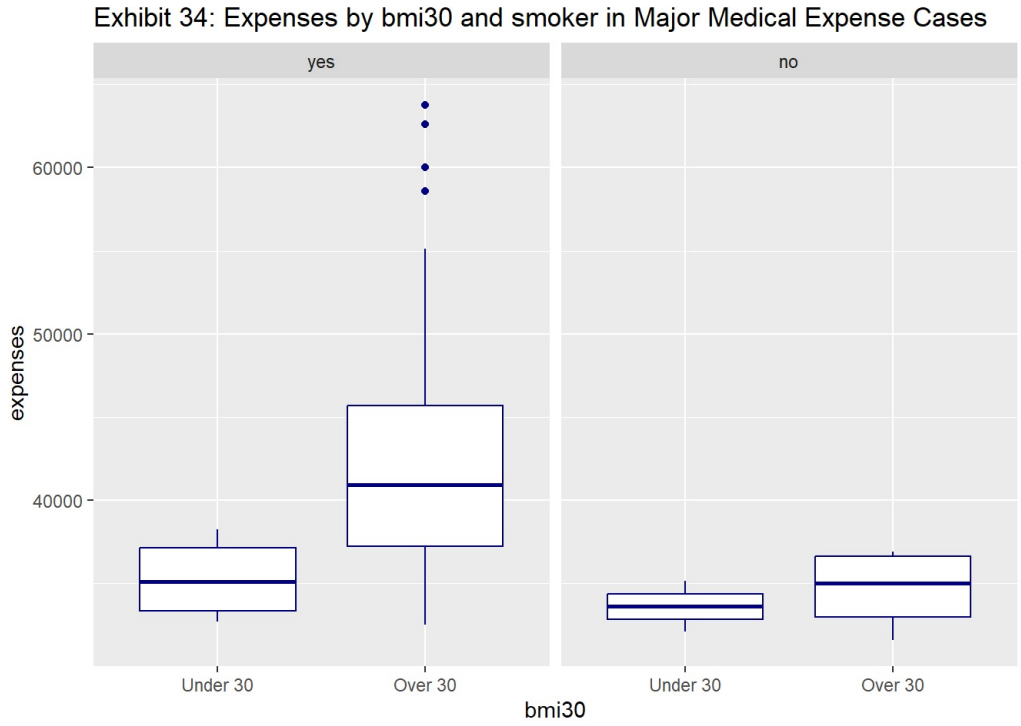


Exhibit 34 shows that the cases with the highest medical expenses of all take place when individuals have a BMI over 30 and are smokers. Medical expenses for these cases are substantially higher than any other scenario.

From subsetting the data and analyzing it, there appears to be evidence of a potential interaction effect between BMI and smoker that affects yearly medical expenses.

Model Selection

Choosing a Model

In this part of my analysis, I will be using a multiple regression model from Chapter 7 in the textbook “Introductory Econometrics: A Modern Approach.” In doing so, I will be able to use both quantitative and qualitative explanatory variables to generate predictions for my dependent variable (yearly medical expenses). This model assumes that there is a linear relationship between the independent variables and the dependent variable, thus estimating the slope and intercept of the line that best fits the data. I chose this model because it enables the use of all the relevant variables included in this data set to make predictions, especially the categorical variables that could be very influential in predicting yearly medical expenses, such as the smoking habits of the policyholder and if their BMI is over the recommended value of 30. Therefore, this model contains the format which allows for the best use of my data, and will therefore yield the best possible predictions for the dependent variable.

Model Equation

The following equation will be used to estimate the yearly medical expenses for each individual using ordinary least squares:

$$expenses = \beta_0 + \beta_1 age + \beta_2 dependants + \beta_3 bmi + \delta_0 sex + \delta_1 smoker + \delta_2 region + \delta_3 bmi30 + \delta_4 smoker * bmi30 + u$$

Explanation of Model Equation

I chose this model to address how a start-up medical insurance company could estimate the yearly claims that they will need to pay, as this model accounts for the most amount of information possible (most amount of variables/attributes), while also giving additional importance to the interaction between the BMI level of each individual and their smoking habits. By choosing this model, I will be using every attribute and the interaction between ‘smoker’ and ‘bmi30’ to generate the most accurate possible value of yearly medical expenses through ordinary least squares regression.

In this model, the β_0 is the intercept, which represents the expected value of yearly medical expenses when all of the independent variables in the model are equal to zero. The β_1 coefficient represents the expected change in medical expenses for every one unit increase in the age of the policyholder, holding all other variables constant. Similarly, the β_2 coefficient represents the expected change in medical expenses for every one unit increase in the dependents on the policy, holding all other variables constant. The β_3 coefficient represents the expected change in medical expenses for every one unit increase in the body mass index of the policyholder, holding all other variables constant. On the other hand, the δ_0 coefficient shows the difference in the yearly medical expenses between males and females, given that all other variables are being held constant. The δ_1 coefficient shows the difference in medical expenses between smokers and nonsmokers, the δ_2 coefficient shows the difference in expenses between each of the four regions, and the δ_3 coefficient shows the difference in expenses between individuals which have a BMI over 30 and those who have a BMI below 30. Lastly, the δ_4 coefficient explains the degree to which the effect of the smoking habits of an individual on the yearly medical expenses depends on the level of that individuals BMI (specifically if the BMI is above or below 30). A positive coefficient for the interaction term between smoking habits and BMI above 30 indicates that the effect of smoking on yearly medical expenses is larger for individuals with a BMI above 30 than for those with a BMI below 30.

