Medical Insurance Cost

Levon Haroutunian
Richa Jain
Jaivardhan Singh
Molly MacKinnon
Nihal Dhillon

1. Introduction

1.1. General Background

America has a unique and convoluted healthcare system compared to other developed nations. One key difference is the prevalence of private insurance companies that have relatively lax government oversight. Although the U.S. spends close to $4 trillion on healthcare costs [1], patients still face many challenges when trying to access care such as an inability to access preventative services and having to pay an abundant amount of money for basic treatments.

The dataset used in this analysis contains the medical expenses of an individual, along with other identifying factors. The individuals are adults ages 18 to 64, where subjects above 64 were not included as they are typically covered by the government. Along with an age and charges, the person's sex, BMI, region of residence, number of children/dependents covered in their plan, and smoking status is included. It is important to note that the charges describe the yearly medical costs accrued by the recipient and any dependents before insurance coverage.

Private insurance companies seek to maximize their profits by making more in premiums than it loses in covering patient medical charges. Finding relationships between what types of individuals are more likely to be costly to the insurance companies can be used to adjust insurance fees based on such factors.

1.2. Goals

The goal of our project is to use patient data to predict the average medical care expenses for such population segments among different ages, regions, and other factors. These estimates could be used to create actuarial tables that set the price of yearly premiums higher or lower according to the expected treatment costs.

We want to find relationships among various personal factors to answer a variety of different questions. We want to (1) determine how much weight each field has on individual charges and (2) find correlations between variables. Finding any relationship - or lack thereof - among these factors may provide valuable insights into which individuals are most costly and burdensome to the insurance company.

1.3. Discussion of Interesting Questions or Research Questions

We, as a group, came about many rather interesting problems for discussion. The small bout of

data we were given prompted many questions that we hope to find answers to throughout the duration of this project. The first question we came across is as follows: What is the direct weight and influence that the number of children as dependents cause on an individual's insurance plan? Is there a linear relationship between the number of dependents and overall charges of a plan, or is it much more complicated than this, potentially factoring in the other variables that have been provided?

Another complicated intricacy that we formulated is how at different stages in life an individual might be charged in relation to certain social standards for an individual at this point in life. An example of this is that in the U.S., it is typical for a woman to get pregnant around ages 25-30. Will insurance companies charge women around these ages more in anticipation of pregnancy-related medical costs?

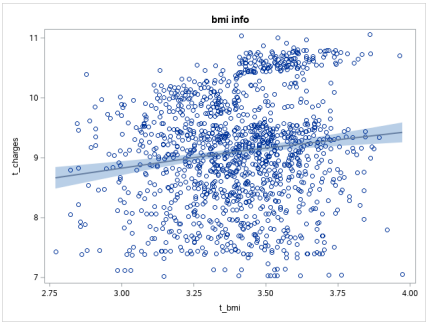
One more difficult question we came up with is related to the given BMIs of individuals. Not everyone with a remarkably low or high BMI necessarily has related health issues, and so is there a way to monitor the health problems directly related to BMI to prevent the extra charges from being applied to individuals who do not have BMI-related health issues? Or do insurance companies just charge across the board for having a drastically lower or higher BMI?
It is through the research and discussion of these rather interesting topics and questions that we, as a group, hope to come to some conclusions on how medical insurance costs in the U.S. are assessed, both for individual policyholders and for the specified groups of interest.

2. Exploratory Data Analysis

Our research began with looking at boxplots for the quantitative variables to look at their distributions. Age is a quantitative variable which was measured in years. BMI is a quantitative variable which is weight (kilograms) divided by square of height (meters). Children is a categorical variable measured by the number of children. Charges is a quantitative variable measured in dollars. Region is a categorical variable measured by 4 categories (southwest, southeast, northwest, northeast). Sex is a categorical variable measured by 2 categories (male and female). Smoker is a categorical variable measured by 2 categories (yes and no). The independent variables are age, BMI, sex, smoker, region and children. Charges is a dependent variable.

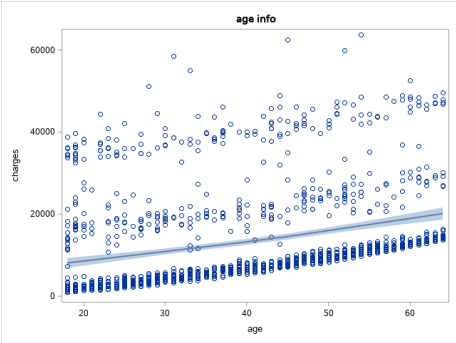
The boxplot for the log transformation of BMI showed a slight right skew. We also saw some potential outliers with high BMIs. The boxplot for age showed a symmetric distribution. There was a bit of a right skew. The boxplot for charges showed a very strong right skew. There were several outliers with very high charges, which could have been due to potential confounding variables.

We then observed the scatterplots between dependent and independent variables.



(Graph 1: LOESS Curve of log(charges) and log(BMI))

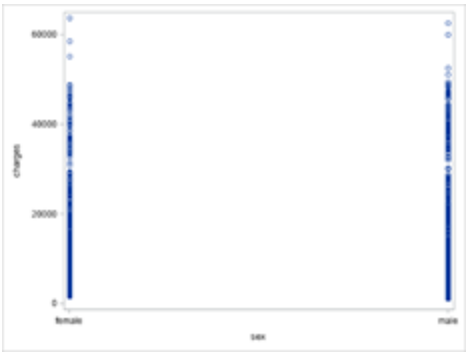
After performing a log transformation on both BMI and charges, which can be found in Graph 1, there is improvement in variance consistency, with the variation in charges being similar throughout the range of BMIs. The distinct subgroups in the data have largely disappeared, although there is still no clear linear pattern. The outliers do not seem as significant with the transformation.

