Factors Affecting Health Insurance Premiums

Data 603 Final Project, L01 Group 1

Bobbi Boyce, Stephen Bonsu, Kennedy Gunderson, Maxwell Paterson, Noah Seminoff

2024-11-09

Contents

Chapter 1

Introduction

1.1 Motivation

1.1.1 Context

A multitude of different factors are used to assess health insurance costs. These calculations generally depend on various characteristics such as an individual's age, income, health status, and type of coverage desired. Linear models are often used to predict claim costs for insurance companies, as they enable the segmentation of individuals based on risk factors (Samson and Thomas, 1987). Additionally, studies have been published using similar data sets that determine what impacts insurance premiums, but also expand their research to suggest possible policy interventions based on their results (Alzoubi et al., 2022). This indicates that there is a need to reduce insurance inequity and create more inclusive policies. However, the ability to perform these calculations may also be useful for individuals to understand better what health and lifestyle choices contribute to their insurance costs. This project aims to determine which variables from our data set can predict insurance charges by employing multiple linear regression (MLR) strategies. By developing a statistically sound predictive model, we hope to identify which variables are key effectors in increasing or decreasing premiums.

1.1.2 Problem

The problem we aim to address is understanding and predicting the factors that significantly influence medical insurance costs. Our focus specifically lies in identifying how key demographic characteristics, such as age and sex, alongside health-related indicators like BMI, impact insurance premiums. By analyzing these elements, our goal is to uncover which variables weigh most heavily in premium calculations. This knowledge has the potential to offer valuable insights both for insurers, in refining their pricing models, and for consumers, who may better understand how their lifestyle and choices affect insurance costs.

1.1.3 Challenges

One of the challenges we faced writing this report was to find variables that cause change in the cost of health insurance costs, rather than variables that are just associated with the change. We

are only looking at a small number of variables that relate to the person's health insurance cost, there of course can be a large amount of other factors that come into play that we do not get to see. For example, someone may be paralyzed, which may raise their insurance premium. Still, since it is not a variable in our data set we would never realize that is the reason their insurance is so much more expensive.

Other difficultues in creating this model is the amount of categorical variables that there are. This will mean that we have to use dummy variables to represent the different categories and will increase the amount of variables that we have. It will also make it so that there are significantly more interaction terms that we will need to look at and consider.

1.2 Objectives

1.2.1 Overview

The objective of this project is to analyze and compare different linear regression models on the health insurance data set, in order to get a better insight into health data and how different factors such as your age or number of children affect insurance costs. We also will find the best model to make future predictions about insurance costs. This will be done by first creating a variety of regression models such as a first order additive model, interaction term model, and higher order model. We will then use various statistical techniques to compare the effectiveness of these models in describing the data set, in order to find the best model that can be used to make future predictions about the cost of insurance.

Our approach involves a structured, data-driven solution to predict and analyze medical insurance costs. First, we'll explore the data, ensuring accuracy and identifying patterns within key variables like age and BMI. We will conduct a correlation analysis to identify significant relationships between insurance cost and our predictor variables, which will guide our modelling phase. Using regression and other machine learning techniques, we'll create a final model that reveals the most influential factors while also predicting costs.

Visualizations will play a key role in illustrating these findings. Histograms and other plots will help in conveying complex relationships, making our findings both engaging and practically informative. This approach will ultimately provide valuable insights into cost-driving factors, assisting both insurers and consumers in understanding the dynamics behind insurance pricing.

1.2.2 Goals & Research Questions

Our main goal is to create a model that can be used to accurately predict the insurance costs for an individual. Our research question aims to determine if an individual's BMI, sex, region, number of children, age and if they are a smoker are significant in predicting charges, and what MLR techniques can be employed to increase predictive power.

Chapter 2

Methodology

2.1 Data

The data set we will be using is a medical insurance costs data set (Rahul Vyas, 2024), containing 2772 rows of data on medical costs for individuals along across 7 attributes describing each person. The data set was uploaded to the website Kaggle in 2024 under the MIT open source license, meaning it is free to use, copy, or modify. The data is stated to have been collected from "multiple online and offline data sets". Although not stated one way or the other, we suspect this may be an artificial data set due to the lack of explicit sources. The data is stated to be a sample of 2772 diverse individuals, across the United States, in the year 2024. Each row in the data set represents an individual person. For a person listed in the data set, the 7 attributes used to describe them are:

- `age`: The person's age, measured in years. Quantitative. Varies from 18 to 64 in this data set.
- `sex`: The person's sex. Qualitative. Can be "male" or "female" in this data set.
- `bmi`: The Body Mass Index. Quantitative. This value is proportional to the ratio of a person's weight to their height squared, and is generally used to tell if a person is a healthy weight, while taking into account the effect of height. For weight W in pounds and height H in inches, the formula is $\text{BMI} = 703 * \frac{W}{H^2}$. Varies from 16 to 53.1 in this data set.
- `children`: The total number of children they have. Quantitative. Varies from 0 to 5 in this data set.
- `smoker`: Whether they smoke. Qualitative. Can be "yes" or "no" in this data set.
- `region`: The geographical sub-region of the United States they reside in, listed by inter cardinal direction. Qualitative. The value can take on one of "northeast", "northwest", "southeast", or "southwest".
- `charges`: The dollar amount they paid for insurance annually. Quantitative. Varies from \$1.12k to \$63.8k in this data set.

The `charges` column will be used as the response variable, and the rest will be used as predictor variables. A glimpse into the data set can be seen in Table B.1.

2.2 Approach

We are going to use R to do all of the statistical analysis of the data as well as creating all the graphs that we need as well. Specifically there are four R libraries that we are going to leverage to help us complete this project. GGplot (Wickham et al., 2024a) is going to be used in order to help us create graphs of our data, especially creating the beginning EDA in order to explore our data more effectively. The library olsrr (Hebbali, 2024) is going to help us with creating step wise models so that we do not have to do it manually but instead can use functions. GGally (Schloerke et al., 2024) is another library that we are going to use. This library is going to help us in determining which variables we should raise to a higher power. It has the ability to easily make a chart of all of our variables and their interactions with one another, also including the correlation coefficients. With these results we are able to determine which variable we should introduce into the model with a higher power first. We are also using tidyr (Wickham et al., 2024b) and dplyr (Wickham et al., 2023) to help assist us with the more general data manipulation and analysis, for example by using their functions such as gather and count to help us create graphs for our EDA.

This approach to the data analysis and visual creation will work well for us due to our experience with using these libraries from DATA 603. We know that these libraries in conjunction with R can create and help assist us with everything we need to do and we are confident that the results that we will get will reflect our understanding.

2.3 Workflow

2.3.1 Numerical Summaries

This step will involve calculating summaries of the data. For each of the quantitative variables, this will include calculating the mean as well as a five point summary (minimum, first quartile, median, third quartile, and maximum). For qualitative variables, this will be a count of the occurrences of each possible value of the variable. Afterwards, the next step will be to analyze the output in the context of the data set to see what we can learn about the data. This process will be relatively straightforward and should cause no issues.

2.3.2 Visual Summaries

This step will involve analyzing the data set by generating and interpreting graphs. In general, we will first look at the distribution of each variable on its own, and then how the response variable `charges` is affected by each of the predictor variables. Afterwards, we will analyze the graphs to learn about the data set. The most difficult part will be deciding which variable or set of variables to plot, or which variable to use for grouping, as well as deciding on the types of graphs to use. If a graph does not work well or does not tell us much information, the steps to try will be to add a grouping, such as colouring by sex or region, to see if we can learn more from that. If that does not work we can try a different type of graph as well.

2.3.3 Full Model

Creating a full model is an important step in beginning the assessment of a linear model. This step includes performing a linear regression test on all necessary variables in the data set, with the intent of determining which should remain in the model as predictors. By intentionally failing to include a variable, we risk the possibility of omitted variable bias (Wilms et al., 2021). This could lead to biased estimates and increase the risk of skewed predictions, especially if the variable being left out is a significant predictor (Wilms et al., 2021). Creating a full model is also important to assess the multicollinearity of all independent variables. Using tests such as the variance inflation factor allows for determining where further model assessment needs to be performed, which may otherwise be missed with the omission of a significant variable. Additionally, a full model helps assess if there are any significant predictors in tests such as ANOVA, which can tell us if further model refinement is necessary. It is important to note that all following tests will be evaluated against a significance level of 0.05.

2.3.4 Multicollinearity Assumption

After creating the full model, we will check that our model passes the multicollinearity test. The multicollinearity assumption is used to check that the independent variables in our model are not linearly correlated. We will check this assumption through calculating the Variance Inflation Factors (VIF). If the VIF equals 1, we can conclude that there is no collinearity between the independent variables. If the VIF falls between 1 to 5, we can conclude that there is moderate collinearity. If the VIF falls between 5 to 10, we can conclude that there is severe collinearity between the independent variables and we should consider removing one of the variables.

2.3.5 ANOVA Test

The ANOVA (Analysis of Variance) test is essential for determining if there are statistically significant variations among groups or model versions, especially when numerous factors or interactions come into play. In our examination, ANOVA is employed to contrast various regression models (e.g., null_model, health_first_order, health_full_model). It allows us to assess whether incorporating additional variables (or interaction terms) enhances the model's capacity to clarify the variability in insurance premiums. If the ANOVA indicates a significant enhancement, it suggests that the extra terms in the advanced model represent important variations in the data. This renders the model comprehensible for practical use —insurers or consumers can identify which life elements (such as age, BMI, or lifestyle) greatly influence premiums. Moreover, ANOVA assists in preventing over fitting by showing whether increasing complexity genuinely improves the model. This ensures that solely important variables and interactions are included, enhancing the model's efficiency, usability, and generalizability.

2.3.6 Individual T-Test

Individual t-tests are important in analyzing health insurance premiums since they facilitate specific, pairwise comparisons to determine which factors significantly affect costs. Although ANOVA shows whether factors such as age, BMI, or smoking status influence premiums, t-tests determine precisely where the differences occur. For instance, if ANOVA indicates that smoking status is

significant, a t-test can verify if there is a notable difference in premiums between smokers and non-smokers.

This specificity improves our understanding, clarifying which demographic or health aspects influence premium variations. It also eases interpretation, enabling us to convey specific results to stakeholders, like "smokers pay more in premiums than non-smokers," derived from important-test outcomes. In conclusion, t-tests enhance your analysis by providing depth and clarity, fine-tuning ANOVA's general results through specific comparisons.

2.3.7 Reduced Model

A reduced model is the result of performing ANOVA and individual t-tests on the full regression model. The reduced model generally contains only the most significant predictors. This allows for easier interpretation, as multiple variables are generally eliminated. The result is a more concise model for predictions, often having a lower residual standard error and a higher adjusted R-squared. By creating a reduced model, we can also remove predictors with high multicollinearity, meaning we mitigate the risk of redundancy in the model. The reduced model can then be improved by adding higher order or interaction terms, which is a simplified process when fewer variables are included.

2.3.8 Partial F-test

Running a partial F-test on the reduced model will confirm the results we obtained from the individual t-tests. First we will state our hypothesis:

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

To run a partial F-test we will construct an ANOVA table using the full model and the reduced model. This table will give us a p-value in which we will compare to our significance level ($\alpha = 0.05$). If the p-value is greater than α, we will fail to reject the null hypothesis and conclude that we should drop the specified variables off the model. This conclusion would confirm the results from the individual t-tests. If there is a case where the p-value is less than α, we will reject the null hypothesis and conclude that we should not drop the specified variables off the model. If this situation arises, we will have to look back at our model and decide on the best fit model before moving forward with our project.

2.3.9 Interaction Model

In order to determine if we are going to have interaction terms in the model we need to do some tests. Interaction terms occur when the variables in our model have an effect on eachother, and not just on the dependent variable. That is to say that the slope of one regression coefficient may be partly dependent on another variable.

We can take the full model with every independent and dependent variable and test the significance of all of the interaction terms that are possible. We will be constructing and testing eighteen different interaction terms.

We will test each one of them against the following alternate and null hypotheses.

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

The t-test statistic as well as the p-values for each of the regression coefficients will be calculated using R functions, but if we were to do it manually the following formula could be used. The t-test statistic for each of the regression coefficients could be calculated using the following formula.

$$t_{calc} = \frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)}$$

From this calculated t value we are able to obtain the p-value for the regression coefficient in question. We are going to use a level of significance of 0.05 and thus if the p-value we find is less than that we will be justified in rejecting the null hypothesis. If the p-value that we find is greater than our level of significance then we will not have sufficient evidence to reject the null hypothesis and will conclude that the regression coefficient is not statistically different from zero.

2.3.10 Higher Order Model

Once we have obtained the final model that includes the interaction terms that we have concluded to be significant, we can begin to look at whether adding higher order terms to our model will help with the accuracy.

In order to discover which variables have the highest likelihood of increasing the accuracy of our model when put to a higher power, we created a pairwise plot displaying all correlation coefficients between the independent variables and the dependent variable. Based on the correlation coefficients we will be able to assess which ones to raise to a higher power first.