Strengths and Weaknesses (Assumptions) of This Model

Like every model, the multiple regression model that I chose for this analysis has its strengths and weaknesses. This model is very strong when it comes to adaptability, as this model can be used in almost any modeling task as long as the data contains a similar format. At the same time, this model is by far the most common approach for modeling numeric data with both categorical and quantitative predictor variables, making its results easy to interpret to a wide audience. At the same time, this model is great at providing estimates of both the size and strength of the relationships among features and the outcome. While the strengths of this model far outweigh its weaknesses, this model does have its shortcomings. This model makes strong assumptions about the data, which could be a problem if these assumptions are not aligned with the data. These assumptions include that the relationship between independent and dependent variables is linear, that observations are independent of each other, that the variance of residuals is constant across all levels of independent variables (homoscedasticity), that the errors are normally distributed, and that the model does not contain multicollinearity or influential outliers that unduly influence the results.Also, this model's form must be specified by the user in advance, and it does not handle missing data. While the assumptions of this model are weaknesses, there are no problems with the model assumptions when it comes to the linearity or independence in the model thus far, and the rest of the assumptions will be tested for later on in this analysis.

Since this is a relatively simple model, the results will be easy to explain to the decision makers of the start-up medical insurance company. Since there are no problems with the model assumptions that compromise the reliability of the results, the model will generate the most accurate possible predictions that can possibly be made. Of course, since there are only 1,338 observations that this model is using to make predictions, the results will reflect the availability of information, meaning the model will improve at making predictions as the company collects more data. That being said, this model's results should be trusted when estimating the yearly medical expenses for each individual, as it incorporates the most amount of information possible (uses all categorical and continuous vairables, along with interactions) to generate results that minimize the sum of squared differences between actual and predicted values.

Model, Analysis, and Results

Carrying Out the Model

The next step in this analysis was to carry out this model. Carrying out this model is a multi-step process, the first of which is to use the lm() function to create a linear model from the entire data set. After creating this model, I printed the results of the coefficients (also known as estimate), its standard error, t-statistic, and p-value. These results, shown below in Exhibit 35, show how each independent variable in the model affects the results of the predictor variable in terms of magnitude and significance.

Presenting and Interpreting the Results of This Model

Exhibit 35: Results of Final Linear Model Regression

term	estimate	std.error	statistic	p.value
(Intercept)	-19399.74	1630.65	-11.90	0.00
age	264.01	9.34	28.26	0.00
dependants	518.98	108.18	4.80	0.00
bmi	1266.62	54.14	23.40	0.00
sexmale	-533.04	261.71	-2.04	0.04
smokerno	20632.79	1618.36	12.75	0.00
regionsoutheast	170.99	370.06	0.46	0.64
regionnorthwest	595.11	374.63	1.59	0.11
regionnortheast	1219.52	375.26	3.25	0.00
bmi30Over 30	3081.11	435.29	7.08	0.00
bmi:smokerno	-1449.93	51.68	-28.05	0.00

The results of the linear regression model shown in Exhibit 35 estimates the yearly medical expenses for individuals based on their age, number of dependents, body mass index (BMI), gender, smoking habits, region, and the interaction between BMI and smoking habits. The results of the regression model are as follows:

- The intercept (β0) is estimated to be -19399.74, which means that the expected value of yearly medical expenses for an individual with zero values for all the other independent variables is -$19,399.74. This value is not practically meaningful, and the interpretation of the intercept should be done with caution.
- The coefficient for age (β1) is estimated to be 264.01, which means that for every one-year increase in age, the expected value of yearly medical expenses increases by $264.01, holding all other independent variables constant.
- The coefficient for dependents (β2) is estimated to be 518.98, which means that for every one additional dependent on the policy, the expected value of yearly medical expenses increases by $518.98, holding all other independent variables constant.
- The coefficient for BMI (β3) is estimated to be 1266.62, which means that for every one-unit increase in BMI, the expected value of yearly medical expenses increases by $1,266.62, holding all other independent variables constant.
- The coefficient for gender (δ0) shows that the expected value of yearly medical expenses for males is estimated to be $533.04 lower than for females, holding all other independent variables constant.

- The coefficient for smoking habits (δ1) shows that the expected value of yearly medical expenses for smokers is estimated to be $20,632.79 higher than for non-smokers, holding all other independent variables constant.
- The coefficients for regions (δ2) show that there are no statistically significant differences in the expected value of yearly medical expenses between the Southeast and the Northwest regions, while the expected value of yearly medical expenses for individuals from the Northeast region is estimated to be $1,219.52 higher than for individuals from the Southwest region, holding all other independent variables constant.
- The coefficient for bmi30 (δ3) shows that the expected value of yearly medical expenses for individuals with a BMI over 30 is estimated to be $3,081.11 higher than for individuals with a BMI below 30, holding all other independent variables constant.
- The coefficient for the interaction between BMI and smoking habits (δ4) is estimated to be -1449.93, which means that the effect of smoking on yearly medical expenses is $1,449.93 lower for individuals with a BMI over 30 than for individuals with a BMI below 30, holding all other independent variables constant.

The p-values for all the coefficients are statistically significant (as their p-values are less than 0.05), indicating that the estimated coefficients are unlikely to be zero. Overall, the regression model suggests that age, number of dependents, BMI, smoking habits, and BMI-smoking habit interaction are important predictors of the expected value of yearly medical expenses for individuals, while gender and region have less impact on the expected value of yearly medical expenses.

While the data exploration analysis section of this report indicated that the BMI-smoking habit interaction would generate higher expenses for smokers with a BMI over 30, the negative effects of being a smoker and having a BMI over 30 are still being accounted for in the ‘smoker’ and ‘bmi30’ variables. As such, this interaction is still important to be included in the final model.

Next, I generated the results of the overall regression’s performance, including its R Squared value, its Adjusted R Squared value, its Akaike Information Criterion (AIC), and its Bayesian Information Criterion (BIC).