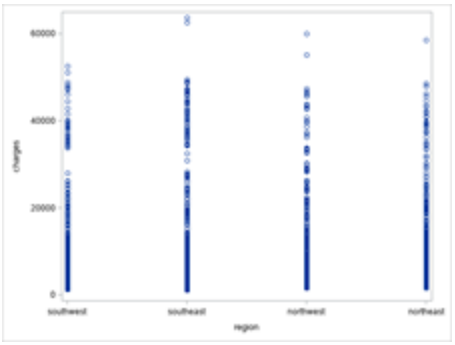


(Graph 2: LOESS Curve of log(charges) and age)

Fitting a linear regression model between age and charges, which can be found in Graph 2, displays an interesting pattern. We can see three separate linear patterns within this model. The data could further be separated on another potential categorical variable (with different levels). This is because it looks like there are three bands of data which each have a linear relationship with the same slope 1 but different intercepts 0.

We believe this pattern may be related to individuals who are healthy, in moderate health, and unhealthy with healthy having lower charges and unhealthy having higher charges. Using smoking status as a marker for health, elementary evidence of this idea is presented and discussed in further sections. There are some outliers with very high charges.

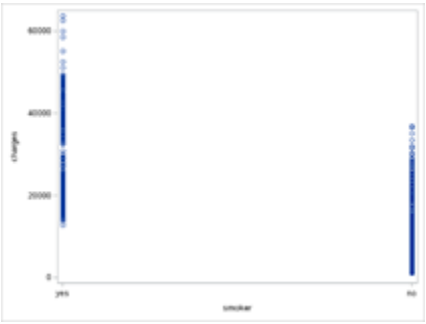


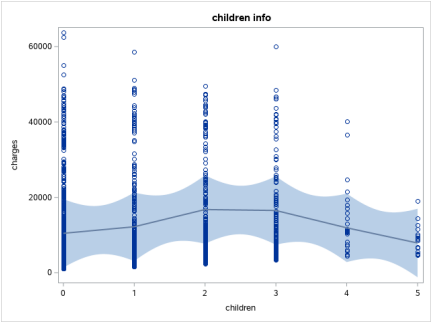


(Graph 3: Scatterplot of log(charges) and sex) (Graph 4: Scatterplot of log(charges) and region)

Looking at the relationship between sex and charges in Graph 3, we can see that both male and females have some outliers with very high charges; however, other than that, it seems like both females and males have similar charges. There is not much of a pattern. However, in the regression analysis later, it is discovered that males do generate, on average, higher charges than females.

Looking at the relationship between region and charges in Graph 4, we can see that each region, except southwest, has some potential outliers with high charges. Despite this, it seems as though all the regions have similar charges on average. Regions northeast and northwest display a constant scatter, whereas southwest and southeast have a "jump" - or lack of observed individuals - having around $3000 of charges. Across the nominal region categories, there is not much of a pattern.





(Graph 5: Scatterplot of log(charges) and smoker) (Graph 6: LOESS Curve of log(charges) and children)

Looking at the relationship between smoker and charges in Graph 5 we can see that those who smoke generally have higher charges than those who do not smoke. There are outliers in smokers with very high charges whereas non-smokers do not have many outliers. There is an overlap in charges where some smokers and nonsmokers have the same charge. This may be due to some confounding variable. Non-smokers have a constant scatter whereas smokers have a jump around $3000.

Looking at the relationship between children and charges in Graph 6 we see that there is a

negative relationship between the two variables which conflicts with the given LOESS curve. We can see some outliers where some individuals with children have higher charges. There are very few individuals who have more than three children. This LOESS curve indicates that charges increase around 2 and 3 children, almost in a bell or hill shape. This could be explained by the nonconstant variance in the data, and by the fact that there are far many more data points for 0-3 children than there are for 4-5 children. For almost every number of children, there is a jump around $3000.

3. Exploring Regression Relationship Using Simple Linear Regression Models

3.1. BMI and Medical Charges

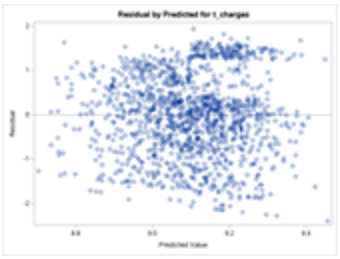
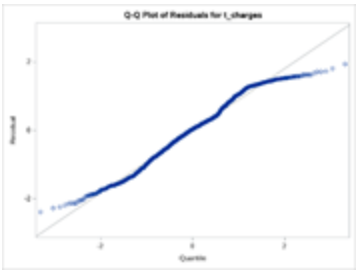
Body mass index (BMI) is a simple benchmark used in determining if an individual is overweight or obese. It is calculated by squaring the quotient of an individual's weight, in kilograms, over their height, in meters. The ideal BMI range is between 18.5 and 24.9, however, there are disputing arguments to the validity of using BMI as an indicator for healthy weight [2], [3].

Transformations on both BMI and charges were most effective in resolving the nonconstant variance in either variable. The table below shows a selection of the transformations attempted on the data.

bmi	*	log(charges)
sqrt(bmi)	*	log10(charges)
sqrt(bmi)	*	log2(charges)
sqrt(bmi)	*	log(charges)
sqrt(sqrt(bmi))	*	log(charges)
(bmi)**-1	*	log(charges)
log10(bmi)	*	log2(charges)

(Table 1: Transformation Attempts for BMI and charges)

Although the transformations above all produced very similar results, the natural log applied to BMI and charges was deemed as the most suitable modification to linearize and correct the data. The effect of this chosen transformation is shown and discussed below.