We will take the variable with the highest correlation coefficient and raise it to a higher power, then look at the summary of the model to determine if the regression coefficient for the higher power variable is significant. When doing this, we also need to ensure the previous lower power variable is still significant when adding the higher power terms. If any of the variables of a lower power are no longer significant, then we must take a step back to ensure that they all stay significant.

For this we will use the following null and alternate hypothesis.

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

Using a level of significance of 0.05, we will look at the p-value returned for each of the higher power variables and determine if we can reject the null hypothesis or not. This will let us determine if we are going to keep the higher power variable in our model.

2.3.11 All Possible Regressions Method

This step will involve using an automatic algorithm to find the best subset of predictor variables to use for the model, in order to later compare to other procedures. The first step in this process is to generate the largest possible model given the data set. This model is then passed into a function from the `olsrr` package (Hebbali, 2024) called `ols_step_best_subset`. Given a set of predictor variables of length n, this function will calculate all $\binom{n}{k}$ possible models of size k, for each $k \in [1, n]$. For each model length k, the function will calculate the value of a statistical criteria for the models, and return the best one. The default statistical criteria is the largest R^2. Then, with the returned list of best models for each length, we can compare other statistical criteria such as R^2_{adj} or AIC, to find which length of model is the best at describing the data set. Since this process looks at every possible model, we are guaranteed to find the best additive model of a certain length. We can then compare this to other methods, to check if results are similar. The hardest step in this process will be comparing the different sized models, since we want to be careful of over fitting, or throwing away too many variables early on, and the statistical criteria can be difficult to compare definitively.

2.3.12 Analysis

After we have created the full model, first-order model, interaction model, higher order model and all possible regressions model, we will choose the best fit model for predictive purposes. To help us make a decision, we will be looking at the residual standard error (RSE), the adjusted R^2, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) for each model. We will choose the model that has the lowest RSE, an adjusted R^2 that is closest to 1, the lowest AIC or the lowest BIC. Before finalizing our decision, we will discuss the potential of over fitting the model. Once we are confident with our choice, we will use the model for predictive purposes.

2.3.13 Assumptions

After we have chosen the best fit model, we will check to see if the model meets the linearity assumption, independence assumption, equal variance assumption, and normality assumption. We will also check for outliers.

Linearity Assumption

The linearity assumption test is used to check that our model has a linear relationship between the cost of insurance and the predictor variables. We will begin by creating a residual plot. To pass the linearity assumption, the residual plot should show no discernible pattern of the residuals.

Independence Assumption

The independence assumption is used to check that our model's error terms are mutually independent. To check this assumption, we will look at the predictor variables to confirm that they are not related to time, space, or group.

Equal Variance Assumption

The equal variance assumption is used to check that our model's error terms have a constant variance. To test our model, we will conduct a Breusch-Pagan test. We will first state our hypothesis:

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2$$
$$H_1 : \text{At least one } \sigma_i^2 \text{ is different than the others, for } i = 1, 2, ..., n$$

Running the `bptest()` function in R will give us a p-value in which we will compare to our significance level ($\alpha = 0.05$). If the p-value is greater than α, we will fail to reject the null hypothesis and conclude that homoscedasticity is present in the model. If the p-value is less than α, we will reject the null hypothesis and conclude that heteroscedasticity is present in the model.

Normality Assumption

The normality assumption is used to check that our model's error terms are normally distributed. Will will test this assumption using two methods. We will begin by looking at the Q-Q plot. If the residuals on the plot are close to the diagonal reference line, our model will pass the normality assumption. If the residuals deviate away from the reference line, then our model will not pass the normality assumption. The second method to test normality will be done by conducting a Shapiro-Wilk Normality test. We will state our hypothesis:

$$H_0 : \text{residuals are normally distributed}$$
$$H_1 : \text{residuals are not normally distributed}$$

Running the `shapiro.test(residuals())` function in R will give us a p-value in which we will compare to our significance level ($\alpha = 0.05$). If the p-value is greater than α, we will fail to reject the null hypothesis and conclude that the residuals are normally distributed. If the p-value is less than α, we will reject the null hypothesis and conclude that the residuals are not normally distributed.

Outliers

The last thing we need to check for is if there are any influential outliers. We will check for outliers using two different methods; the Residual vs Leverage plot and the Cook's Distance method. Both the Residual vs Leverage plot and the Cook's Distance plot show us values above 0.5. If the plots show values over 0.5, we will remove them from the analysis. Removing any influential outliers could help normalize our model. If the plots show that there are no values above 0.5, indicating there are no significant outliers, then we will not remove any outliers from the data.

2.3.14 Prediction

After we have our final best model, we will use it to make an estimate of a fictional person's medical costs, based on the relevant factors that are significant in the model.

2.4 Contributions

In the introduction, each group member will write one of the 5 subsections. The data summary and approach will be written by Noah and Max respectively. Following this, we will split up the methodology and main analysis steps as well as their respective write-ups as follows; Noah will work on the numeric and visual EDA and the best-subsets model. Kennedy will work on the initial first order additive model and the reduced model, as well as write interpretations for all models. Stephen will work on the analysis of these models with ANOVA and individual t-tests. Bobbi will work on the partial F-test analysis and the final comparison of all models. Max will work on the interaction model and the higher order model. Checking assumptions will be done by Bobbi. Making predictions will be done by Noah. The conclusion and future work steps will be written by Max and Kennedy respectively. Combining individual work and formatting the document and references will be done by Noah. Work on the oral presentation of the report will be distributed based on the above workload distribution, due to familiarity with each section.

Chapter 3

Results

3.1 Numerical Summaries

The results of the summaries for the quantitative variables can be seen in Table B.2. Each variable shows the mean, as well as the 5 point summary (minimum, Q1, Q2, Q3, and maximum), which gives us an idea of how they are each distributed. We can see that the median person is 39 years old, with a BMI of 30.45, has 1 child, and spends about 9 thousand dollars a year in medical expenses. We can also see that all four of these variables have a mean that is larger than the median, meaning there are likely a few large outliers in each that skew the distribution to the right.

The summaries for the qualitative variables can be seen in Table B.3. Notable results are that there is approximately the same number of males as females in the dataset (to within 3%), far more nonsmokers than smokers (2208 versus 564, almost 4x), and roughly 670 people in each region, with the exception of southeast at 766. In addition, since each column sums to 2772 (the number of rows in the dataset), we can note that none of the rows have missing data for these three columns.

3.2 Visual Summaries

From Figure C.1 we can see the distributions of the numeric predictors. There appears to be more 18 and 19 year olds than any other age, but the rest are fairly uniform. The BMI variable is bell shaped, with the most common BMI being about 30. The charges and children variables are both skewed to the right, indicating that a small number of children and less annual medical costs is most common, but there are a small number of large values that skew the data to the right.

From Figure C.2 showing the distribution of categorical data, it appears that the region variable is distributed fairly evenly, with southeast being slightly more common. The distribution of sex is fairly even as we would expect from a random sample. There are also significantly fewer smokers than nonsmokers in the data set, which may have an impact on comparisons across the full data set that don't take into account this difference.

In Figure C.3 we see the distribution of medical charges, divided into smokers and nonsmokers. We see a stark difference here, where the higher medical charges are almost entirely from smokers, and the lower values are almost entirely from nonsmokers. The distribution of charges for smokers

appears to be bimodal, indicating there may be another factor at play here. Conversely, when we split the distribution by sex in Figure C.4, the distribution of medical charges does not seem to vary by sex.

In Figures C.5 and C.6, we see the distribution of medical charges by number of children and subregion respectively. From a glance, it does not appear there is much of an effect on charges by number of children and region, or if there is, it may be a small one.

Figure C.7 shows us the effect of BMI on medical costs. For nonsmokers, there does not appear to be much of an effect. However, for smokers there is a clear increase in charges for those with a BMI below 30 and those with a BMI above 30. This split is likely what is responsible for the bimodal distribution seen earlier in Figure C.3.

Lastly, we can see how the medical costs vary by age in Figure C.8. We can see three "bands" in the plot. From the colouring based on smoking status, we can see the lower band is nonsmokers, and the higher band is smokers, and the middle band is mixed. Overall, charges seem to increase as age increases, which is not unsurprising.

3.3 Full Model

To begin the process of creating an appropriate model for predicting medical insurance costs, we began by creating a full model with all variables present in the data set. In this case, sex, region, and whether or not the individual is a smoker were all categorical variables, while age, BMI, and number of children were continuous. As such, we add the categorical predictors to the model using the `factor()` function.

Running the summary function (Appendix A.4) gives the coefficients for each variable, and the following equation:

$$\begin{aligned} Y_{charges} = &- 11635.451 + 255.577X_{age} + 330.015X_{bmi} + 506.343X_{children} + 23976.197X_{smoker} \\ &- 56.944X_{sex} - 331.841X_{northwest} - 1078.362X_{southeast} - 1055.254X_{southwest} \end{aligned}$$

At this point, the resulting model has a residual standard error (RSE) of 6073 on 2763 degrees of freedom and an Adjusted R-squared of 0.7502. More specifically, this means that 75.02% of the variation in the charges variable is explained by the predictors present in the model. However, we need to run an ANOVA test and individual t-tests to determine which variables should remain in the model, and which need to be removed to make our model more robust.

3.3.1 Full Model Coefficient Interpretations

- β_0: Negative \$11,635.451 is the insurance cost when all predictors are zero. This occurs when the individual is female, the individual is not a smoker, and the region is northeast.
- β_1: For each additional increase in an individual's age, the insurance price increases by \$255.577, while holding all other variables constant.
- β_2: For each single unit increase in BMI, the insurance charges increase by \$330.015, while holding other variables constant.

- β_3: For each additional child, the insurance charges increase by $506.343, when holding other variables constant.
- β_4: The difference in the cost of insurance between smokers and non-smokers. When an individual is a smoker, they have a $23,976.197 increase in insurance charges compared to non-smokers, when all other variables are held constant.
- β_5: The difference in the cost of insurance between males and females. The cost of insurance for males is approximately $56.944 lower than for females when other variables are held constant.
- β_6: Living in the Northwest region results in a $331.841 decrease in insurance prices when compared to those living in the Northeast, when all other variables are held constant.
- β_7: Living in the Southeast region results in a $1,078.362 decrease in insurance prices when compared to those living in the Northeast, when all other variables are held constant.
- β_8: Living in the Southwest region results in a $1,055.254 decrease in insurance prices when compared to those living in the Northeast, when all other variables are held constant.

3.4 Multicollinearity Assumption

Now that we have created our full model, we will check that our model passes the multicollinearity test through calculating the Variance Inflation Factors (VIF).

The results of this test (Appendix A.5) show that VIF's are between 1.004 and 1.6807 for each of the independent variables. Based on these results, we can conclude that there is no collinearity between the independent variables.

3.5 ANOVA Test

The results of the ANOVA test shows a p-value which is way below alpha(0.05), which is $2x10^{-16}$. Due to this, we reject the null hypothesis that all of the variables equal zero and therefore come to the conclusion that at least one does not equal zero.

Running the ANOVA code (Appendix A.6) will provide the results Analysis of Variance Table. The results are significant and suggest that adding variables like age, BMI, or smoking status meaningfully improves the model, indicating these factors affect premiums.

3.6 Individual T-Test

The results of the individual t-test highlight particular factors influencing health insurance premiums by analyzing the means of two groups. Significant p-values suggest that disparities between groups (e.g., smokers compared to non-smokers) hold statistical significance. The orientation of the average difference indicates which group possesses higher premiums, while the effect size emphasizes the magnitude of this difference. These insights provide clear, data-supported explanations for premium differences, aiding in decision-making regarding pricing and understanding cost factors.

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

We are going to use the above as hypotheses to look into each individual predictor, using a significance level, alpha of 0.05.

Age

Age has a p-value of $2x10^{-16}$, which is less than 0.05, therefore we reject the null hypothesis and confirm that it is significant in our model.

BMI

It also has a p-value of $2x10^{-16}$. Since it is less than alpha, we will reject the null hypothesis and conclude that BMI is not equal to 0

Children

With a p-value of $1.12x10^{-7}$, we reject the null hypothesis since it has a p-value way less than significance level.

Smoker

P-value of smoker is $2x10^{-16}$ and this is much smaller than 0.05 and so we reject the null hypothesis.

Sex

This is the first predictor to be dropped from our model because it has a p-value of 0.80602 which is higher than the significance level. Therefore, we fail to reject the null hypothesis and conclude that $\beta_{sex} = 0$.

Region-Northwest

It has a p-value = 0.32109. Since p-value > alpha, we fail to reject the null hypothesis. This is also insignificant in our model, but we will have to check the other levels of the categorical variable before deciding to get rid of it.

Region-Southwest

The p-value of Region-Southwest is 0.00155, which is less than 0.05. We will therefore reject the null hypothesis.

Region-Southeast

It has a p-value of 0.00128. Since this is less than the significance level, we reject the null hypothesis and conclude that $\beta_{Southeast} \neq 0$, and therefore we need to keep all the region levels in the model.