Exhibit 36: Results of Final Linear Model Regression

statistic	value
r.squared	0.85
adj.r.squared	0.85
AIC	26469.45

The results of the linear regression model shown in Exhibit 36 display the following statistics:

- The R-squared value of 0.85 indicates that 85% of the variance in the dependent variable (yearly medical expenses) can be explained by the independent variables (age, number of dependents, BMI, sex, smoking habits, region, and BMI-smoking habit interaction).
- The adjusted R-squared value is also 0.85, which means that this value is taking into account the number of independent variables and their contribution to the model.
- The AIC (Akaike Information Criterion) value of 26469.45 is a measure of the goodness of fit of the model, where a lower AIC value indicates a better fit. This value can be used to compare this model with other models to determine which one fits the data best. This will be used later in the analysis when comparing other models to this one.

Creating a Train-Test Split to Further Explain Results of the Model

Next, I created a train-test split to partition the data set into a training set and a testing set. This allowed me to train the linear model on the training data set to make predictions about the testing set of data.

I split 75% of the data set into training data and 25% into testing data. After doing so, I conducted a test for equal means on continuous variables and a test for equal proportions on categorical variables. Doing these tests would allow me to conclude whether the 75/25 train-test split would generate accurate results for the linear regression model later on.

First, I did a two-sample t-test, which a method used to test whether the means of two unknown samples are equal. That means that in this data set, I would be testing for whether the average value of each continuous variable in the train set is equal to the average value of that same variable in the test set. The t-test calculates a t-statistic, which measures the difference between the means of the two samples relative to their variation. If the calculated t-statistic is larger than the critical value of 1.96 or smaller than the critical value of -1.96, and the p-value is less than 0.05, then the null hypothesis stating that the there is no statistical difference between the means of the two samples is rejected in favor of the alternative hypothesis stating that there is a significant difference between the means of the two samples. In conducting this test, I was was expecting to fail to reject the null hypothesis, as I would expect that the means of the two samples for each variable would be close to being exactly the same. This would mean that the training and testing data contain data that is similar enough to yield accurate results when the model is trained and then evaluated.

The following were the results of the two-sample t-tests performed on the continuous variables:

Exhibit 37: Results of Test for Equal Means

Variable	T-Statistic	P-Value
age	0.07	0.95
dependants	0.90	0.37
bmi	0.81	0.42
expenses	1.37	0.17

From the results of Exhibit 37, I can see that all of the variables had a p-value above the significance level of 0.05 and a t-statistic below 1.96, meaning that there I failed to reject the null hypothesis. This means that the means of each variable are similar enough in the testing and training data sets, meaning that the accuracy of the results in my linear regression model will not be affected by the train-test split.

Next, I conducted a chi-squared test for equal proportions on categorical variables. This test examines the train and test data sets for equal proportions in categorical variables, with the null hypothesis being that the proportions are equal between the two samples. If the p-value that this test produces is above the significance level of 0.05, then the null hypothesis is rejected for the alternative hypothesis stating that there is a meaningful difference between the proportions of the categorical variable in the two samples. Like the two-sample t-test, I was expecting to fail to reject the null hypothesis, as it would mean that the accuracy of the results in my linear regression model would not be affected by the train-test split.

The following were the results of the chi-squared tests performed on each categorical variable:

Exhibit 38: Results of Test for Equal Proportions

Variable	Chi-squared Statistic	P-Value
sex	3.47	0.06
smoker	2.91	0.09
region	1.79	0.62
bmi30	1.29	0.26

From the results in Exhibit 36, I can see that I fail to reject the null hypothesis for every categorical variable, meaning that the accuracy of my results should not be affected by the train-test split, as there appears to be no meaningful difference proportions of each categorical variable.

After making sure that the accuracy of the results from my model would not be affected by the train-test split, I was able to generate predictions of medical expenses on the testing data. After doing so, I created a visualization which plotted the actual observations of the test set against the predictions made by the linear model. This visualization is displayed in Exhibit 38 below.