(Graph 7: Residual Plot of log(charges) and log(BMI)) (Graph 8: Q-Q Plot of log(charges) and log(BMI))

The residual plot against the log transformed charges in Graph 7 shows that the transformation made considerable improvements in removing any associations among the residuals and stabilizing its variance. The residual plot is centered around zero and scattered in a mostly random fashion. Compared to the raw data, this suits the assumptions of SLR much better.

The quantile-quantile plot (Q-Q plot) for the transformed data is represented in the graph to the right. The log transformation of BMI improves the distribution of the residuals, showing a more linear and symmetrical Q-Q plot. However, it is not completely ideal due the presence of short tails. Possible contributing factors may be related to the sample size of the data (n=1338) or other types of departures. While more investigation is needed, the new diagnostic plots provide evidence to deem the transformation successful enough to continue with a SLR.

The new fitted model, $\widehat{log(charges)} = 6.696732 + 0.62632(log(BMI))$, is much more appropriate. The p-value associated with t_bmi is <.0001, which therefore shows us that this log transformation of BMI is in fact significant.

Making improvements to the nonconstant variance, nonlinearity, and nonnormality of the observed residuals makes the model more robust. This is because nonconstant variance of the error terms causes less efficient estimates of beta0 and beta1 and produces an invalid estimate of sigma squared. The slope of the regression, 0.62632, represents the natural log of the average expected increase in medical charges per one unit increase in the natural log of BMI. In other words,β_1 shows the percent change in Y, which represents the BMI, while X, which represents the charges, increases by one percent.

3.2. Charges and Age

Age typically comes with health and financial changes, making it worth considering as an explanatory variable for predicting medical charges.

Constructed from our data, the fitted linear regression model is $\widehat{log(charges)} = 3165.88501 + 257.72262(age)$. The estimated intercept does not provide useful information about charges from a newborn because the sample only considers adults ages 18 to 64. The p-value associated with age is <.0001, which therefore shows us that this parameter of age is in fact significant.

As mentioned in earlier sections, there are three distinguishable subgroups all following the same slope but starting at different intercepts. The fitted model fails to capture the observations in the uppermost group, as most of those data points fall outside of the 95% prediction interval.The validity of a SLR is challenged by the scatterplot of residuals against Ŷ.

(Graph 9: Residual Plot of log(charges) and age)

Graph 9 also represents this pattern that is seen in the original data, where there are three different subgroups that follow the same slope given in the regression analysis. This further supports the interesting linear relationship that is presented in the analysis of age versus charges. There are a few outliers with very high charges that do not fall into any of these three subgroups, and so this could be a deviation from our assumptions of the model.

Although the fitted model does not appear to capture all the data, we intend to further investigate the three groups that make up the plot. Perhaps there lies strong relationships between such subgroups and charges, however, the presence of such groups is obstructed by transformations.

3.3. Medical Charges and Categorical Variables

3.3.1. Medical Charges and Sex

In the dataset, sex is described as a binary of male or female. The following regression is generated using the indicator variable "isMale" having two possible values, 0 and 1. An isMale value of 0 represents female individuals and a value of 1 are male individuals.

We have that $\widehat{log(charges)} = 12570 + 1387.17233(isMale)$, given $X = 0$ for a female and $X = 1$ for a male. b_0 represents the average expected cost for females, whereas b_1 represents that, on average, males are charged $1387.17 more than females. The p-value associated with isMale is <.0361, which therefore shows us that this binary for sex might not be significant significant.

The intercept represents the reference group which is age=female. The estimated intercept is the expected charges for the female group. The isMale parameter represents the expected difference in expected charges going from the female to male group. Here, the p-value is statistically

significant for the alpha test of 0.05 so this may be a good categorical variable to include in our final regression mode.

3.3.2. Medical Charges and Region

Information from individuals in the dataset also included the region of the US in which they lived, described as northeast, northwest, southeast, or southwest. Across such broad geographic areas, there are known political, economic, social, and other differences which may or may not be apparent in medical charges. The following 4 SLR models seek to observe any possible associations between living in any one of the 4 regions or living outside that given region.

Our fitted model here is $\widehat{log(charges)} = 12347 + 1059.44717(NE) + 70.63800(NW) + 2388.47406(SE)$. The intercept β_0 represents the effect of the reference group southwest (SW) on expected charges. The parameter estimates for each region (NE, NW, SE) are the changes in expected charges for each region compared to the reference region SW. The p-values are relatively high for NE and NW indicating they do not have a statistically significant effect on the expected charges when going from the reference region (SW) to those regions. However, the region SE appears to have a p-value of 0.0097 which can be considered statistically significant based on alpha=0.05, and therefore there is a statistically significant effect on going from region SW to SE regarding the expected charges.

3.3.3. Medical Charges and Smoker

Smoking is a known risk factor for many chronic and life-threatening diseases. The harm smoking causes to human health is also reflected in the cost of care associated with managing such health implications [4]. The model below seeks to quantify and predict the average medical costs using smoking as a predictive independent variable.

The fitted model is $\widehat{log(charges)} = 8434.26830 + 23616(yes\_smoker)$, given $X = 0$ for a non-smoker and $X = 1$ for a smoker. b_0 represents the average expected cost for non-smokers, whereas b_1 represents that, on average, smokers are charged $23616 more than non-smoker.

The p-value for the effect of being a smoker compared to a non-smoker is <.0001. Therefore, this indicates that this categorical variable is a good use in our final regression model due to its statistical significance.

3.3.4. Medical Chargers and Children

The charges sent towards the insurance company not only stems from the individual patient, but

also any dependents covered in their plan. The sampled individuals had between zero and five children. It is natural to assume that the more children in a covered plan, the more charges would be accrued by the plan. The regression below uses evidence from the data to investigate this idea.