3.7 Reduced Model

After running both an ANOVA test and individual t-tests, we determined that sex was not a significant predictor to be included in our model. In the context of both the topic and the data set, this result is unsurprising, as many insurance companies use standardized pricing models, which

fail to differentiate between sexes. Additionally, even though the northwest region is insignificant, we include it in the model as it is significant for the southeast and southwest regions. The code to produce the final reduced model can be seen in Appendix A.7.

By excluding sex in our model, the RSE decreases a single unit, with a new value of 6072 on 2764 degrees of freedom, and an adjusted R-squared of 0.7503. In other words, the predictors in the model explain 75.03% of the variation present in the charges variable.

$$\begin{aligned}\hat{Y}_{charges} = &-11659.114 + 255.637X_{age} + 329.809X_{bmi} + 505.941X_{children} + 23970.504X_{smoker} \\ &- 331.005X_{northwest} - 1078.176X_{southeast} - 1054.856X_{southwest}\end{aligned}$$

3.7.1 Reduced Model Coefficient Interpretations

- β_0: Negative \$11,659.114 is the insurance cost when all predictors are zero. This occurs when the individual is female, the individual is not a smoker, and the region is northeast.
- β_1: For each additional increase in an individual's age, the insurance price increases by \$255.637, while holding all other variables constant.
- β_2: For each single unit increase in BMI, the insurance charges increase by \$329.809, while holding other variables constant.
- β_3: For each additional child, the insurance charges increase by \$505.941, when holding other variables constant.
- β_4: The difference in the cost of insurance between smokers and non-smokers. When an individual is a smoker, they have a \$23,970.504 increase in insurance charges compared to non-smokers, when all other variables are held constant.
- β_6: Living in the Northwest region results in a \$331.005 decrease in insurance prices when compared to those living in the Northeast, when all other variables are held constant.
- β_7: Living in the Southeast region results in a \$1,078.176 decrease in insurance prices when compared to those living in the Northeast, when all other variables are held constant.
- β_8: Living in the Southwest region results in a \$1,054.856 decrease in insurance prices when compared to those living in the Northeast, when all other variables are held constant.

3.8 Partial F-test

Running a partial F-test will confirm our results from the individual t-tests indicating the reduced model is currently the best fit model. Our hypothesis states:

$$\begin{aligned}H_0 :&\beta_{age} = \beta_{bmi} = \beta_{children} = \beta_{smoker} = \beta_{region} = 0 \\ H_1 :&\text{at least one of } \beta_{age}, \beta_{bmi}, \beta_{children}, \beta_{smoker}, \beta_{region} \neq 0\end{aligned}$$

To find our p-value we will create an ANOVA table using our full model and our reduced model (Appendix A.8).

$\alpha = 0.05$

p-value = 0.806

Since p-value $> \alpha$, we fail to reject the null hypothesis and conclude that we should drop the categorical variable sex off the model. This conclusion is consistent with the results found in the individual t-tests.

3.9 Interaction Model

To find and test all of the interaction terms in our model at once, we can create a new linear model that contains all of our independent variables and square the whole variable term. This will create a linear model that contains all the possible interaction terms.

We are going to use the following hypotheses when looking at each of the interaction term's t values and corresponding p-values.

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

We are going to use a significance value of 0.05.

Once we call a summary on this new model, we will have the ability to see the t value calculated for each of the interaction terms, as well as the p-value and thus be able to conclude whether we accept or reject null.

From the results in Appendix A.9, we see that there are a total of five interaction terms whose p-values are less than our significance level of 0.05, and therefore we are going to choose to keep in our model as we have tested them to be significant as they have sufficient evidence to be able to reject the null hypothesis and accept the alternate hypothesis.

`Age*factor(region)southeast`

This interaction term has a calculated t value of 2.444 and a corresponding p-value of 0.0146.

`age*factor(region)southwest`

This interaction term has a calculated t value of 2.300 and a corresponding p-value of 0.0215.

`bmi*factor(smoker)yes`

This interaction term has a calculated t value of 37.790 and a corresponding p-value of less than $2\text{x}10^{-16}$.

`bmi*factor(region)southeast`

This interaction term has a calculated t value of -4.634 and a p-value of $3.75\text{x}10^{-6}$

`bmi*factor(region)southwest`

This interaction term has a calculated t value of -2.161 and a p-value of 0.0308.

Because region is a categorical variable, it has dummy variables that are created for its different categories. We see that only some of the regions are significant when interacting with BMI and

age, however we can not only include some of the regions and thus we are going to keep all the interaction terms between region and age as well as region and BMI.

The final model (Appendix A.10) we obtain from this is as follows:

$$\begin{aligned}\widehat{Y}_{charges} = &-3616.125 + 233.858X_{age} + 105.4839X_{bmi} + 509.491X_{children} - 20918.581X_{smoker} \\ &- 461.785X_{northwest} + 2733.186X_{southeast} - 280.224X_{southwest} + 8.105X_{age} * X_{northwest} \\ &+ 51.309X_{age} * X_{southeast} + 41.7735X_{age} * X_{southwest} + 1457.361X_{bmi} * X_{smoker} \\ &- 13.709X_{bmi} * X_{northwest} - 191.661X_{bmi} * X_{southeast} - 92.140X_{bmi} * X_{southwest}\end{aligned}$$

3.9.1 Interaction Model Coefficient Interpretations

- β_0: Negtive \$3,616.125 is the insurance cost when all predictors are zero. This occurs when the individual is female, the individual is not a smoker, and the region is northeast.
- β_1: For each additional increase in an individual's age, the insurance price increases by \$233.858, while holding all other variables constant.
- β_2: For each single unit increase in BMI, the insurance charges increase by \$105.4839, while holding other variables constant.
- β_3: For each additional child, the insurance charges increase by \$505.941, when holding other variables constant.
- β_4: The difference in the cost of insurance between smokers and non-smokers. When an individual is a smoker, they have a \$20,918.581 decrease in insurance charges compared to non-smokers, when all other variables are held constant, and before accounting for interaction terms.
- β_6: Living in the Northwest region results in a \$461.785 decrease in insurance prices when compared to those living in the Northeast, when all other variables are held constant, and before accounting for interaction terms.
- β_7: Living in the Southeast region results in a \$2,733.19.176 increase in insurance prices when compared to those living in the Northeast, when all other variables are held constant, and before accounting for interaction terms.
- β_8: Living in the Southwest region results in a \$280.224 decrease in insurance prices when compared to those living in the Northeast, when all other variables are held constant, and before accounting for interaction terms.
- β_9: The age and Northwest interaction term produced a value of 8.105. This means that for each additional increase in age, the insurance price increases by \$8.11 in the Northwest region when compared to the Northeast region.
- β_{10}: The age and Southeast interaction term produced a value of 51.309. As such, for each additional increase in age, the insurance charge increases by \$51.31 more in the Southeast region than in the Northeast region.
- β_{11}: The age and Southwest interaction term produced a value of 41.7735. As such, for each additional increase in age, the insurance charge increases by \$41.7735 more in the Southwest region than in the Northeast region.
- β_{12}: The BMI and smoker interaction term produced a value of 1457.361. This means that for each additional increase in BMI, the insurance charge increases by \$1,457.361 more for smokers than for non-smokers.

- β_{13}: The BMI and Northwest interaction term produced a value of -13.709. This means that for each additional increase in BMI, the insurance price decreases by $13.709 when compared to the Northeast
- β_{14}: The BMI and Southeast interaction term produced a value of -191.661. This means that for each additional increase in BMI, the insurance price decreases by $191.661 when compared to the Northeast
- β_{15}: The BMI and Southwest interaction term produced a value of -92.140. This means that for each additional increase in BMI, the insurance price decreases by $92.140 when compared to the Northeast

3.10 Higher Order Model

Our first step in order to see which variables we should put to a higher power is going to be creating a graph of all of the interactions between our variables. This will show us the correlation coefficients between all the dependent and independent variables which will help us determine which ones we should try to put to a higher power.

We have not included any of the categorical variables in this analysis as they will not return us with any sort of correlation coefficient relating to the independent variable and thus we have left them out of the graph.

From Figure C.9 we see that the variable that has the highest correlation coefficient with the dependent variable is age. Therefore, the first higher power variable that we are going to create is going to be age squared. When we do this we see that the p-value for age squared is $4.06\text{x}10^{-11}$, however the p-value for age has gone up and is now 0.20611. Because adding the squared age variable has made our original age variable not significant, we are not going to include age squared.

The next variable that we are going to look at is BMI as it has the second highest correlation coefficient in relation to the dependent variable. When adding BMI squared we see that it has a p-value of 0.00437 which is less than our level of significance of 0.05, and BMI has a p-value of 0.000349 which is also less than our significance level and thus we can keep BMI squared in our model and move on to looking at BMI cubed. Additionally the models adjusted r squared value has gone from 0.8413 to 0.8417 which shows that adding BMI squared has had a small positive impact on the models fit. We are now going to add BMI cubed, and we see that it has a corresponding p-value of $5.03\text{x}10^{-09}$, and BMI squared as well as BMI p-values are also both still under our significant level. Therefore, we are going to include BMI cubed as well. The adjusted r squared value is now 0.8436, which is a slight increase over the last model with BMI squared being the highest order. Moving on, we are going to add a BMI to the power of four to the model, and when we do we see that it has a p-value of 0.000121, and all the previous BMI variable's p-values are also under the significance level so we are going to include BMI to the power of four as well. The adjusted r squared value has now increase slightly to 0.8444. We also are going to try BMI to the power of five, but the p-value that we get for it is 0.06620 and thus is not significant and is not going to be included in the model.

Lastly, we are going to introduce the variable children squared to see if it is significant and if it increases the accuracy of our model. When adding children squared we see that it has a corresponding p-value of 0.393794 and is therefore not significant and is not going to be included in the final mode.

From the output in Appendix A.11 we can now state that the final model, with higher order variables added as:

$$\begin{aligned}\widehat{Y}_{charges} = 97940 &+ 233.6X_{age} - 12610X_{bmi} + 568.4X_{bmi}^2 - 10.76X_{bmi}^3 + 725.6X_{bmi}^4 + 527.9X_{children} \\ &- 20860X_{smoker} + 380.7X_{northwest} + 1484X_{southeast} - 23.32X_{southwest} \\ &+ 11.38X_{age} * X_{northwest} + 50.07X_{age} * X_{southeast} + 44.59X_{age} * X_{southwest} + 1456X_{bmi} * X_{smoker} \\ &- 47.56X_{bmi} * X_{northwest} - 148.3X_{bmi} * X_{southeast} - 106.1X_{bmi} * X_{southwest}\end{aligned}$$

3.10.1 Higher Order Model Coefficient Interpretations

As a result of the higher-order terms, we cannot meaningfully interpret the coefficients beyond understanding that this model would produce a concave upward slope based on the signs of the higher-order terms. Additionally, since it is not meaningful for the age, and BMI to be equal to zero when predicting insurance costs for an individual, we can not interpret this value either.

3.11 All Possible Regressions Method

Running the stepwise best subsets function from olsrr (Hebbali, 2024), we get back a list of the subset of the model with (by default) the largest R^2 value for each number of predictors from $n = 1$ to $n = 6$. We do not have to use the adjusted R^2 to compare here, because we are comparing only models with the same number of variables at each step. The results of this process can be seen in Table B.4. From the best model for each number of predictors, we can compare their statistical criteria. Here, we want to look at the adjusted R^2, because we are comparing across different numbers of predictors. The largest adjusted R^2 is for the $n = 5$ model, which also has the smallest AIC, and the smallest Mallows C_p criteria. We can generate the coefficients for this model with `summary` (Table B.5). From the code output (Appendix A.12) the best model for medical charges for person i will be:

$$\begin{aligned}\hat{Y}_{charges} = &- 11659.114 + 255.637X_{age} + 329.809X_{bmi} + 505.941X_{children} + 23970.504X_{smoker} \\ &- 331.005X_{northwest} - 1078.176X_{southeast} - 1054.856X_{southwest}\end{aligned}$$

Where the categorical variable $smoker_i$ is 1 for smoker, 0 for nonsmoker, and the categorical variables $region_{ji}$ for $j \in [1, 3]$ is 000 for northeast, 100 for northwest, 010 for southeast, and 001 for southwest. These results are as expected, the best-subsets criteria reiterated what we found out from the reduced model, that sex is not a significant predictor for medical charges.

3.11.1 All Possible Regression Model Coefficient Interpretations

Since this model is the same as the reduced model (Section 3.7), the interpretation of coefficients will not change here.

3.12 Analysis

Now that we have created the full model, first order model, interaction model, higher order model and all possible regressions model, we will choose the model that is the best fit model for prediction purposes. When looking at Table B.6, we can see that the reduced model and all possible regressions model have the same criterion. This is because the possible regression model and the reduced model dropped the sex variable due to its insignificance when predicting health insurance costs. Based on Table B.6 we can see that the higher order model has the lowest RSE, AIC, and BIC, as well as having the highest adjusted R^2. This model meets the requirements for every criterion we are using to choose the best model. Although it is clear that the higher order model is the best fit, we are cautious of the potential of overfitting the model.