Exhibit 39: Predicted Expenses vs Observed Expenses

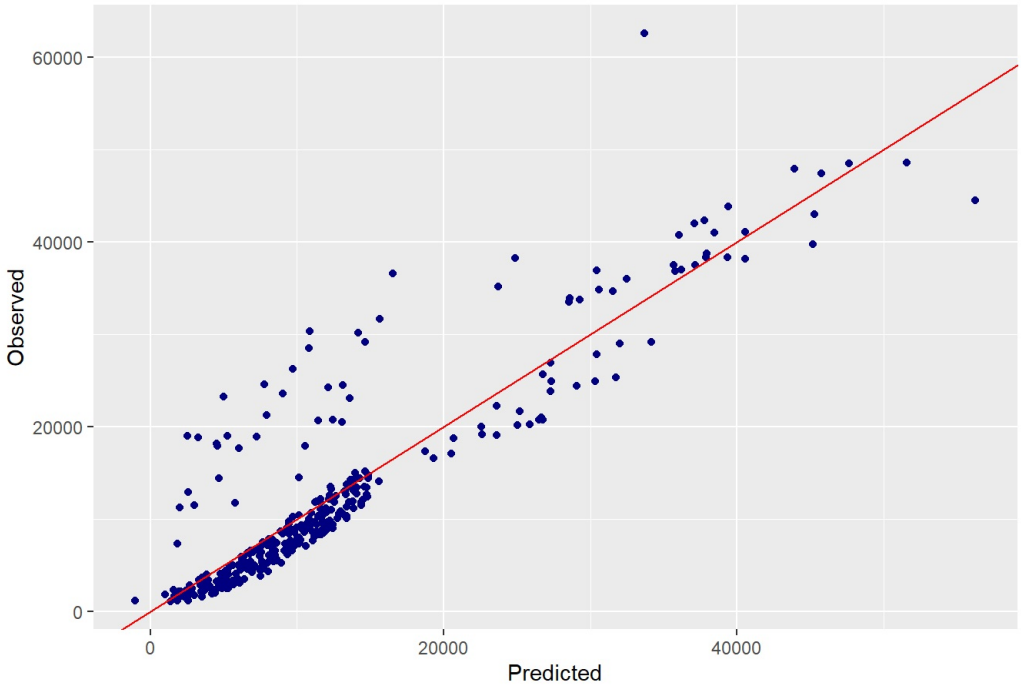


Exhibit 39 shows the relationship between observed (actual) values and the values predicted by the model. In this visualization, I would ideally like to have every observation be plotted on the red line going through y = x, signifying that the predicted value and actual value were equal. If observation is above the y=x line, it means that the actual (or observed) value was higher than the linear model predicted and if the observation is below that line, it means that the actual value was lower than the linear model prediction. Exhibit 39 shows that the majority of observations are very close to the y=x line, with only a few observations being significantly above or below this line. This makes sense, as it is difficult for the linear model to accurately predict cases where major medical attention was required when the characteristics of individuals indicate that it likely for that individual to have minor or moderate medical attention and expenses. This explains why the majority of observations which are significantly over the y=x line occur where there is a cluster of observations with minor medical expenses.

To put a quantitative figure on the accuracy of the predictions, I generated a Mean Absolute Percentage Error (MAPE) value which is used to measure the difference between actual and predicted values as a percentage of the actual values. Specifically, this statistic displays the average percentage error between actual and predicted values, with a lower MAPE indicating a more accurate forecast.

Exhibit 40: Mean Ablsolute Percentage Error Value of Final Model

	MAPE
MAPE	29.9

Exhibit 40 shows that the MAPE value obtained from this model is 29.8968%, which means that on average, the model's predictions for expenses are off by around 29.9% relative to the actual expenses. While this might look extremely high, suggesting that the model is not performing well in terms of predicting expenses accurately, this high value can be explained by way this value is calculated. Since the MAPE uses an average calculate to generate the final statistical value, the few observations which have unexpectedly high expenses relative to their predictions increase the average significantly.

With medical insurance, the unpredictability of medical expenses is likely the source of this high MAPE value. As such, it is likely that this is the lowest MAPE value that can possibly be obtained from this data. This will be tested in later on on this analysis against the MAPE value of other models that can be obtained in this data set.

Addressing Assumptions of the Model

Like previously mentioned, this model assumes that the relationship between independent and dependent variables is linear, that observations are independent of each other, that the variance of residuals is constant across all levels of independent variables (homoscedasticity), that the residuals are normally distributed, and that the model does not contain multicollinearity or influential outliers that unduly impact the results.

I took the following steps to address these assumptions:

1. Conducted a Ramsey RESET test to detect potential non-linearity and confirm that the relationship between the independent variable and the dependent variable is linear.
2. Conducted a Breusch-Pagan test to detect potential heteroscedasticity in the model and confirm that the variance of errors is constant accross all levels of independent variables
3. Conducted a Variance Inflation Factor test to detect potential multicollinearity.
4. Generated Cook's distance value to measure the influence of each observation on the regression results
5. Conducted a Shapiro-Wilk test to test for normality in the residuals.

Step 1: Ramsey RESET Test

The Ramsey RESET test is a statistical test used to detect nonlinearity in a regression model. The idea behind the Ramsey RESET test is to check whether the regression model fits the data well enough, or whether there is some remaining nonlinearity in the model that has not been accounted for. The test involves adding one or more squared or cubed terms of the independent variables to the regression equation, and then testing whether the new model has significantly better fit than the original model.

The results of the Ramsey RESET test conducted on the current model are displayed in Exhibit 41 below.

Exhibit 41: Results of Ramsey Reset Test

	RESET	P_value
RESET	2.58	0.11

The results from Exhibit 41 show the test statistic for the Ramsey RESET test is 2.58, which generates a p-value of 0.11. Since this p-value is greater than the significance level of 0.05, the results from this test suggests that there is not enough evidence to reject the null hypothesis that the current form of the model (which is linear) is correctly specified.

Step 2: Breusch-Pagan Test for Heteroscedasticity

The Breusch-Pagan test is a statistical test that can be used to formally test for homoscedasticity. The test involves adding a squared term of the predicted values (or other relevant independent variables) to the regression equation and testing whether this term is statistically significant. If the squared term is not significant, then the assumption of homoscedasticity is met. However, if the squared term is significant, then this suggests that there may be heteroscedasticity in the model.

The results of the Breusch-Pagan test conducted on this model will be displayed below in Exhibit 42.

Exhibit 42: Results of Breusch-Pagan Test

	BP_Statistic	P_value
BP	4.58	0.92

The results from Exhibit 42 show that the Breusch-Pagan test statistic is 4.58 with a respective p-value of 0.9175. This suggests that there is no significant evidence of heteroscedasticity in the linear model, thus confirming that the variance of errors is constant across all levels of independent variables.

Step 3: Variance Inflation Factor (VIF) Test for Multicollinearity

The VIF test is used to determine if the predictor variables in a regression model are too highly correlated with each other, which can lead to unstable and unreliable estimates of the regression coefficients. A high VIF indicates that the variance of the estimated regression coefficient is increased due to the presence of correlation among the predictor variables. Generally, a VIF value greater than 5 is considered to indicate high multicollinearity and can suggest that some of the predictor variables should be removed from the model.

The results from the Variance Inflation Factor (VIF) test are displayed below in Exhibit 43.