The fitted model is as follows:

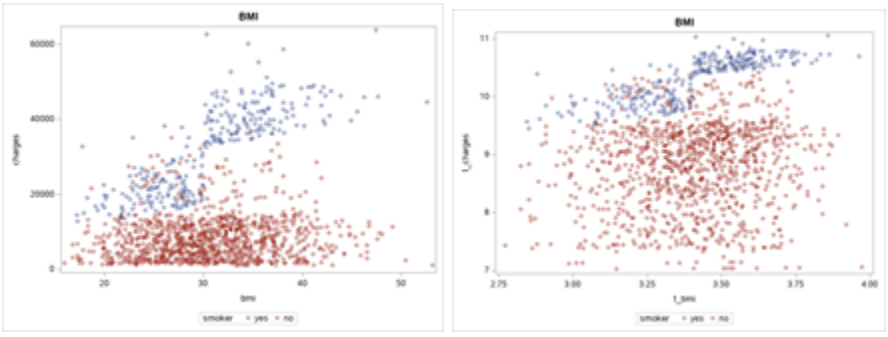
$$\widehat{log(charges)} = 12366 - 365.19623(children\_1) + 2707.58813(chidren\_2) + 2989.34277(children\ 3) + 1484.68071(children\_4) - 3579.94035(children\_5),$$

where $X = 0$ is having no children, and this same logic holds up until having 5 children. The p-values associated with three of the dependent parameters, those being children 1, 4-5, are large numbers, which therefore shows us that this parameter of children might not be significant. Children 2-3 could possibly be significant, as their p-values are smaller numbers, .0035 and .0060 respectively. However, we will come to the conclusion later on that the parameter of children does not have any significance in this model.

4. Multiple Linear Regression Analysis

4.1. Grouped Scatter Plots - Newfound Associations with Grouped Variables

When grouped with certain categorical variables, many clear relationships can be seen when distinguished over the quantitative independent variables age and BMI. Scatterplots that were previously messy and hard to interpret suddenly showed very interesting relationships. The results from the following groupings open many new questions and areas to explore that may potentially expand our understanding of the nuances around medical charges and the broader strategies of health insurance financings.

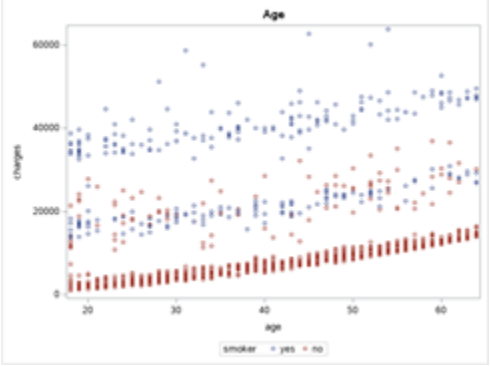


(Graph 10: Scatterplot of charges and BMI Colored by smoker) (Graph 11: Scatterplot of log(charges) and log(BMI) Colored by smoker)

A clear example of such patterns can be seen when the binary smoker category is grouped within the scatterplots of BMI and charges; both untransformed (Graph 10) and both transformed (Graph 11). Previously mentioned in the model using untransformed BMI, there were two vaguely identifiable trends seen within the scatterplot; one staying flat and close to zero, and the other in a somewhat-straight line that increases in a positive direction. When smokers (colored in blue) and non-smokers (colored in red) are identified, these trends strikingly stand out.

From the data, it can be said that non-smokers tend to generate consistent charges - mostly between $0 and $20000 - *regardless of their BMI*. Considering only non-smokers, there is little graphical evidence to support a linear relationship between BMI and charges. The same cannot be said for smokers. On the contrary, smokers with higher BMIs tend to generate higher charges than smokers with lower BMIs. As BMI increases among smokers, charges are seen to increase as well. Looking at the transformed scatterplot, the positive association between smokers' BMI and charges is remarkably consistent as well: the variance among smokers is tight, further supporting the relationship. On the same plot, the variance among non-smokers is high and randomly scattered, supporting the idea that BMI is not an important indicator of charges for non-smokers.

Data regarding the variables of age and smokers can also be investigated, as when put into a comparison by scatterplot, there is an alarmingly distinct relationship between these two variables that assist us in making further assumptions about our model.



(Graph 12: Scatterplot of log(charges) and age Colored by smoker)

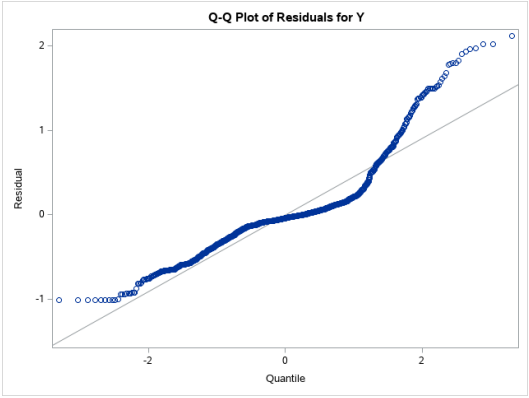
As given in Graph 12, there is a clear correlation between the variables of age and smokers, which follows our interesting discovery of the three subcategories within age displayed within the scatterplot. The increase in charges is consistent with the increase in age, regardless of smoking status. However, those of low age, who have relatively lower charges than those of higher age, among smokers are charged, on average, much higher than nonsmokers of this same age. There is an overlap in the charges of smokers and nonsmokers that lies within the middle of the data, where the highest charged young nonsmokers begin to match the charges of the lowest charged middle-aged smokers. This further supports our findings that, when comparing age and smoker status together against charges, that there is a clear relationship between smoker status and charges on a plan.

4.2. Base Model

With the discovery of these newfound relationships between some of our quantitative and qualitative variables through the grouped scatter plots above, we are now able to begin to formulate our multiple linear regression models, and this begins with the construction of our base model.