3.13 Assumptions

Now that we have chosen the best fit model, we will check to see if the model meets the linearity assumption, independence assumption, equal variance assumption, and normality assumption.

Linearity Assumption

When analyzing the residual plot (Figure C.10), we can see that there is no discernible pattern of the residuals and therefore the model meets the requirements of the linearity assumption.

Independence Assumption

The model we are working with does not have variables that are related to time, space or group, and therefore we can assume that the error terms are mutually independent.

Equal Variance Assumption

Running an equal variance test will tell us if our model's error terms have a constant variance. Our hypothesis states:

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2$$

$$H_1 : \text{At least one } \sigma_i^2 \text{ is different than the others where i = 1,2,...,n}$$

The output of the code can be seen in Appendix A.13, where we get a result of:

$\alpha = 0.05$

P-value = 0.0019

Since p-value < alpha, we reject the null hypothesis and conclude that heteroscedasticity is present. This tells us that our model does not pass the Breusch-Pagan test (equal variance assumption).

Normality Assumption

When looking at our QQ Plot of residuals (Figure C.11), we can see that the residuals do not fall close to the diagonal reference line and therefore we can conclude that the model does not pass the normality assumption.

The second test we will run is the Shapiro-Wilk Normality test. This test will also tell us if our model's error terms are normally distributed. Our hypothesis states:

H_0 :residuals are normally distributed

H_1 :residuals are not normally distributed

The result of the code output can be seen in Appendix A.14, where we see a result of:

$\alpha = 0.05$

P-value = $2.2\text{x}10^{-16}$

Since p-value $< \alpha$, we reject the null hypothesis and conclude that the residuals are not normally distributed.This tells us that our model does not pass the Shapiro-Wilk Normality test (normality assumption). This conclusion is consistent with the findings from analyzing the Q-Q residual plot.

Outliers

Lastly, we will check for any influential outliers. Based on the Residual vs Leverage plot, graph (Figure C.12), there are no values greater than 0.5 and therefore there are no influential outliers. We can see that the Cook's Distance plot, graph (Figure C.13), also shows that there are no values above 0.5. Based on the two outlier tests, we can conclude that there are no influential outliers.

3.14 Prediction

With the final best model chosen to be the higher order model, we can make a few predictions. First, we define the attributes of a handful of fictional people, and then use the model to predict their annual medical costs. The input values and 95% confidence interval prediction results can be seen in Table B.7. For example, we can be 95% confident that for a 26 year old with a BMI of 28, no children, and who does not smoke, from the northwestern united states, that their annual medical costs will be between \$3924.31 and \$5039.65. We can also see that in the last row the effect of smoking: A young person with a low BMI and no children, who would otherwise seem healthy, has annual medical costs around 10 thousand dollars, almost equivalent to the medical costs in the third row, a much older person with multiple children who does not smoke. The reliability of these predictions should be taken into account, however, as many of the assumptions were not met.

Chapter 4

Conclusion

4.1 Approach

The approach that we took to modelling our data looks quite promising. The data that we used for the model has a nice relationship between the dependent and independent variables and thus it makes it easier to construct a model that resembles this relationship. Due to the nature of our data we knew that there would be a significant multiple linear regression that we could model.

One thing that we could have done to make this approach better would be to transform our model to try and make it so that it satisfies all of the assumptions. As it stands right now our final model does not pass the equal variance test or the normality test.

4.2 Future Work

Future steps may include testing other types of models such as logistic regression. In this case, the probability that an individual is a smoker or not based on other insurance variables. Additionally, since our model failed to pass many of the assumptions, it may be meaningful to perform a transformation on the data to determine if this improves the model fit. Depending on how the logistic regression model performs, we may find that this model is better for predictive purposes in respect to the dependent variable selected. Alternatively, it may be meaningful to perform the same workflow and steps on an additional insurance data set, as the data we selected likely has biases due to its artificial nature.

References

H. M. Alzoubi, N. Sahawneh, A. Q. AlHamad, U. Malik, A. Majid, and A. Atta. Analysis of cost prediction in medical insurance using modern regression models. In *2022 International Conference on Cyber Resilience (ICCR)*, pages 1–10, 2022. doi: 10.1109/ICCR56254.2022.9995926.

A. Hebbali. *olsrr: Tools for Building OLS Regression Models*, 2024. URL https://olsrr.rsquaredacademy.com/. R package version 0.6.0.

T. Hothorn, A. Zeileis, R. W. Farebrother, and C. Cummins. *lmtest: Testing Linear Regression Models*, 2022. URL https://CRAN.R-project.org/package=lmtest. R package version 0.9-40.

M. U. Imdad, M. Aslam, S. Altaf, and A. Munir. Some new diagnostics of multicollinearity in linear regression model. *Sains Malaysiana*, 48(9):2051–2060, 2019. URL http://dx.doi.org/10.17576/jsm-2019-4809-26.

M. Imdadullah, M. Aslam, and S. Altaf. mctest: An r package for deteection of collinearity among regressors. *The R Journal*, 8(2):499–509, 2016. URL https://journal.r-project.org/archive/2016/RJ-2016-062/index.html.

I. U. Muhammad and A. Muhammad. *mctest: Multicollinearity Diagnostic Measures*, 2020. URL https://CRAN.R-project.org/package=mctest. R package version 1.3.1.

M. Rahul Vyas. Medical insurance cost prediction, 2024. URL https://www.kaggle.com/datasets/rahulvyasm/medical-insurance-cost-prediction.

D. Samson and H. Thomas. Linear models as aids in insurance decision making: The estimation of automobile insurance claims. *Journal of Business Research*, 15(3):247–256, 1987. ISSN 0148-2963. doi: https://doi.org/10.1016/0148-2963(87)90027-0. URL https://www.sciencedirect.com/science/article/pii/0148296387900270.

B. Schloerke, D. Cook, J. Larmarange, F. Briatte, M. Marbach, E. Thoen, A. Elberg, and J. Crowley. *GGally: Extension to ggplot2*, 2024. URL https://ggobi.github.io/ggally/. R package version 2.2.1.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

H. Wickham, R. François, L. Henry, K. Müller, and D. Vaughan. *dplyr: A Grammar of Data Manipulation*, 2023. URL https://dplyr.tidyverse.org. R package version 1.1.4.

H. Wickham, W. Chang, L. Henry, T. L. Pedersen, K. Takahashi, C. Wilke, K. Woo, H. Yutani, D. Dunnington, and T. van den Brand. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, 2024a. URL https://ggplot2.tidyverse.org. R package version 3.5.1.

H. Wickham, D. Vaughan, and M. Girlich. *tidyr: Tidy Messy Data*, 2024b. URL https://tidyr.tidyverse.org. R package version 1.3.1.

C. O. Wilke. *ggridges: Ridgeline Plots in ggplot2*, 2024. URL https://wilkelab.org/ggridges/. R package version 0.5.6.

R. Wilms, E. Mäthner, L. Winnen, and R. Lanwehr. Omitted variable bias: A threat to estimating causal relationships. *Methods in Psychology*, 5:100075, 2021. ISSN 2590-2601. doi: https://doi.org/10.1016/j.metip.2021.100075. URL https://www.sciencedirect.com/science/article/pii/S2590260121000321.

Y. Xie. knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, and R. D. Peng, editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014. ISBN 978-1466561595.

Y. Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015. URL https://yihui.org/knitr/. ISBN 978-1498716963.

Y. Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2024. URL https://yihui.org/knitr/. R package version 1.48.

A. Zeileis and G. Grothendieck. zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6):1–27, 2005. doi: 10.18637/jss.v014.i06.

A. Zeileis and T. Hothorn. Diagnostic checking in regression relationships. *R News*, 2(3):7–10, 2002. URL https://CRAN.R-project.org/doc/Rnews/.

A. Zeileis, G. Grothendieck, and J. A. Ryan. *zoo: S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations)*, 2023. URL https://zoo.R-Forge.R-project.org/. R package version 1.8-12.

Appendix A

Code

A.1 Setup

```
knitr::opts_chunk$set(
  echo = TRUE,
  fig.width=10,
  fig.height=6)
```

A.2 Load Libraries

```
library(ggplot2)
library(GGally)
library(ggridges)
library(dplyr)
library(tidyr)
library(olsrr)
library(lmtest)
library(mctest)
library(knitr)
```

A.3 Load Data

```
medical_ins = read.csv('medical_insurance.csv', stringsAsFactors = TRUE)
head(medical_ins, 10)
```

```
##    age    sex    bmi children smoker    region   charges
## 1   19 female 27.900        0    yes southwest 16884.924
```

```
## 2   18   male 33.770        1     no southeast  1725.552
## 3   28   male 33.000        3     no southeast  4449.462
## 4   33   male 22.705        0     no northwest 21984.471
## 5   32   male 28.880        0     no northwest  3866.855
## 6   31 female 25.740        0     no southeast  3756.622
## 7   46 female 33.440        1     no southeast  8240.590
## 8   37 female 27.740        3     no northwest  7281.506
## 9   37   male 29.830        2     no northeast  6406.411
## 10  60 female 25.840        0     no northwest 28923.137
```

A.4 Full Model Summary

```
health_full_model = lm(charges ~ age + bmi + children + factor(smoker)
    + factor(sex) + factor(region), data = medical_ins)

summary(health_full_model)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + factor(smoker) +
##     factor(sex) + factor(region), data = medical_ins)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11489  -2789  -1016   1340  29867
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -11635.451    686.885 -16.939  < 2e-16 ***
## age                        255.577      8.268  30.913  < 2e-16 ***
## bmi                        330.015     19.869  16.609  < 2e-16 ***
## children                   506.343     95.164   5.321 1.12e-07 ***
## factor(smoker)yes        23976.197    288.461  83.118  < 2e-16 ***
## factor(sex)male            -56.944    231.866  -0.246  0.80602
## factor(region)northwest   -331.841    334.380  -0.992  0.32109
## factor(region)southeast  -1078.362    334.418  -3.225  0.00128 **
## factor(region)southwest  -1055.254    333.121  -3.168  0.00155 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6073 on 2763 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7502
## F-statistic:  1041 on 8 and 2763 DF,  p-value: < 2.2e-16
```

A.5 Multicollinearity

```
imcdiag(health_full_model,method = "VIF")
```

```
##
## Call:
## imcdiag(mod = health_full_model, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                              VIF detection
## age                       1.0182         0
## bmi                       1.1143         0
## children                  1.0040         0
## factor(smoker)yes         1.0135         0
## factor(sex)male           1.0099         0
## factor(region)northwest   1.5306         0
## factor(region)southeast   1.6807         0
## factor(region)southwest   1.5501         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

A.6 Anova (Full Model)

```
null_model = lm(charges~1, data = medical_ins)

print(anova(null_model, health_full_model))
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ 1
## Model 2: charges ~ age + bmi + children + factor(smoker) + factor(sex) +
##     factor(region)
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   2771 4.0918e+11
## 2   2763 1.0191e+11  8 3.0727e+11 1041.3 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A.7 Reduced Model

```
health_first_order = lm(charges ~ age + bmi + children + factor(smoker) +
                          factor(region), data = medical_ins)

subsetsummary = summary(health_first_order)$coefficients

summary(health_first_order)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + factor(smoker) +
##     factor(region), data = medical_ins)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11515  -2792  -1011   1360  29843
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -11659.114    679.978 -17.146  < 2e-16 ***
## age                         255.637      8.262  30.939  < 2e-16 ***
## bmi                         329.809     19.848  16.617  < 2e-16 ***
## children                    505.941     95.133   5.318 1.13e-07 ***
## factor(smoker)yes         23970.504    287.479  83.382  < 2e-16 ***
## factor(region)northwest    -331.005    334.306  -0.990  0.32220
## factor(region)southeast   -1078.176    334.360  -3.225  0.00128 **
## factor(region)southwest   -1054.856    333.061  -3.167  0.00156 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6072 on 2764 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7503
## F-statistic:  1190 on 7 and 2764 DF,  p-value: < 2.2e-16
```

A.8 Partial F-test

```
print(anova(health_first_order, health_full_model))
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ age + bmi + children + factor(smoker) + factor(region)
## Model 2: charges ~ age + bmi + children + factor(smoker) + factor(sex) +
```

```
##     factor(region)
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1   2764 1.0192e+11
## 2   2763 1.0191e+11  1   2224676 0.0603  0.806
```