Exhibit 43: Results of Variance Inflation Factor (VIF) Test

variable	vif
age	1.02

dependants	1.00
bmi	3.00
sex	1.01
smoker	3.00
region	1.11
bmi30	2.79

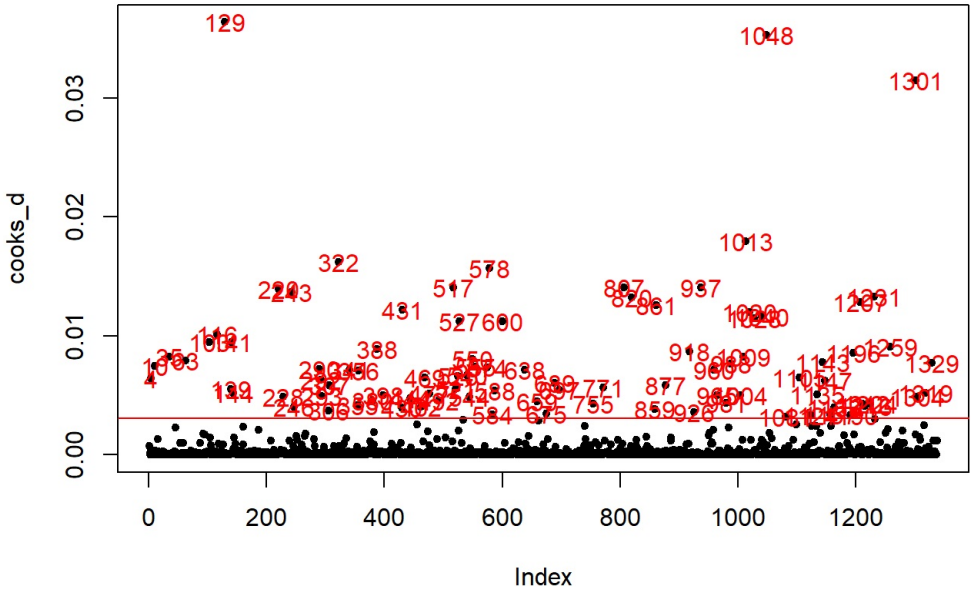
Exhibit 43 shows the results of the Variance Inflation Factor (VIF) test conducted on the final model. A VIF of 1 indicates no multicollinearity, while values above 5 or 10 are generally considered problematic. As such, there does not appear to be any issues regarding multicollinearity in the final model.

Step 4: Cook's Distance Value

Cook's distance is a measure of the influence of each observation on the regression results. It is calculated by comparing the regression coefficients with and without each observation. An observation with a high Cook's distance value indicates that it has a strong influence on the results and may be an influential outlier. A rule of thumb is that observations with a Cook's distance value that is greater than 4/n, where n is the sample size, may be considered an influential outlier.

To measure if this model contains a considerable amount of influential outliers, I will plot the cook's distance values and highlight the observations that exceed the Cook's distance rule of thumb in red.

Exhibit 44: Cook's distance plot



The results from Exhibit 44 show that there is a significant amount of influential outlier observations in this model, violating the assumption that the model does not contain influential outliers that unduly impact the results of the linear regression. This is not surprising at all, as the boxplot for the medical expenses variable in the exploratory analysis of this analysis showed that there were a significant amount of outliers in this data set, especially for major medical expenses. Like mentioned previously, this issue could be fixed by excluding values which are expected to be outliers. However, this would not be realistic for a start-up medical insurance company who is trying to generate the greatest amount of sales possible, and have as much clientele as it possibly can. Because a medical insurance is unlikely to stop providing medical insurance to so many people which are considered outliers by the Cook's distance plot, it does not make sense to remove these observations from the model. As stated previously in this analysis, this issue exemplies the factor of unpredictability that medical insurance companies have to deal with in order to stay competitive in their industry, and therefore it is a factor that must be dealt with by charging high premiums for insuring its customers. Because of this reality, I will maintain the outlier values in the data set as part of the linear regression model.

Step 5: Shapiro-Wilk Test for Residual Normality

The Shapiro-Wilk test is a statistical test that can be used to formally test for normality. The test calculates a test statistic called W, and the null hypothesis is that the data is normally distributed. If the test statistic's respective p-value is greater than the significance level of 0.05, then the assumption of normality is met. However, if the p-value is less than the significance level, then this suggests that the residuals may not be normally distributed. Because the assumption for the model is that the residuals are normally distributed, it is beneficial if the p-value is above 0.05.

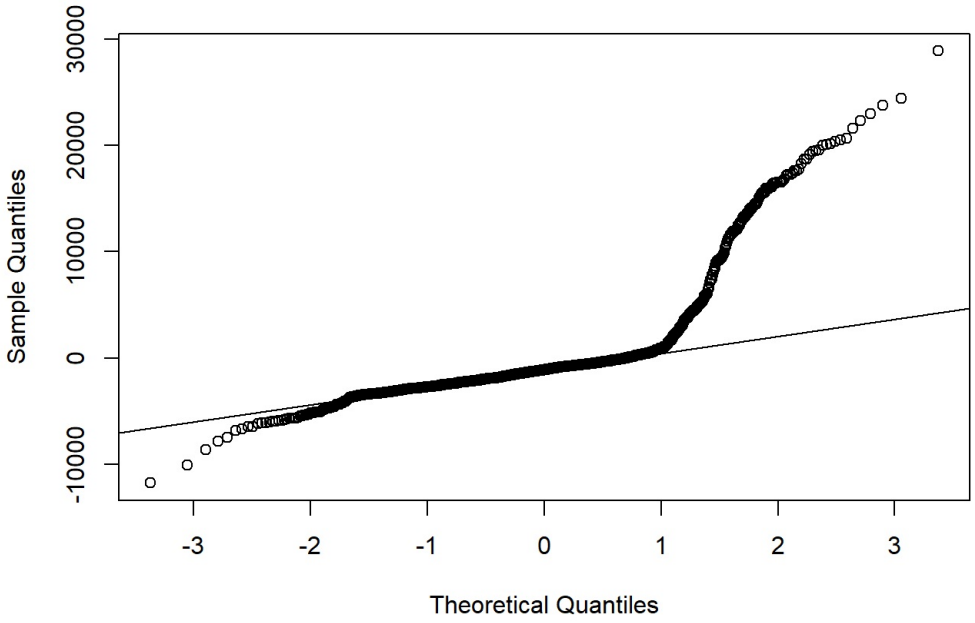
The results of the Shapiro-Wilk test for normality conducted on this model will be displayed below in Exhibit 43.