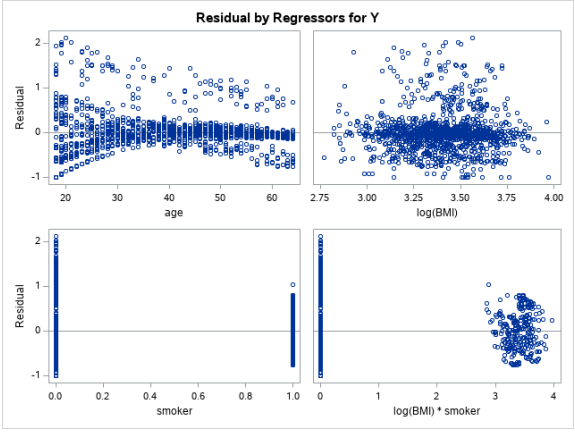
Our base model that we chose is represented by $Y = b_0 + b_1(age) + b_2(log(BMI)) + b_3(smoker) + b_4(log(BMI) * smoker) + \varepsilon$. We chose this model in an attempt to further follow our instincts that log(BMI) and smoker are heavily related, which is one of the correlations that we found in our grouped scatter plots. Thus, we created the new variable of log(BMI)*smoker, which represents the interaction of these two variables.

From our data, we cannot conclude any results, as our assumptions of an MLR model are not met, as we can observe through our findings below.



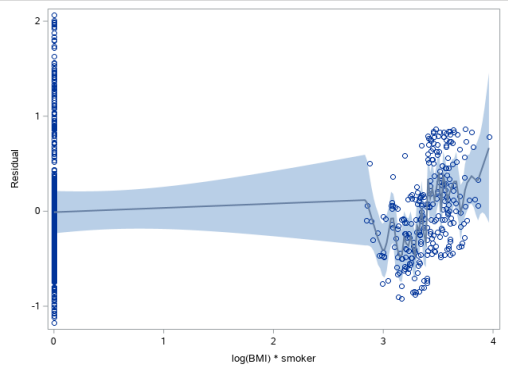
(Graph 13: Q-Q Plot of the Base Model)

As we can see from the Q-Q plot displayed in Graph 13 is not a straight line, therefore we can conclude that the MLR is not normally distributed and the assumptions for MLR are not met.



(Graph 14: Individual Residual Plots of the Base Model)

There are clear patterns for age and log(BMI) in Graph 14. They are not symmetrically distributed as age has much more points on the positive side compared to the negative side and log(BMI) is heavily concentrated in the middle. The others are insignificant as they are qualitative predictors.



(Graph 15: Residual Plot of log(charges) and log(BMI)*smoker)

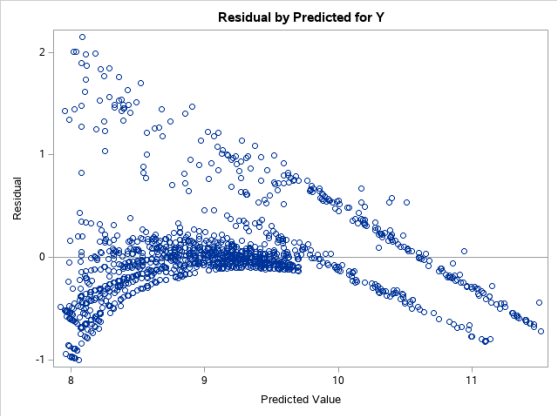
The residual of log(BMI)*smoker represented in Graph 15 shows us that our interaction variable between log(BMI) and smoker clearly has some sort of relationship that needs to be investigated, as this residual violates our assumptions of the regression model very distinctly.

4.3. Exploratory Models

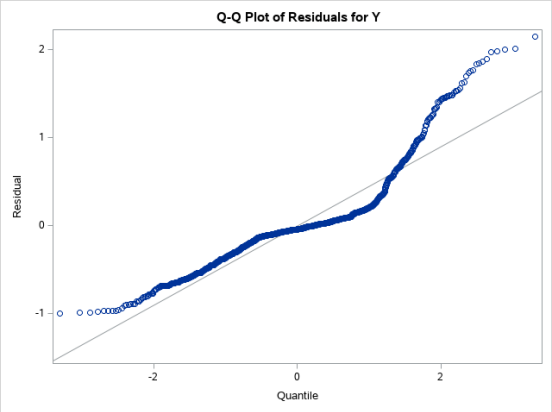
Model Numbers	Model
2	Base + log(BMI)*sex, sex
3	Base + log(BMI)*region, region
4	Base + log(BMI)*children, children
6	Base + age*sex, sex
7	Base + age*region, region
8	Base + age*children, children

(Table 2: Attempted Exploratory Models)

All of the models attempted in Table 2 did not result in any significant conclusions. All of the residual plots and Q-Q plots for each model had the same or similar patterns compared to each other. Graph 16 represents what each model's residuals look like and Graph 17 represents what each model's Q-Q plots look like.. None of the assumptions of the models were met, so we cannot draw any conclusions from them.



(Graph 16: Residual Plot for Models 2-4, 6-8)

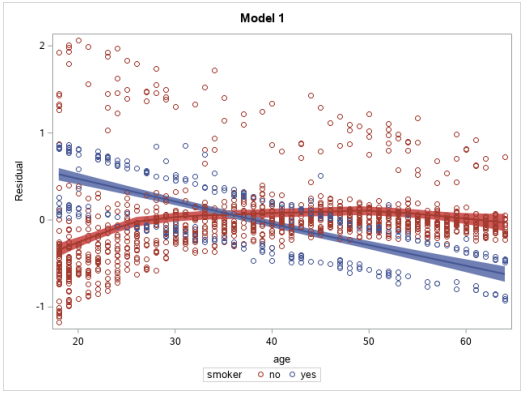


(Graph 17: Q-Q Plot for Models 2-4. 6-8)

4.4. Model 1

Model 1 is represented by
$Y = b_0 + b_1(age) + b_2(log(BMI)) + b_3(smoker) + b_4(log(BMI) * smoker + b_5(sex) + b_6(region) + b_7(children) + \varepsilon$. We want to test if sex, region, and children are significant. Note that this Model 1 is just the Base Model with the additional parameters of sex, region, and children.