A.9 Interaction Model

```
health_int = lm(charges~(age+bmi+children+factor(smoker)+factor(region))^2,
            data = medical_ins)

summary(health_int)
```

```
##
## Call:
## lm(formula = charges ~ (age + bmi + children + factor(smoker) +
##     factor(region))^2, data = medical_ins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11842.5  -2037.4  -1233.1   -209.6  30112.4
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                            -2.372e+03  1.705e+03  -1.391
## age                                     2.018e+02  3.613e+01   5.586
## bmi                                     5.232e+01  5.604e+01   0.933
## children                                5.898e+02  4.510e+02   1.308
## factor(smoker)yes                      -2.078e+04  1.338e+03 -15.531
## factor(region)northwest                -4.378e+02  1.568e+03  -0.279
## factor(region)southeast                 3.463e+03  1.490e+03   2.325
## factor(region)southwest                 1.238e+02  1.510e+03   0.082
## age:bmi                                 1.180e+00  1.123e+00   1.050
## age:children                           -2.732e+00  5.928e+00  -0.461
## age:factor(smoker)yes                  -2.350e-02  1.662e+01  -0.001
## age:factor(region)northwest             7.072e+00  1.916e+01   0.369
## age:factor(region)southeast             4.658e+01  1.906e+01   2.444
## age:factor(region)southwest             4.454e+01  1.937e+01   2.300
## bmi:children                            6.445e+00  1.322e+01   0.487
## bmi:factor(smoker)yes                   1.470e+03  3.891e+01  37.790
## bmi:factor(region)northwest            -2.031e+01  4.907e+01  -0.414
## bmi:factor(region)southeast            -1.953e+02  4.214e+01  -4.634
## bmi:factor(region)southwest            -1.007e+02  4.659e+01  -2.161
## children:factor(smoker)yes             -3.548e+02  1.982e+02  -1.790
## children:factor(region)northwest        2.324e+02  2.240e+02   1.037
## children:factor(region)southeast       -1.874e+02  2.251e+02  -0.833
```

```
## children:factor(region)southwest          -3.902e+02  2.125e+02  -1.836
## factor(smoker)yes:factor(region)northwest -3.311e+02  6.780e+02  -0.488
## factor(smoker)yes:factor(region)southeast -9.214e+02  6.425e+02  -1.434
## factor(smoker)yes:factor(region)southwest  1.058e+03  6.859e+02   1.542
##                                          Pr(>|t|)
## (Intercept)                                0.1643
## age                                       2.55e-08 ***
## bmi                                        0.3507
## children                                   0.1910
## factor(smoker)yes                          < 2e-16 ***
## factor(region)northwest                    0.7801
## factor(region)southeast                    0.0201 *
## factor(region)southwest                    0.9347
## age:bmi                                    0.2936
## age:children                               0.6449
## age:factor(smoker)yes                      0.9989
## age:factor(region)northwest                0.7121
## age:factor(region)southeast                0.0146 *
## age:factor(region)southwest                0.0215 *
## bmi:children                               0.6260
## bmi:factor(smoker)yes                      < 2e-16 ***
## bmi:factor(region)northwest                0.6790
## bmi:factor(region)southeast               3.75e-06 ***
## bmi:factor(region)southwest                0.0308 *
## children:factor(smoker)yes                 0.0735 .
## children:factor(region)northwest           0.2997
## children:factor(region)southeast           0.4051
## children:factor(region)southwest           0.0664 .
## factor(smoker)yes:factor(region)northwest  0.6253
## factor(smoker)yes:factor(region)southeast  0.1517
## factor(smoker)yes:factor(region)southwest  0.1232
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4830 on 2746 degrees of freedom
## Multiple R-squared:  0.8434, Adjusted R-squared:  0.842
## F-statistic: 591.7 on 25 and 2746 DF,  p-value: < 2.2e-16
```

A.10 Final Interaction Model

```
health_final_int = lm(charges~age+bmi+children+factor(smoker)+factor(region)
  +age*factor(region)+bmi*factor(smoker)+bmi*factor(region), data = medical_ins)
summary(health_final_int)
```

```
##
```

```
## Call:
## lm(formula = charges ~ age + bmi + children + factor(smoker) +
##     factor(region) + age * factor(region) + bmi * factor(smoker) +
##     bmi * factor(region), data = medical_ins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12577.1  -1968.3  -1280.6   -241.8  29932.7
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -3616.125   1065.717  -3.393 0.000701 ***
## age                             233.858     13.663  17.116  < 2e-16 ***
## bmi                             105.483     33.456   3.153 0.001634 **
## children                        509.491     76.011   6.703 2.47e-11 ***
## factor(smoker)yes             -20918.581   1156.713 -18.085  < 2e-16 ***
## factor(region)northwest         -461.785   1545.909  -0.299 0.765181
## factor(region)southeast         2733.186   1446.701   1.889 0.058963 .
## factor(region)southwest         -280.224   1482.198  -0.189 0.850060
## age:factor(region)northwest        8.105     19.161   0.423 0.672328
## age:factor(region)southeast       51.309     18.386   2.791 0.005296 **
## age:factor(region)southwest       41.773     19.266   2.168 0.030224 *
## bmi:factor(smoker)yes           1457.361     36.789  39.614  < 2e-16 ***
## bmi:factor(region)northwest      -13.709     49.062  -0.279 0.779938
## bmi:factor(region)southeast     -191.661     42.140  -4.548 5.65e-06 ***
## bmi:factor(region)southwest      -92.140     46.559  -1.979 0.047919 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4840 on 2757 degrees of freedom
## Multiple R-squared:  0.8421, Adjusted R-squared:  0.8413
## F-statistic:  1051 on 14 and 2757 DF,  p-value: < 2.2e-16
```

A.11 Higher Order Model

```
higher_order = lm(charges~age + I(bmi^2)+I(bmi^3)+I(bmi^4)+ bmi
                  +children+factor(smoker)+factor(region)+age*factor(region)
                  +bmi*factor(smoker)+bmi*factor(region),
                data = medical_ins)
summary(higher_order)
```

```
##
## Call:
## lm(formula = charges ~ age + I(bmi^2) + I(bmi^3) + I(bmi^4) +
##     bmi + children + factor(smoker) + factor(region) + age *
```

```
##     factor(region) + bmi * factor(smoker) + bmi * factor(region),
##     data = medical_ins)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9505  -2100  -1227    -77  29759
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   9.794e+04  1.989e+04   4.924 8.96e-07 ***
## age                           2.336e+02  1.354e+01  17.251  < 2e-16 ***
## I(bmi^2)                      5.684e+02  1.226e+02   4.635 3.74e-06 ***
## I(bmi^3)                     -1.076e+01  2.515e+00  -4.278 1.95e-05 ***
## I(bmi^4)                      7.256e-02  1.885e-02   3.849 0.000121 ***
## bmi                          -1.261e+04  2.587e+03  -4.873 1.16e-06 ***
## children                      5.279e+02  7.531e+01   7.010 2.98e-12 ***
## factor(smoker)yes            -2.086e+04  1.147e+03 -18.195  < 2e-16 ***
## factor(region)northwest       3.807e+02  1.535e+03   0.248 0.804188
## factor(region)southeast       1.484e+03  1.535e+03   0.967 0.333733
## factor(region)southwest       2.332e+01  1.480e+03   0.016 0.987429
## age:factor(region)northwest   1.138e+01  1.898e+01   0.599 0.548952
## age:factor(region)southeast   5.007e+01  1.825e+01   2.743 0.006133 **
## age:factor(region)southwest   4.459e+01  1.909e+01   2.337 0.019532 *
## bmi:factor(smoker)yes         1.456e+03  3.647e+01  39.916  < 2e-16 ***
## bmi:factor(region)northwest  -4.756e+01  4.882e+01  -0.974 0.330047
## bmi:factor(region)southeast  -1.483e+02  4.527e+01  -3.275 0.001071 **
## bmi:factor(region)southwest  -1.061e+02  4.648e+01  -2.284 0.022473 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4793 on 2754 degrees of freedom
## Multiple R-squared:  0.8454, Adjusted R-squared:  0.8444
## F-statistic: 885.7 on 17 and 2754 DF,  p-value: < 2.2e-16
```

```r
# for pairwise ggpairs plot
numerical_values = data.frame(medical_ins$charges, medical_ins$age,
                        medical_ins$bmi, medical_ins$children)
```

A.12 All Possible Regressions

```r
# largest additive model to start
n_fullmodel = lm(charges~age+sex+bmi+children+smoker+region, data=medical_ins)

# calculate best subsets
```

```
n_bestsubsets = ols_step_best_subset(n_fullmodel, details=FALSE)
n_bestsubsets
```

```
##              Best Subsets Regression
## ------------------------------------------------
## Model Index    Predictors
## ------------------------------------------------
##      1         smoker
##      2         age smoker
##      3         age bmi smoker
##      4         age bmi children smoker
##      5         age bmi children smoker region
##      6         age sex bmi children smoker region
## ------------------------------------------------
##
##                                                                                                    Subsets Regression Summary
## ------------------------------------------------------------------------------------------------------------------------------
##                         Adj.        Pred
## Model    R-Square    R-Square    R-Square      C(p)          AIC           SBIC          SBC
## ------------------------------------------------------------------------------------------------------------------------------
##   1        0.6222      0.6220      0.6214    1423.3204    57316.1052    49447.9261    57333
##   2        0.7227      0.7225      0.7219     310.0106    56460.4974    48593.2897    56484
##   3        0.7470      0.7467      0.7461      42.5289    56208.3465    48341.6525    56237
##   4        0.7496      0.7492      0.7485      16.3342    56182.3578    48315.7398    56217
##   5        0.7509      0.7503      0.7493       7.0603    56173.0767    48302.5203    56226
##   6        0.7509      0.7502      0.7491       9.0000    56175.0162    48304.4666    56234
## ------------------------------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

```
# chosen model through best subsets method
all_possible_regressions = lm(charges~age+bmi+children
                              +factor(smoker)+factor(region),
                              data = medical_ins)
summary(all_possible_regressions)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + factor(smoker) +
##     factor(region), data = medical_ins)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -11515  -2792  -1011   1360  29843
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -11659.114    679.978 -17.146  < 2e-16 ***
## age                         255.637      8.262  30.939  < 2e-16 ***
## bmi                         329.809     19.848  16.617  < 2e-16 ***
## children                    505.941     95.133   5.318 1.13e-07 ***
## factor(smoker)yes         23970.504    287.479  83.382  < 2e-16 ***
## factor(region)northwest    -331.005    334.306  -0.990  0.32220
## factor(region)southeast   -1078.176    334.360  -3.225  0.00128 **
## factor(region)southwest   -1054.856    333.061  -3.167  0.00156 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6072 on 2764 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7503
## F-statistic:  1190 on 7 and 2764 DF,  p-value: < 2.2e-16
```

A.13 BP Test

```
bptest(higher_order)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  higher_order
## BP = 38.809, df = 17, p-value = 0.0019
```

A.14 Shapiro-Wilk Test

```
shapiro.test(residuals(higher_order))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(higher_order)
## W = 0.68032, p-value < 2.2e-16
```

A.15 Model Selection

```
# code for model statistics table
Models = c("Full Model",
           "Reduced Model",
           "Interaction Model",
           "Higher Order Model",
           "All Possible Regressions Model")
RSE = c(summary(health_full_model)$sigma,
              summary(health_first_order)$sigma,
              summary(health_final_int)$sigma,
              summary(higher_order)$sigma,
              summary(all_possible_regressions)$sigma)
R_Squared = c(summary(health_full_model)$r.squared,
              summary(health_first_order)$r.squared,
              summary(health_final_int)$r.squared,
              summary(higher_order)$r.squared,
              summary(all_possible_regressions)$r.squared)
Adjusted_R_Squared = c(summary(health_full_model)$adj.r.squared,
                       summary(health_first_order)$adj.r.squared,
                       summary(health_final_int)$adj.r.squared,
                       summary(higher_order)$adj.r.squared,
                       summary(all_possible_regressions)$adj.r.squared)
AIC = c(AIC(health_full_model),
        AIC(health_first_order),
        AIC(health_final_int),
        AIC(higher_order),
        AIC(all_possible_regressions))
BIC = c(BIC(health_full_model),
        BIC(health_first_order),
        BIC(health_final_int),
        BIC(higher_order),
        BIC(all_possible_regressions))

model_stats = data.frame(Models,
                         RSE,
                         R_Squared,
                         Adjusted_R_Squared,
                         AIC,
                         BIC)
```

A.16 Prediction

```r
newdata = data.frame(
  age=c(26, 38, 50, 39, 18),
  bmi=c(28, 24, 22, 30.45, 20),
  children=c(0, 2, 3, 1, 0),
  smoker=c('no', 'yes', 'no', 'no', 'yes'),
  region=c('northwest','northeast','southwest','southeast', 'northwest'))