Exhibit 45: Results of Shapiro-Wilk Test

	W_Statistic	P_value
W	0.67	0

The results from Exhibit 45 show that the Shapiro-Wilk on the final model results in a test statistic of W = 0.67. Since the respective p-value for the test statistic is 2.2e-16 (basically zero), the null hypothesis of normality is rejected. This indicates that the residuals from the model are not normally distributed, which is not surprising considering the amount of influential outliers in observations where policyholders have major medical expenses. This is exemplified by the QQ plot of residuals in Exhibit 46 (displayed below), which compares the quantiles of residuals to the quantiles of a normal distribution.

Exhibit 46: Normal QQ Plot of Residuals



The results from Exhibit 46 show that the results are not normally distributed, as the points on the plot do not follow on the straight line that would represent a normal distribution. Since the results deviate significantly from the straight line, particularly at the higher quantiles, it suggests that the assumption of normality is not valid.

This can be explained by the the Cook's distance plot in Exhibit 45, which shows that there are an influential amount of outliers that are prohibiting the residuals to be normally distributed. Because I previously stated that I cannot remove outlier values without compromising the firm's ability to generate sales in this scenario. As such, this assumption cannot be met for this model

Directly Addressing Concerns From Assumptions

The assumptions of normality and non-influencial observations in the residuals are the only concerns for this model. However, like I mentioned in the analysis of these assumptions, there is only one solution for meeting these assumptions, which is to remove outlier values from the model. While this would be ideal, it is not realistic for me to remove these outlier values from this model, as this would be the equivalent of a start-up medical insurance company refusing to provide coverage to a significant portion of their potential clients. I believe that this is unrealistic because a start-up medical insurance company is unlikely to refuse to provide medical insurance to potential clients due to potential lost sales and therefore potential lost profits. While insuring its entire clientele might be more risky for the company, the medical insurance industry is filled with risk by nature, which is percisely the main driver for the amount of influential outliers in the model. As the start-up medical insurance company grows in both experience and clientele, it will be able to generate data with less outliers as a percentage of the overall number of observations, and therefore generate better predictions. As such, I believe that keeping the model as it is for a start-up medical insurance company is key for its success as the company matures.

Explaining Why This Model Generates the Best Possible Results

This final model with the estimating equation
$expenses = \beta_0 + \beta_1 age + \beta_2 dependants + \beta_3 bmi + \delta_0 sex + \delta_1 smoker + \delta_2 region + \delta_3 bmi30 + \delta_4 smoker * bmi30 + u$ generates the best possible results for predicting the value of medical expenses. This is because this model uses the most amount of logical descriptive attributes (or variables) to generate predictions about each individual's yearly medical expenses.

As stated earlier, this final model generates and R-squared of .85, an Aikaike Information Criterion (AIC) of 26,469, and a Mean Absolute Percentage Error Value (MAPE) of 29.90%. To confirm that this final model generated the best possible results, I compared these results to other simplified models. In doing so, I would be able to confirm that the final model had a higher R-squared value, a lower AIC, and a lower MAPE than simplified models.

I will generate three models to compare to the final model, the first of which has the regression equation of:
$expenses = \beta_0 + \beta_1 age + \beta_2 dependants + \beta_3 bmi + \delta_0 sex + \delta_1 smoker + \delta_2 region + u$

This model, which will be refereed to as Model 1, removes the indicator variable of bmi30 and the smoker-bmi30 interaction, as these two variables were generated through exploring the dataset.

The results from Model 1 are displayed below in Exhibit 47.

Exhibit 48: Results of Model 1

	R_squared	AIC	MAPE

0.75	27115.42	43.04

The results displayed by Exhibit 48 show that the final model has an R-squared which is 10% higher than Model 1, which indicates that the final model explains significantly more of the variation in the data than Model 1. The AIC of the final model is 26,469, while the AIC of Model 1 is 27,115. The lower AIC of the final model indicates that it is a better fit for the model since lower AIC values indicate a better model fit. Lastly, the MAPE of the final model is 29.9%, which is lower Model 1's MAPE of 43.03%, indicating that the final model has a better predictive accuracy. Therefore I concluded that the final model is a better fit for the data than Model 1.

Next, I will compare the final model to Model 2, which has a regression equation of: $expenses = \beta_0 + \beta_1 age + \beta_2 dependants + \beta_3 bmi + \delta_0 smoker + u$

Model 2 removes the sex and region variables from Model 1, as these variables had less predictive capacity according to p-values and coefficients generated in the regression results of the final model.

The results from generating a linear regression from Model 2 are displayed in Exhibit 49 below.