From our data, we cannot conclude any results, as our assumptions of an MLR model are not met, as we can observe through our findings below.

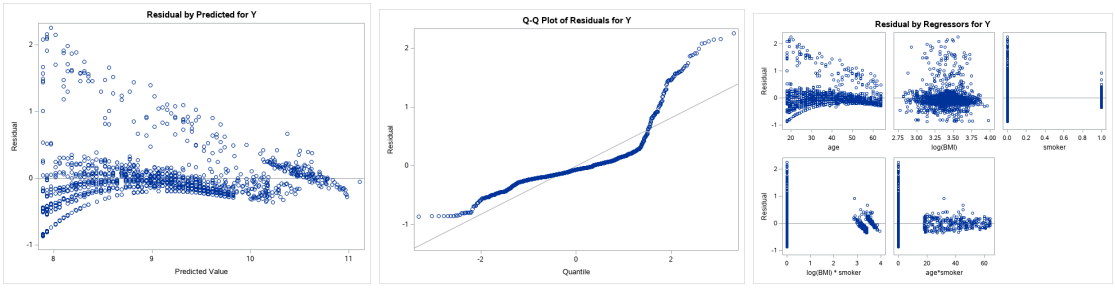


(Graph 18: Residual Plot of log(charges) and age Colored by smoker)

Referring back to our preliminary findings in 4.1, we know that there is some sort of relationship between smoker and age, and so we decided to color the residuals of age and log(charges) against smoker status in order to try and differentiate some new patterns, which is shown in Graph 18. Using Model 1 gives us the most extensive set of data, as there are the most regression coefficients included, and so using this Model we are able to see that there is some sort of pattern regarding age and smoker, and so this shall be noted and invested further, specifically in Model 5.

4.5. Model 5

Model 5 is represented by
$Y = b_0 + b_1(age) + b_2(log(BMI)) + b_3(smoker) + b_4(log(BMI) * smoker) + b_5(age * smoker) + \varepsilon$
, which can be viewed alternatively as the Base Model with the addition of the interaction effect between age and smoker. Looking at Graph 19 we can see that there is some variance in the residuals, compared to previous models, as there is a little more of a spread. When we also look at Graph 20, we can also see a little more deviation around quantile = -2, and it looks less normally distributed. However, when we looked at Graph 21, we see that age*smoker is more spread out around 0 towards yes_smoker. Log(BMI)*smoker presented a new pattern around yes _smoker and there are two parallel residual patterns. This was an interesting relationship that we chose to explore further.

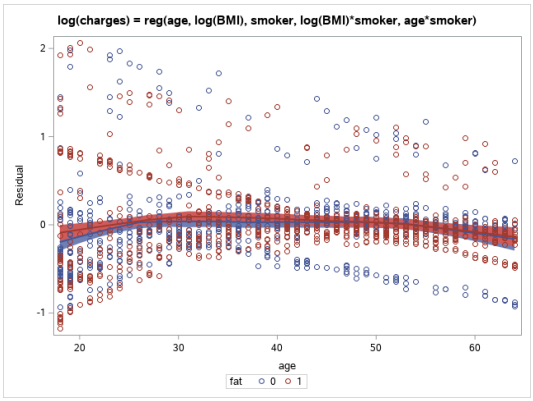


(Graph 19: Residual Plot of Model 5) (Graph 20: Q-Q Plot of Model 5) (Graph 21: Individual Residual Plots of Model 5)

4.6. Model 5.5

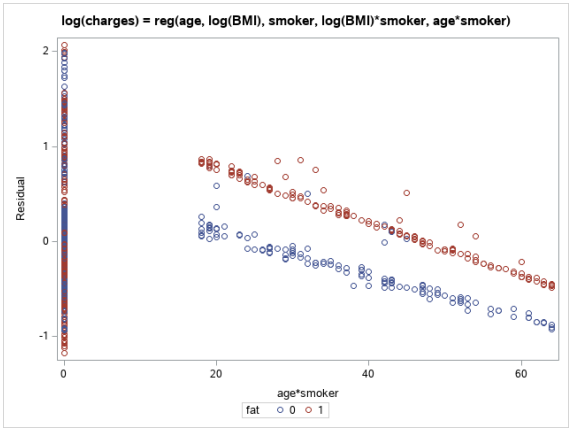
Since we have noted that Model 5 deviates the most from the other exploratory models, we choose this model to investigate our data. Noting that there is a new interesting pattern in the scatterplot between log(BMI)*smoker in Model 5, we need to find a new way to dissect this pattern. One way that we formulated was to create a new binary for log(BMI) that involves the mean for log(BMI), where 1 represents when log(BMI) is above or equal to this mean, and 0 represents when log(BMI) is below this mean. We named this new binary variable “fat”, and this mean is 3.40295.

Using this new fat variable in coloring some of our Model 5 graphs, there is much to be noted regarding the fact that this fat explained the model’s variation almost perfectly. The residual plot of age against log(charges) colored by fat, which is Graph 20, explains Graph 18 in a much better light using this new variable of fat. The residuals are much more focused around the 0 axis and there is less of a pattern than in Graph 18.