# prediction value, lower and upper bounds
predout = predict(higher_order, newdata, interval='confidence', level = 0.95)

# combine input data with output prediction in one table
predtotal = cbind(newdata, predout)
```

Appendix B

Tables

Table B.1: A table showing the first few rows of the dataset

age	sex	bmi	children	smoker	region	charges
19	female	27.900	0	yes	southwest	16884.924
18	male	33.770	1	no	southeast	1725.552
28	male	33.000	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.471
32	male	28.880	0	no	northwest	3866.855
31	female	25.740	0	no	southeast	3756.622
46	female	33.440	1	no	southeast	8240.590
37	female	27.740	3	no	northwest	7281.506
37	male	29.830	2	no	northeast	6406.411
60	female	25.840	0	no	northwest	28923.137

Table B.2: A table of five-number summaries along with the mean for each quantitative predictor variable.

age	bmi	children	charges
Min. :18.00	Min. :15.96	Min. :0.000	Min. : 1122
1st Qu.:26.00	1st Qu.:26.22	1st Qu.:0.000	1st Qu.: 4688
Median :39.00	Median :30.45	Median :1.000	Median : 9333
Mean :39.11	Mean :30.70	Mean :1.102	Mean :13261
3rd Qu.:51.00	3rd Qu.:34.77	3rd Qu.:2.000	3rd Qu.:16578
Max. :64.00	Max. :53.13	Max. :5.000	Max. :63770

Table B.3: A table of counts of each value for qualitative predictors.

sex	smoker	region
female:1366	no :2208	northeast:658
male :1406	yes: 564	northwest:664
		southeast:766
		southwest:684

Table B.4: Table of the best possible model for n=1 through n=6 predictors, and their corresponding statistical criteria.

n	predictors	rsquare	adjr	cp	aic
1	smoker	0.6221792	0.6220428	1423.320356	57316.11
2	age smoker	0.7227172	0.7225170	310.010615	56460.50
3	age bmi smoker	0.7470093	0.7467351	42.528929	56208.35