Exhibit 49: Results of Model 2

R_squared	AIC	MAPE
0.75	27113.95	43.33

Based on the results displayed by Exhibit 49, the final model has a higher R-squared value and a lower AIC value compared to Model 2, indicating that the final model is a better fit for the data. At the same time, the MAPE value of the final model of 29.90% is significantly lower than Model 2's 43.33%, indicating that the final model has a better predictive accuracy. Therefore, I concluded that the final model is a better fit for the data than Model 2.

Lastly, I will compare the final model to Model 3, which has a regression equation of: $expenses = \beta_0 + \beta_1 age + \beta_2 dependants + \beta_3 bmi + u$

Model 3 removes the sex and region variables and only displays continuous vairiables in efforts to simplify the linear regression.

The results from generating a linear regression from Model 3 are displayed in Exhibit 50 below.

Exhibit 50: Results of Model 3

R_squared	AIC	MAPE
0.12	28793.96	112.4

The results displayed by Exhibit 50 show that the final model has an R-squared which is 73% higher than Model 1, which indicates that the final model explains significantly more of the variation in the data than Model 3. This is likely due to the fact that smoker is explains a substantial portion of the variation in the data. The AIC of the final model is 26,469, while the AIC of Model 1 is 28,793. The lower AIC of the final model indicates that it is a better fit for the model since lower AIC values indicate a better model fit. Lastly, the MAPE of the final model is 29.9%, which is much lower Model 3's MAPE of 112.4%, indicating that the final model has a better predictive accuracy as Model two has forcasted values which are substantially smaller than actual values and therefore indicating that percentage errors were above 100%. Therefore I concluded that the final model is a substantially better fit for the data than Model 3.

The comparisons between the results of the final model and Models 1,2, and 3 show that the final model generates the best possible results in terms of goodness of fit and predictive accuracy when forecasting yearly medical expenses. While the final model still has its flaws, its results show that these are the best possible results to alternative models.

Acknowledging Weaknesses of the Model

In carrying out the model and conducting tests on its predictive accuracy and ability to meet assumptions, I was able to identify a few weaknesses in this specific model. First, the predictive accuracy of the model is not ideal, as the MAPE is almost 30%, meaning that the model's predictions are off by 30% from actual values. Of course, this figure is exacerbated by high outlier values which are very difficult for the model to predict accurately, therefore exaggerating the inaccuracy of the model. Also, the outlier values that the model has a difficult time predicting lead to heteroscedasticity, violating the model's assumption that the residuals are normally distributed. These are the two clear flaws exhibited in the model which cannot be combated through further manipulation of the data, as doing so would comprise the realistic application of the model in the real world.

While these weaknesses are important to acknowledge, it is also important to acknowledge that no model is perfect, especially in a risk-filled industry like medical insurance. The fact that this model explains roughly 85% of the variation in the data is actually very impressive, even if the 15% of the variation in the data can have a substantial impact on the predictive results of the model. This model is expected to give a baseline approach to estimating medical expenses, so its imperfections should be taken into account when applying it in real situations. While its imperfections can seem unattractive to decision makers, there should be no doubt that using the results of a model which is able to explain 85% of the variation in a data set is better than not using the model to predictions at all or using oversimplified models that explain less of the variation and have a far inferior prective accuracy.

Conclusion

For a start-up medical insurance company using available information to determine how much money it will have to pay out in claims every year by estimating yearly medical expenses, using the final linear regression model created in this analysis will generate the best possible results for the decision-makers of the firm.

The results from the linear regression model show the expected yearly medical expenses would be for an individual. In calculating this figure, the decision makers at the firm can calculate how much they can charge in premiums to that individual in order to earn a profit.

For example, if the company has an existing or prospective male client with the age of 37, 2 dependants which need to be covered, has a BMI of 28.9 (meaning that his BMI is under 30), is not a smoker, and is from the northwestern region of the United States, the company can almost instantly estimate how much they would expect to pay out in claims on a yearly basis for covering this individual based on their traits. In doing so, they can estimate how much they want to charge in premiums to that individual to ensure that the company is profitable. For example, if the company wants to charge premiums of that are 30% higher than the estimated medical expenses of the individual policyholder, they would obtain a 30% profit from covering that individual if the model has an exact prediction of the medical expenses for that policyholder. This is of course assuming that the policyholder accepts to be covered at that rate. However, even with this 30% figure, it is likely that some individuals will have significantly higher expenses than estimated by the model, which will lead to a loss in covering those individuals at the given rate. Overall, the majority of predictions using this model are likely to overestimate the medical expenses of policyholders as shown by Exhibit 39, which will lead to profitability in the majority of cases when using this method. This means that this linear regression model can be used as a tool for decision makers at a start-up medical insurance company to understand how what rate they should charge their current and prospective clients to ensure that the gains that they capture on the majority of cases where they generate substantial profits more than offset the losses that they take in on the few cases where predicting medical expenses is difficult and the company will take a hit on profitability.

I applied this theory to several examples, starting with an individual with the characteristics mentioned in the previous paragraph. In this application of theory, I predicted their expenses using their final model, calculated what the premiums the company would received if they charged different percentage rates, and compared those figures to the real yearly medical expenses incurred by that individual in the data set.

The results of this analysis are displayed in Exhibit 51 below, where “Estimated_Expenses” show what the linear model predicted the individual will have in yearly medical expenses, “Premiums_10” show what the company would charge an individual in order to obtain a 10% profit if they estimated yearly expenses exactly right, “Premiums_20” represents the figure charged to obtain a 20% profit from the estimated expenses, “Premiums_30” represents the figure charged to obtain a 30% profit from the estimated expenses, “Actual_Expenses” show the actual yearly medical expenses generated by the individual, and “Percentage_Difference” shows the percentage difference between the expected expenses and the actual expenses.