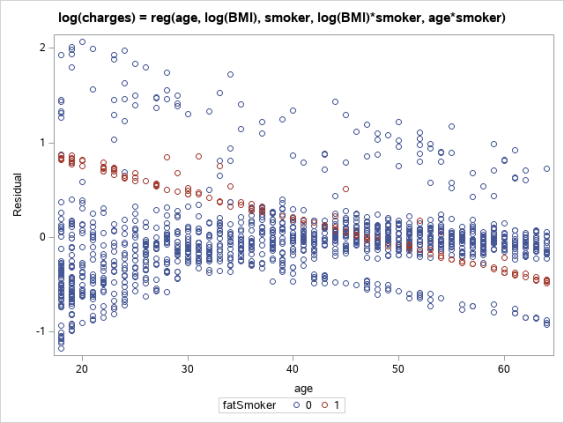
(Graph 22: Residual Plot of Model 5 Colored by fat)

Next, we colored the residual plot of age*smoker and log(charges) with the fat variable in Graph 22 and this showed a significant pattern. This new pattern displays the same pattern that the residual of log(BMI)*smoker and log(charges) from Graph 21 does, but in a much clearer manner, and it clearly displays that the fat variable was responsible for causing this pattern. Those with a fat level below the average log(BMI) consistently are grouped together in terms of charges and age when they are an active smoker, and the same can be said for those with a fat level above the average log(BMI).



(Graph 23: Residual Plot of log(charges) and age*smoker Colored by fat)

Going even further, we made a new variable called fatSmoker, which is just our binary fat variable multiplied by our binary smoker variable. Graph 24 displays the residual plot of age and log(charges) colored by this new fatSmoker variable, and the results are telling. This new graph, when comparing it to Graphs 18 and 22, displays that these patterns that we have seen in these two graphs can be explained even further by fatSmoker. When fatSmoker is equal to 1, the residuals for age stay grouped in a downward sloping linear pattern. On the other hand, when fatSmoker is equal to 0, the residuals for age are much more scattered and random, however there is a slight upward pattern hovering around zero. This new residual is much more randomly scattered and follows the assumptions of a multiple linear regression model. Thus, Model 5.5 is our most accurate model and we would like to explore this model further.



(Graph 24: Residual Plot of log(charges) and age Colored by fatSmoker)

4.7. Continuation of Model 5.5

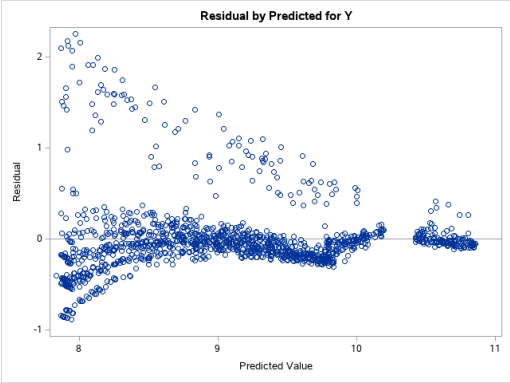
With our new fat and fatSmoker variables, the question now becomes: How can we fit Model 5.5 to accommodate these new findings?
We have Model 5.5, which we know is represented as
$Y = b_0 + b_1(age) + b_2(log(BMI)) + b_3(smoker) + b_4(log(BMI) * smoker) + b_5(age * smoker) + \varepsilon$
, which is Model 5, plus this new factor of fat. We need to find a way to incorporate fat or fatSmoker into Model 5 in order to form our final Model 5.5.
There are a few ways in which this is attempted.

Model Numbers	Model
5.5.1	Model 5 + fat
5.5.2	Age + log(BMI) + smoker + fat + fatSmoker + age*smoker
5.5.3	Model 5 + fat + fatSmoker
5.5.4	Age + log(BMI) + smoker + fatSmoker + age*smoker

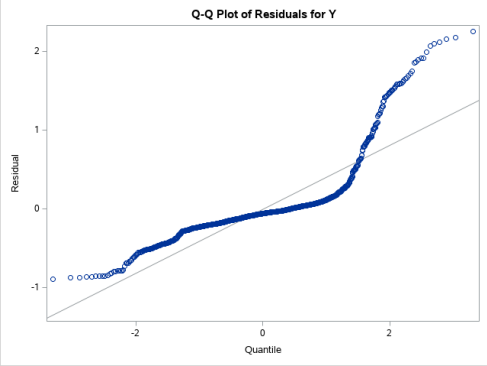
(Table 3: Attempted Continuations of Model 5)

After attempting the models in Table 3, we came to the conclusion that Model 5.5.2 is the most adequate model that we can draw conclusions from. Model 5.5.4 is also adequate, as it also has low VIF values, however, it is the same model as 5.5.2 just without the fat variable, so we chose to look more at Model 5.5.2 because it has the added benefit of an extra variable. The VIF values are all below 10, and so they demonstrate no multicollinearity between the different predictors that we looked at. However, this is not a strong enough indication that the MLR assumptions are

met. The residual plots for Model 5.5.2 are still displaying patterns, almost more than the original Model 5 did. This can be seen in Graphs 25 and 26.



(Graph 25: Residual Plot of Model 5.5.2)



(Graph 26: Q-Q Plot of Model 5.5.2)

Thus, our conclusion is that although we attempted numerous models, it was difficult to find one particular model that perfectly framed our data to fit an MLR. Therefore, further exploration of this data within Model 5.5 is necessary to find a significant model to draw conclusions.

5. Goals and Conclusions

5.1. Determine how much weight each field has on individual charges.

We found through our data analysis that age, log(BMI), and smoker were significant in determining medical cost charges. Region, children, and sex were insignicant in doing so.

5.2. Find correlations between variables.

Age and smoker in relation to log(charges) were correlated. Log(BMI) and smoker in relation to log(charges) were also correlated. This was demonstrated through interaction effects during our model analysis.