4	age bmi children smoker	0.7495509	0.7491888	16.334193	56182.36
5	age bmi children smoker region	0.7509277	0.7502969	7.060314	56173.08
6	age sex bmi children smoker region	0.7509332	0.7502120	9.000000	56175.02

Table B.5: Linear model parameters based on best subsets method.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-11659.1136	679.978084	-17.1463080	0.0000000
age	255.6372	8.262486	30.9394951	0.0000000
bmi	329.8092	19.848151	16.6166209	0.0000000
children	505.9411	95.133340	5.3182314	0.0000001
factor(smoker)yes	23970.5036	287.479368	83.3816486	0.0000000
factor(region)northwest	-331.0052	334.305893	-0.9901267	0.3221989
factor(region)southeast	-1078.1757	334.360112	-3.2245942	0.0012762
factor(region)southwest	-1054.8559	333.060750	-3.1671576	0.0015561

Table B.6: Comparison of model statistics for all models.

Models	RSE	R_Squared	Adjusted_R_Squared	AIC	BIC
Full Model	6073.308	0.7509332	0.7502120	56175.02	56234.29
Reduced Model	6072.275	0.7509277	0.7502969	56173.08	56226.42
Interaction Model	4840.426	0.8421341	0.8413324	54923.05	55017.89
Higher Order Model	4793.143	0.8453717	0.8444172	54871.61	54984.23
All Possible Regressions Model	6072.275	0.7509277	0.7502969	56173.08	56226.42

Table B.7: Prediction input and output

age	bmi	children	smoker	region	fit	lwr	upr
26	28.00	0	no	northwest	4481.983	3924.314	5039.653
38	24.00	2	yes	northeast	22176.990	21426.319	22927.660
50	22.00	3	no	southwest	11362.784	10505.915	12219.653
39	30.45	1	no	southeast	8343.260	7913.567	8772.953
18	20.00	0	yes	northwest	10835.258	9627.685	12042.831

Appendix C

Graphs

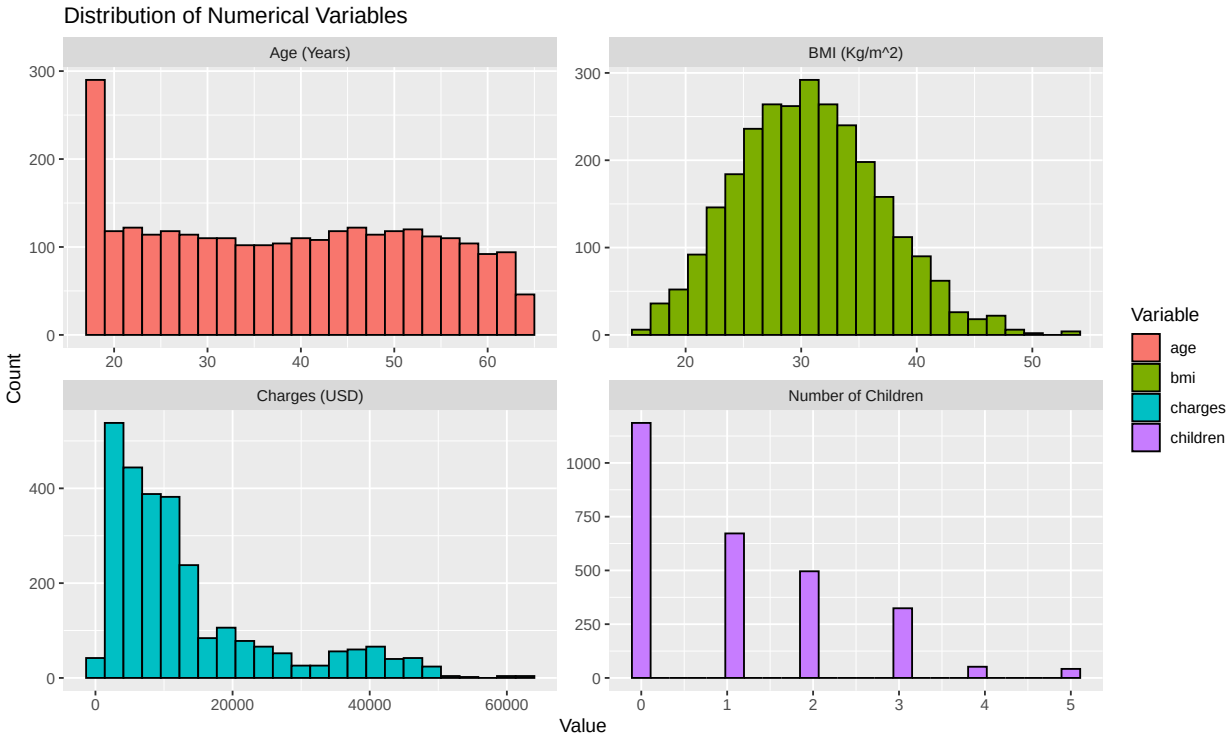

Figure C.1: A set of plots showing the distributions of the numerical variables from the dataset.