Exhibit 51: Results of Individual Policyholder Example 1

Estimated_Expenses	Premiums_10	Premiums_20	Premiums_30	Actual_Expenses	Percentage_Difference
7428.33	8171.16	8913.99	9656.82	6406.41	0.14

Exhibit 51 shows that the estimated expenses were 14% higher than the actual expenses. Therefore, if the decision makers at the start-up medical insurance company used a rate of 10% higher than the estimated expenses to charge the policyholder for insurance coverage, the profitability for the company would have been 24%, or roughly $1,765 in this case. When using rates of 20% and 30% higher than the estimated yearly expenses, the profitability would have been 34% and 44% respectively, making this a very profitable client.

I did two more examples like this to show that no case is the same, as some are more profitable than others, and it is the cumulative effect of this process that really matters.

In Example 2, I used the example of a policyholder with the following characteristics: Age = 56 Dependants = 2 Sex = female BMI = 39.8 bmi30 = over 30 Smoker = no Region = southeast

The results from conducting this same process on Example 2 are displayed in Exhibit 52 below:

Exhibit 52: Results of Individual Policyholder Example 2

Estimated_Expenses	Premiums_10	Premiums_20	Premiums_30	Actual_Expenses	Percentage_Difference
13012.11	14313.32	15614.53	16915.74	11090.72	0.15

Exhibit 52 shows that the estimated expenses were 15% higher than the actual expenses. Therefore, if the decision makers at the start-up medical insurance company used a rate of 10% higher than the estimated expenses to charge the policyholder for insurance coverage, the profitability for the company would have been 25%, or roughly $3,223 in this case. This is why I mentioned that the cumulative effect of this process is what is most impactful, as a 1% difference in profitability in Example 2 from Example 1 leads to roughly 1,500 dollars more in profit When using rates of 20% and 30% higher than the estimated yearly expenses, the profitability would have been 35% and 45% respectively, making this an incredibly profitable client for the firm.

Lastly, to show how incredibly unexpectedly high medical expenses can affect a firm, I created Example 3 which contained information about a policyholder with the following chacteristics:

Age = 28 Dependants = 1 Sex = male BMI = 36.4 bmi30 = Over 30 Smoker = yes Region = southwest

The results from conducting this final process on Example 3 are displayed in Exhibit 53 below:

Exhibit 53: Results of Individual Policyholder Example 3

Estimated_Expenses	Premiums_10	Premiums_20	Premiums_30	Actual_Expenses	Percentage_Difference
37164.54	40881	44597.45	48313.9	51194.56	-0.38

The results from Exhibit 53 show one of the cases where actual yearly medical expenses were much higher than predicted. In this case, the predicted medical expenses were 38% lower than the actual medical expenses. This means that even by charging 10% higher than the estimated expenses were for an individual, the company would have taken a 28% loss on covering this policyholder or the equivalent of 10,313.56 dollars. This figure is more than the profits of the previous two examples combined (specifically when they charge 10% higher than the estimated yearly medical expenses).

I conducted this analysis on these three examples to show how the model could be used to see how profitable some cases were relative to others, and how important the cumulative profits of all cases really is. The real take away from these examples is that the rate charged to individuals when using the model to make predictions is important for the profitability of the company. This means that the yearly medical expenses estimated by the model are the baseline that decision makers can use to make decisions about how much they should charge individuals given the profitability level they desire and the competitiveness of their coverage rates.

To conclude this analysis, I further strengthened the argument that the final model's "baseline" estimate can be incredibly informative for the decision makers by breaking down how this model generates predictions that overestimate actual values by a modest amount for the majority of cases and underestimates the actual values by a significant amount for a small portion of the total cases, resulting in a baseline estimate where the profits from the overestimating and the losses from underestimating cancel out. As such, the company can theoretically charge 10% above the estimated value for each policy and expect to generate a ten percent profit on all cases, even if it earns a small profit (say below 25%) on some cases and a large loss (say above 50%) in other cases.

This logic is explained below in Exhibit 54.

Exhibit 54: Breakdown of Baseline Estimate

No..of.Overestimated.Policies	1015.00
No..of.Underestimated.Policies	323.00
Avg..Gain.from.Overestimated.Policies	1854.66
Avg..Loss.from.Overestimated.Policies	-5828.11
Total.Profit.from.Overestimated.Policies	1882479.90
Total.Losses.from.Underestimated.Policies	-1882479.53
Losses.Incurred.from.Estimates	0.00
Profit.on.All.Cases.Charging.10..Above.Baseline	1775582.52

Exhibit 54 shows the process that the linear model uses in order to come up with a baseline estimate of yearly medical expenses that leads to $0 in losses for the company if it was to charge only the estimated expenses. Because 1,015 cases are overestimated by average values of 1,854.66 dollars, charging the estimated amount to a client leads to a profit of 1,882,749.90 dollars. However, the 323 cases where the model underestimates the actual values lead to losses of 5,282.11 on average, meaning that total losses of 1,882,749.53 dollars. This means that the profits and losses incurred from overestimating and underestimating respectively cancel out, and the estimate becomes accurate for all of the cases combined. Knowing this logic, a start-up medical insurance company could use the breakdown of this model to find a rate above the estimated expenses to charge its customers for insurance, like 10%, where the company could generate roughly 1.8 million dollars in profit and still offer competitive pricing.

In conclusion, this model is an effective tool for decision makers at a start-up medical insurance company working with minimal data to make predictions about how much they should charge a customer for yearly insurance coverage. Therefore, this could become a primary tool that the insurance company could use for estimating the benchmark for the profitability of the business over fiscal periods, and serve as a process that could be used to explore which individual characteristics are generating the highest claims (or losses) for the business. Overall, this final linear model is a tool that provides decision makers with a large amount of useful information that can be implemented into the operations of the business.