5.3. Direct weight and influence that the number of children as dependents cause on an individual's insurance plan.

We found that the number of children was insignificant in relation to medical charges so we are unable to draw any conclusions about its influence on insurance plans.

5.4. Is there a linear relationship between number of dependents and overall charge?

Based on a LOESS curve from earlier in the paper, there was no linear relationship between dependents and overall charges. Additionally, when we looked at p-values we found no significance between number of dependents and log(charges).

5.5. Do charges change at different age stages in life due to certain social standards?

Early in our analysis, we observed an interesting relationship in a scatter plot between age and log(charges). We found three different linear patterns which we hypothesize may have something to do with different age groups such as “young,” “middle aged,” and “old.” This is something that would require further analysis.

6. Individual Part

6.1. Limitations of the Data and the Models/Statistical Techniques

When we first received this dataset, we wanted to approach the data from a social angle and understand the medical-industrial complex and how it impacts certain individuals. One of our biggest goals was to understand how insurance charges change at different age stages in life. For example, do women get charged higher when they are at the age of most likely getting pregnant? We were able to produce a scatter plot of log(charges) vs. age. We hypothesized that the three linear patterns shown in the graph had something to do with different age groups; however, we did not have the statistical techniques to investigate the patterns further. In reality, insurance companies are not allowed to increase your premium due to sex or health condition (including pregnancy), so it would have been interesting to prove or disprove this using statistical models. We attempted multiple different MLR models to try to address questions such as this one, but came out shorthanded each time. No model is perfect and we did find some interesting models, but nothing answered our questions. It is possible that in order to address these questions, we still have to learn more statistical skills. It is also likely that the data does not allow for a perfect or ideal model since it is real data and we cannot shape it to perfectly fit our needs.

6.2. Improvements I Would Make

I would have liked to interrogate the findings of the scatterplot mentioned above more. Like I said, insurance companies are not allowed to increase premiums due to sex or health condition. I would have liked to prove or disprove this idea with the data; however, we did not have enough data or statistical skills to go more in depth. The models we created told us nothing significant to answer this question. I would have liked to create more models revolving age, sex, and charges to see if we missed something from our earlier attempts. Perhaps, it would have been worth searching online for more related data so we could create a better model.

I also think we did not stick to our goals. We lost sight of what we had set out to do at the beginning of the project and started working on many different models just to see what we could find and just to find one working model. I think this was a fun approach because we had the opportunity to try different things, but I think if we had structured our project around our goals we would have possibly had a better outcome.

6.3. What Did I Learn?

This project taught me that data is messy. Especially real world data. No data is perfect and it will never perfectly fit the model or idea that you are trying to work on because there will always be limitations. This means that there will also never be a perfect answer. We were unable to properly answer any of our original questions, which is okay. In my opinion, the purpose of this data analysis was to teach us that sometimes we start out with one idea and at the end the project just ends up being something completely different.

This project also taught me how to work with a group. Working with a group has benefits and drawbacks but it taught me how to compromise and communicate in order to come up with an end product. I am planning on working for an insurance company once I graduate and I know that there is a lot of group work and collaboration and I feel much more comfortable for that after this project.

Finally, I was able to learn a new technical skill. I had worked in SAS a little during an internship, but I had a GUI doing most of the work for me. This time, I was actually able to learn how to code and apply statistical analysis through SAS. I know my post-graduate job requires knowledge of SAS and I am glad that I will be able to point to this project to show my experience in it.

7. Peer Assessment

7.1. Molly

Molly conducted much of the preliminary SLR investigations through SAS and interpreted them. Molly also did a lot of the interpretations of the MLR models while assisting Levon with the SAS code. Molly wrote a good amount of each of the reports and investigated our final models including the continuation of Model 5.5 and our conclusions. Molly also helped create the powerpoint for our presentation.

7.2. Richa

Richa conducted much of the preliminary SLR investigations through SAS and interpreted them. Richa also did a lot of the interpretations of the MLR models while assisting Levon with the SAS code. Richa wrote a good amount of each of the reports and helped determine conclusions.

7.3. Levon

Levon handled the SAS code, especially for the MLRs. He constructed the MLR models and contributed to the SLR investigations. He wrote a good amount of each report, but he was very much involved with the SAS code.

7.4. Nihal

Nihal investigated MLR model 1 and 1.5 on her own. She missed opportunities to participate, but became involved in the MLR models and read over the report to catch up.

7.5. Jai

Jai missed opportunities to participate in the project, but became involved towards the end a little. He assisted with some of the SLR models early on and with the Base MLR model towards the end.

8. Resources

8.1. [1] https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical#:~:text=U.S.%20health%20care%20spending%20grew,spending%20accounted%20for%2017.7%20percent.

8.2. [2]: Kok P, Seidell JC, Meinders AE. De waarde en de beperkingen van de 'body mass index' (BMI) voor het bepalen van het gezondheidsrisico van overgewicht en obesitas [The value and limitations of the body mass index (BMI) in the assessment of the health risks of overweight and obesity]. Ned Tijdschr Geneeskd. 2004 Nov 27;148(48):2379-82. Dutch. PMID: 15615272.

8.3. [3]: Frankenfield DC, Rowe WA, Cooney RN, Smith JS, Becker D. Limits of body mass index to detect obesity and predict body composition. Nutrition. 2001 Jan;17(1):26-30. doi: 10.1016/s0899-9007(00)00471-8. PMID: 11165884.

8.4. [4]: Hall, Wayne, and Chris Doran. “How Much Can the USA Reduce Health Care Costs by Reducing Smoking?.” *PLoS medicine* vol. 13,5 e1002021. 10 May. 2016, doi:10.1371/journal.pmed.1002021