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

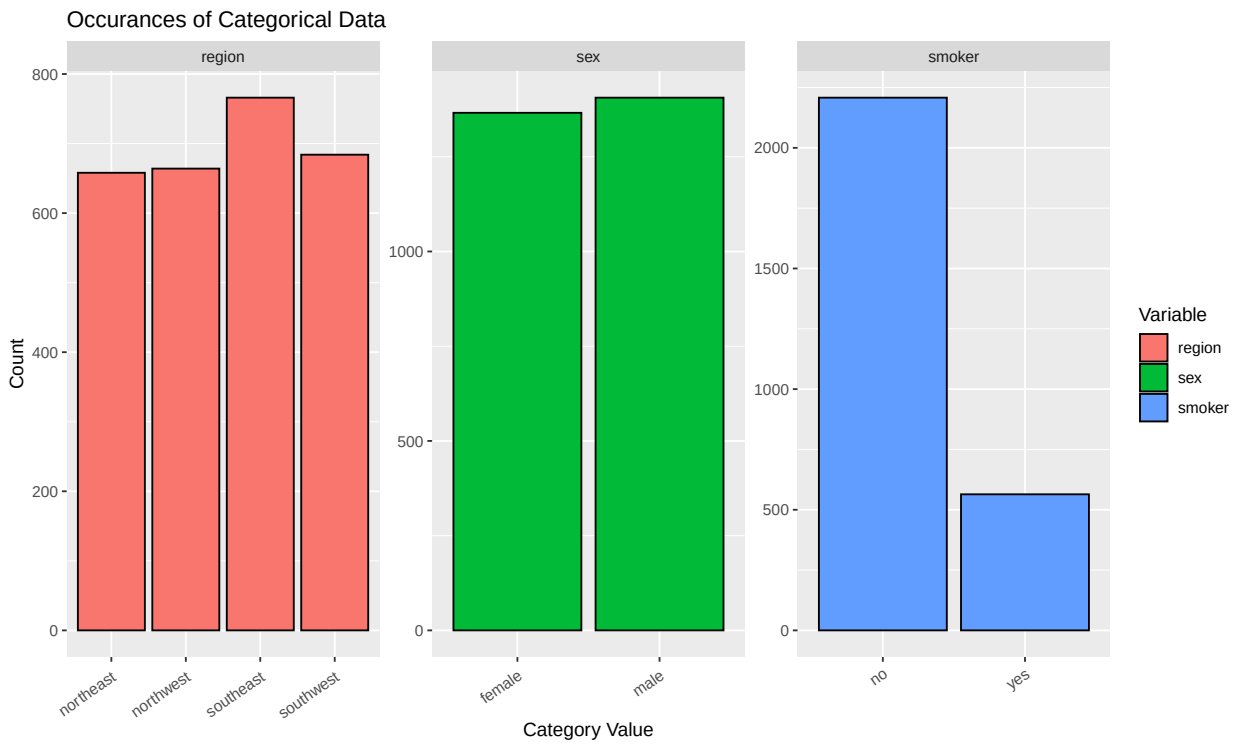

Figure C.2: Bar charts showing the distribution of the occurrences of each possible value, over the three categorical predictors.

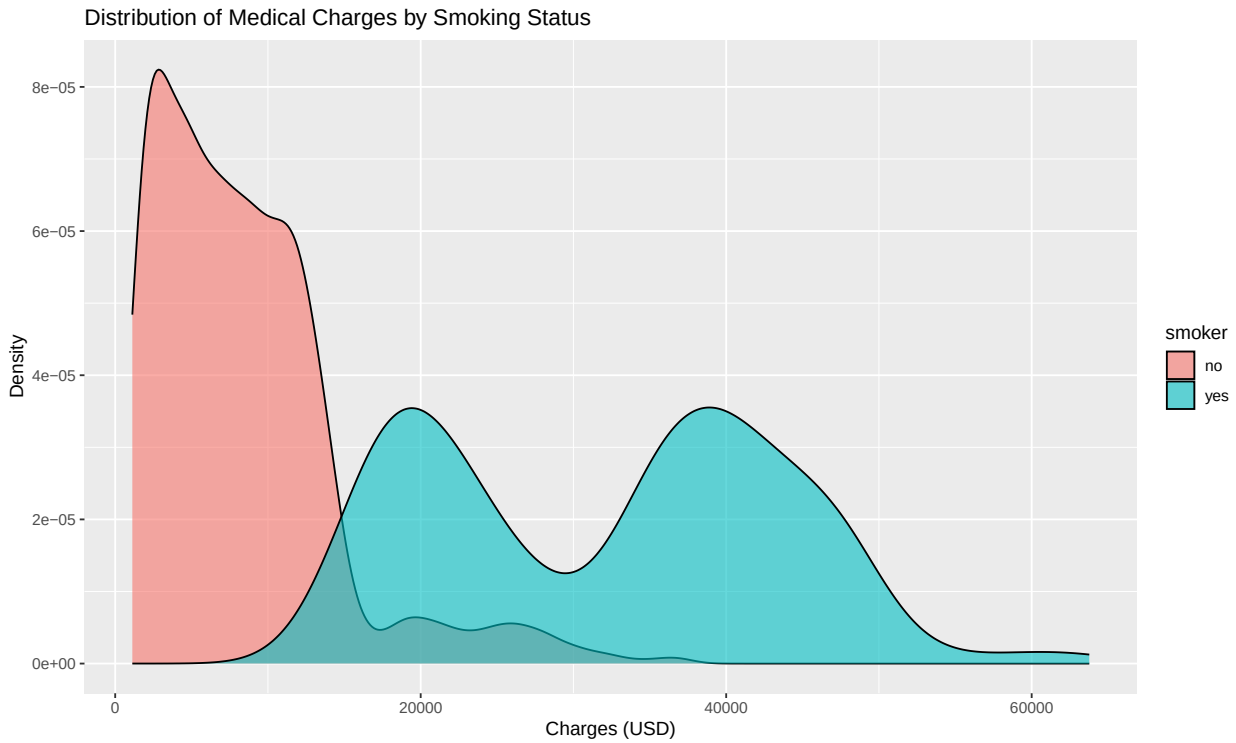

Figure C.3: Split distribution of medical charges in USD, for smokers and nonsmokers.

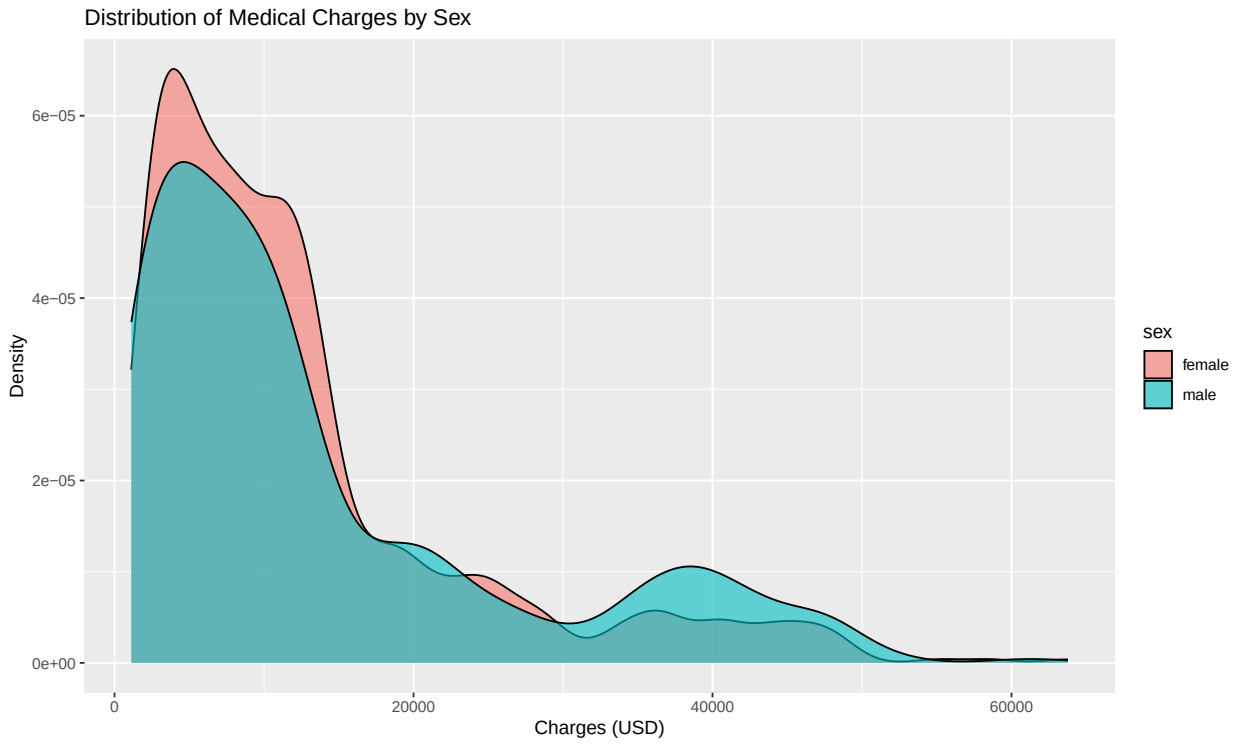

Figure C.4: Distribution of medical charges in USD, split by sex.

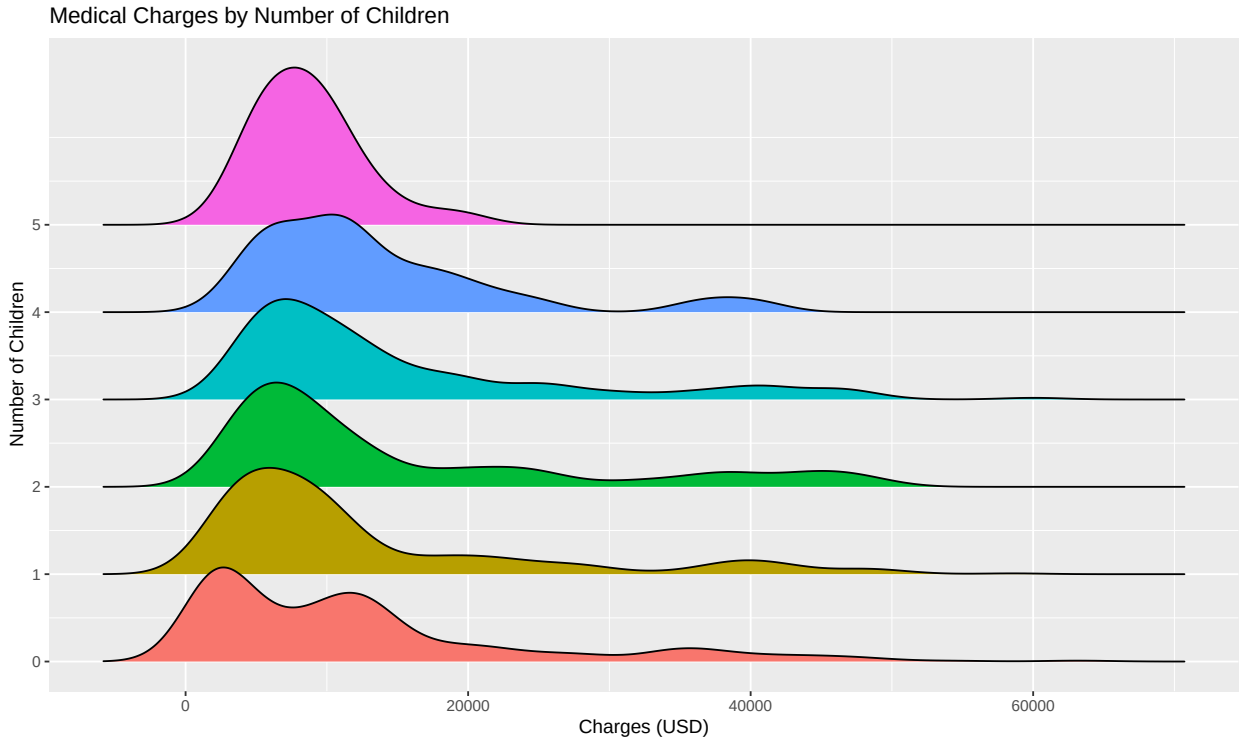

Figure C.5: Ridgeplot showing the distribution of medical charges over varying number of children.

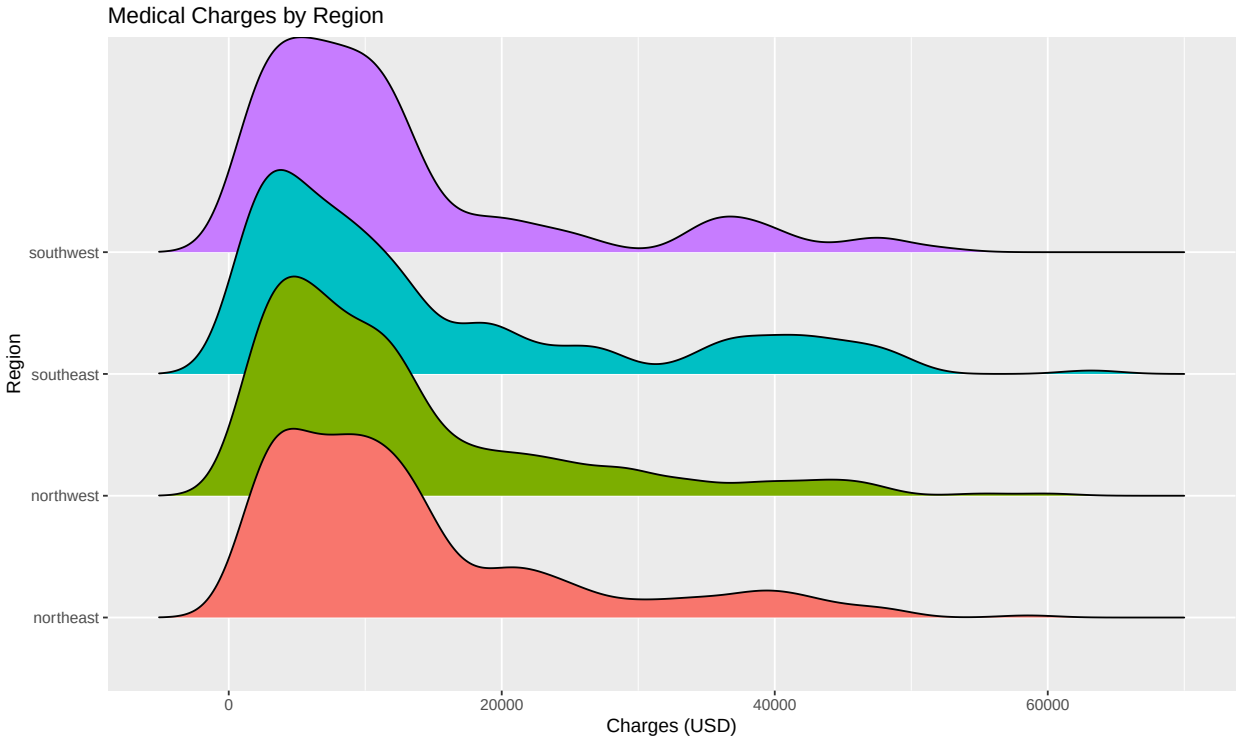

Figure C.6: Ridgeplot showing distribution of medical charges by region.

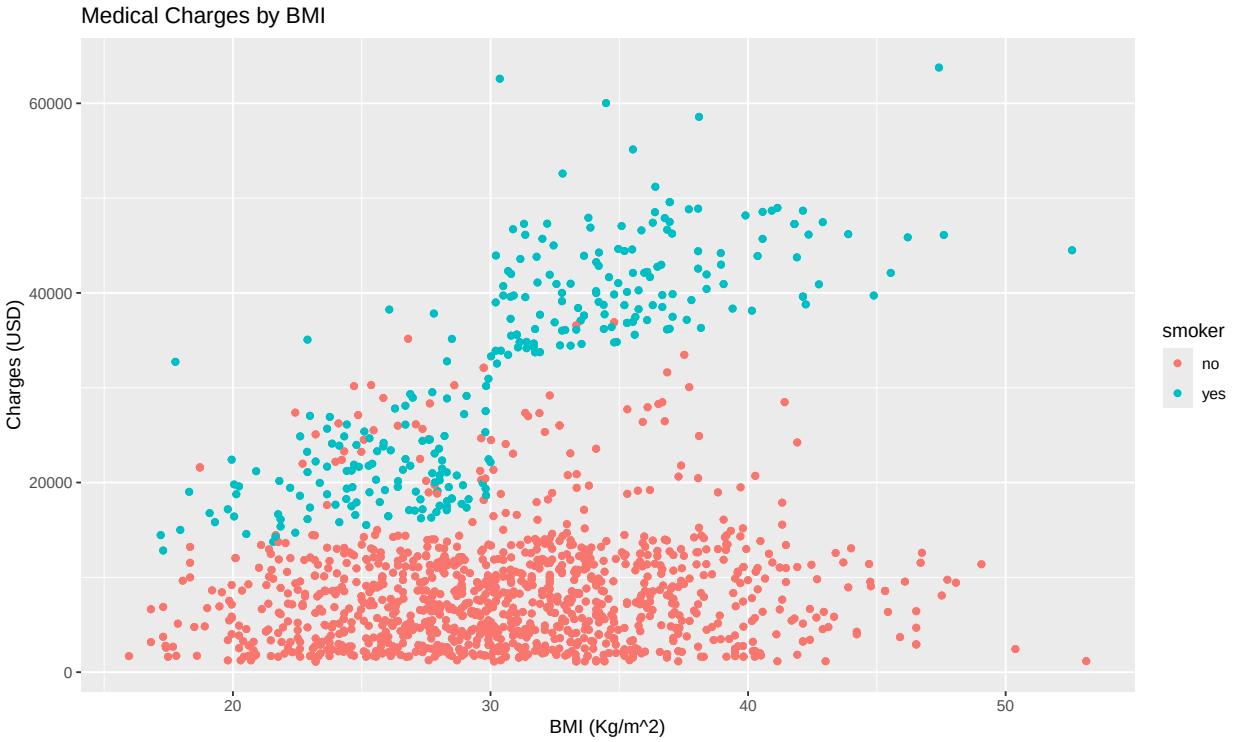

Figure C.7: Scatterplot of medical charges by BMI and smoking status.

Figure C.8: Scatter plot of medical charges by age and smoking status.

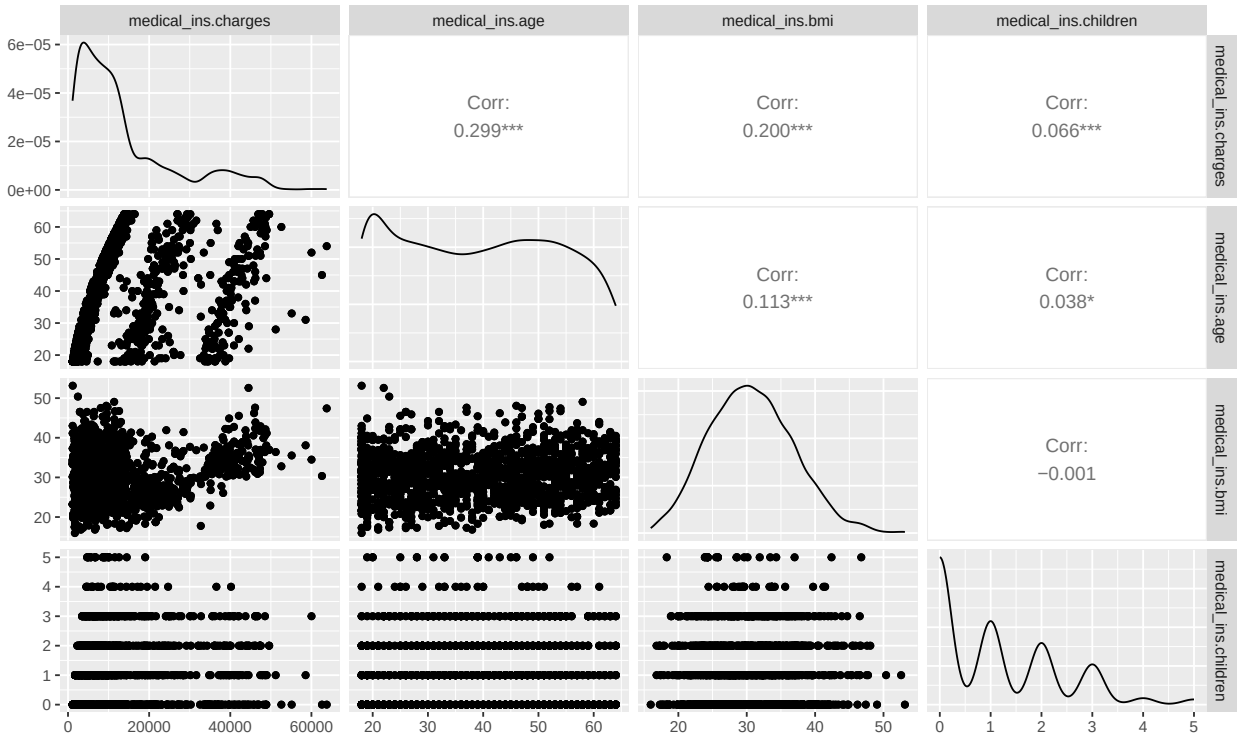

Figure C.9: Pairwise scatter plot displaying the correlation between each relevant variable for higher order term identification.

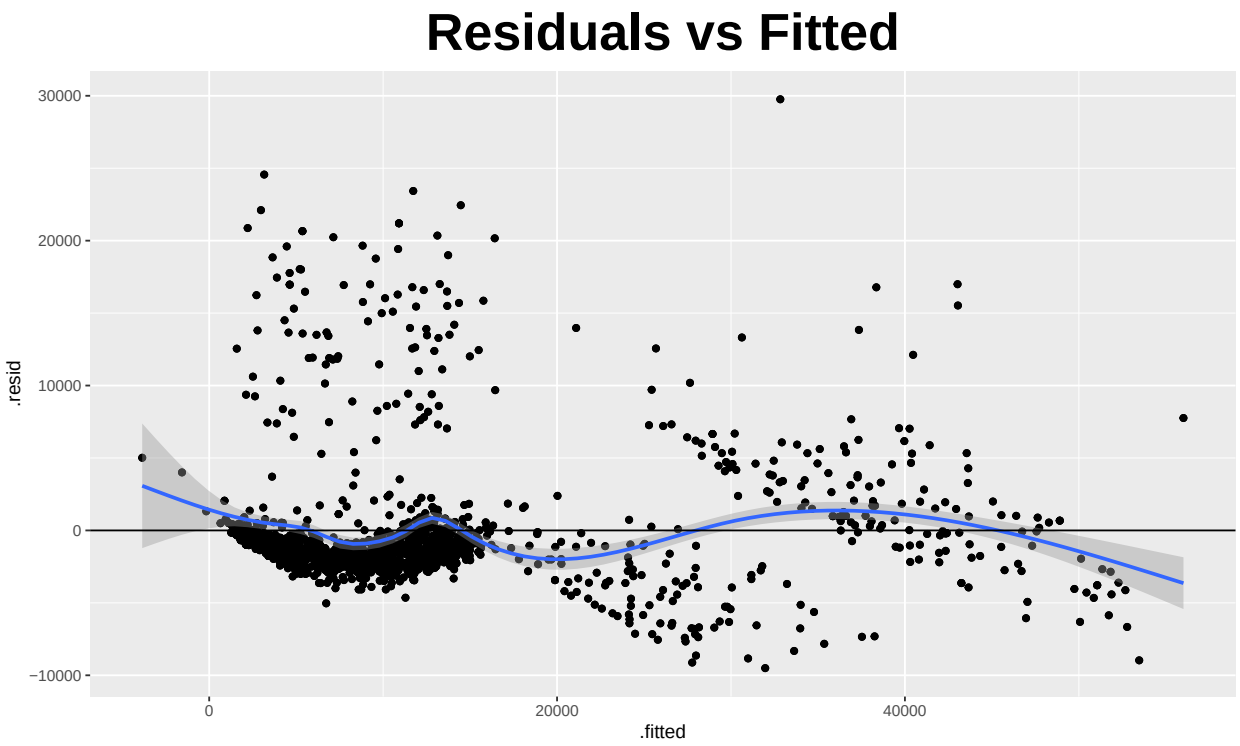

Figure C.10: Residuals vs Fitted Plot for the higher order model for checking independency and homoscedasticity assumptions.

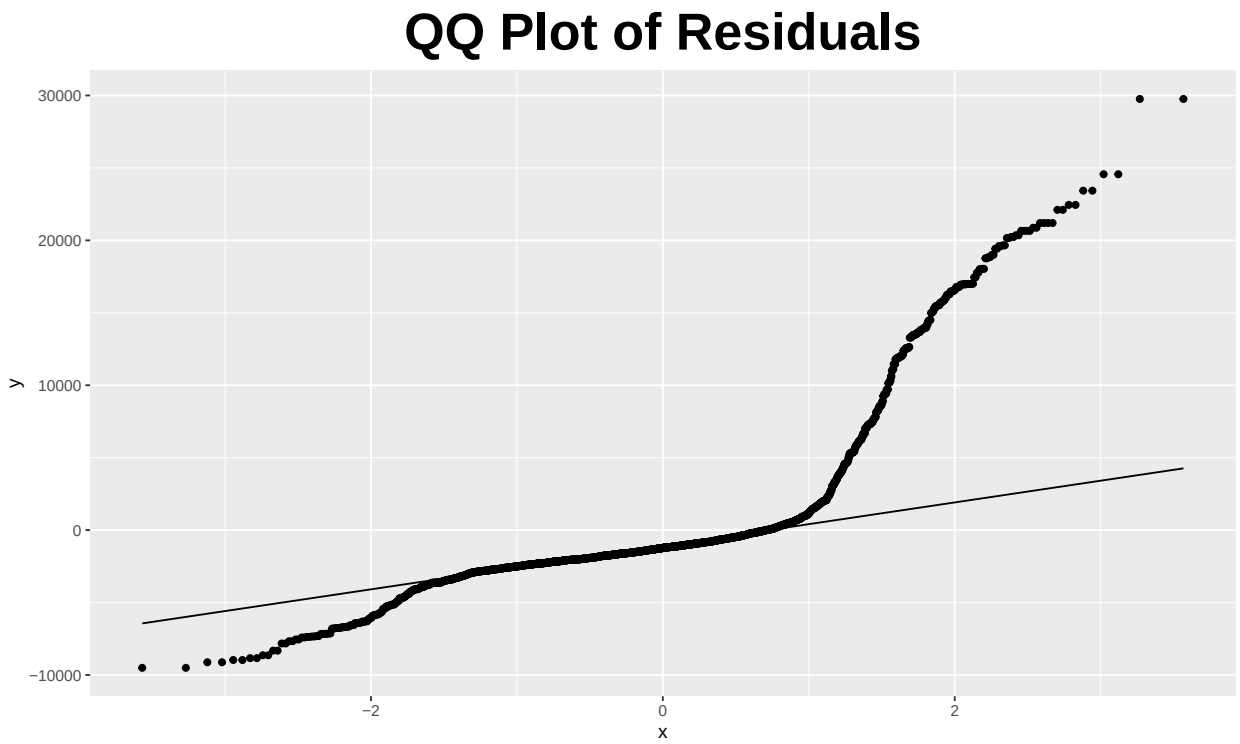

Figure C.11: QQ Plot of Residuals for the higher order model to check normality assumption.

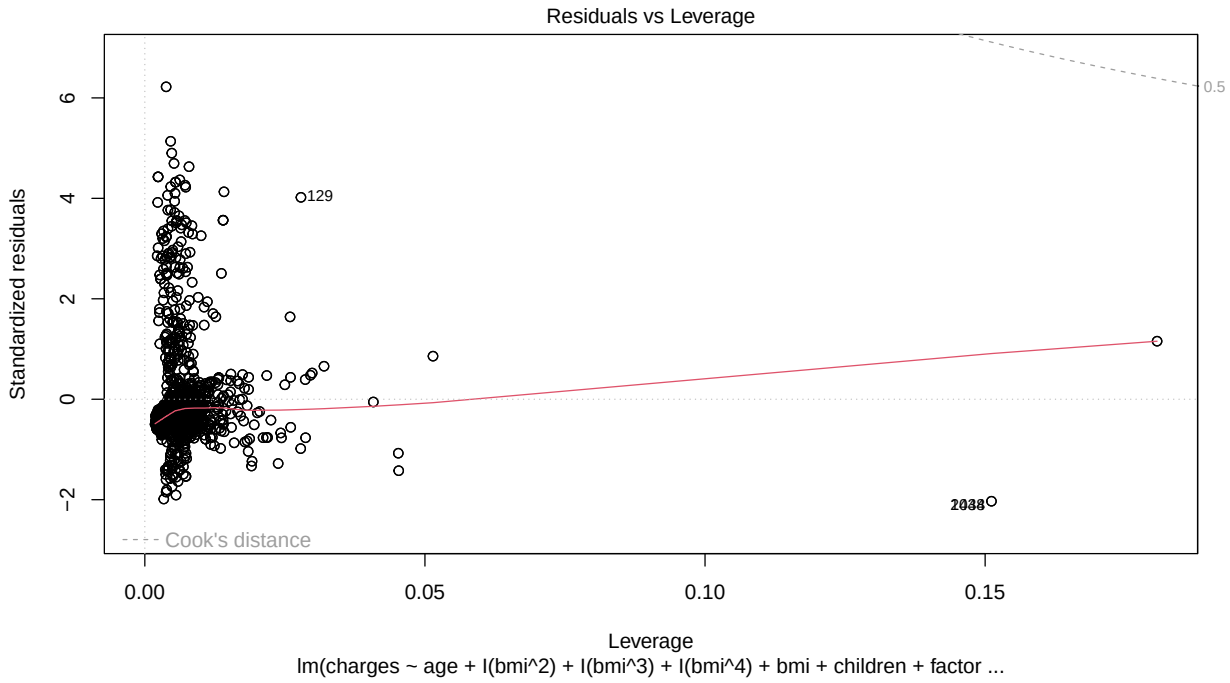

Figure C.12: Residuals vs Leverage Plot for the analysis of outliers from the higher order model.

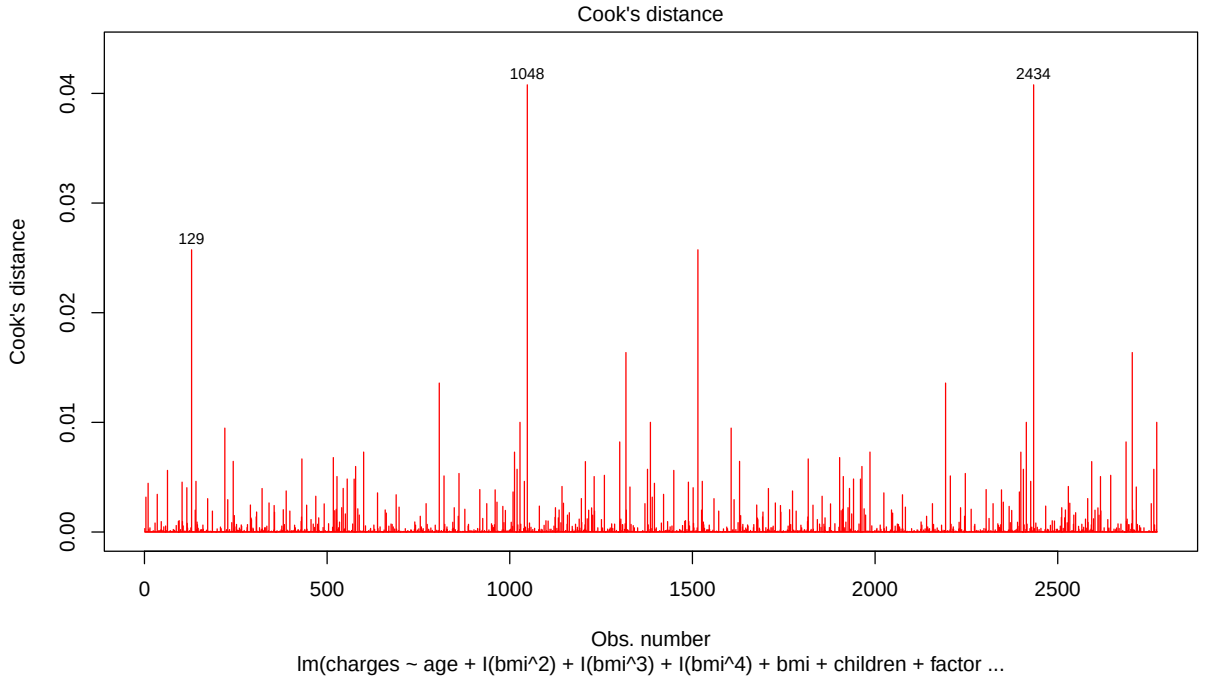

Figure C.13: Cook's Distance Plot for assessment of outliers from the higher order model.