UNIVERSITY OF CALGARY

Multiple Regression Analysis: Insurance Charges

DATA 603: Statistical Modelling with Data

Philip P. Owusu (30137210)

April 4, 2021

Contents

I Introduction

Universal health care is provided as a public service in many countries. In others, access to the health care system is reliant on the purchase of health insurance. One of such countries is the United States of America. The question of whether health care has an impact on health is a contested topic. Researchers confront the challenge of separating observable factors from the relationship between good health and health insurance.[1] While that relationship is more difficult to determine, the correlations between insurance charges and various personal attributes can be investigated by applying statistical techniques to collected insurance data. It is important to investigate this relationship as many Americans choose to forgo insurance in favor of saving money (**Figure 1**).

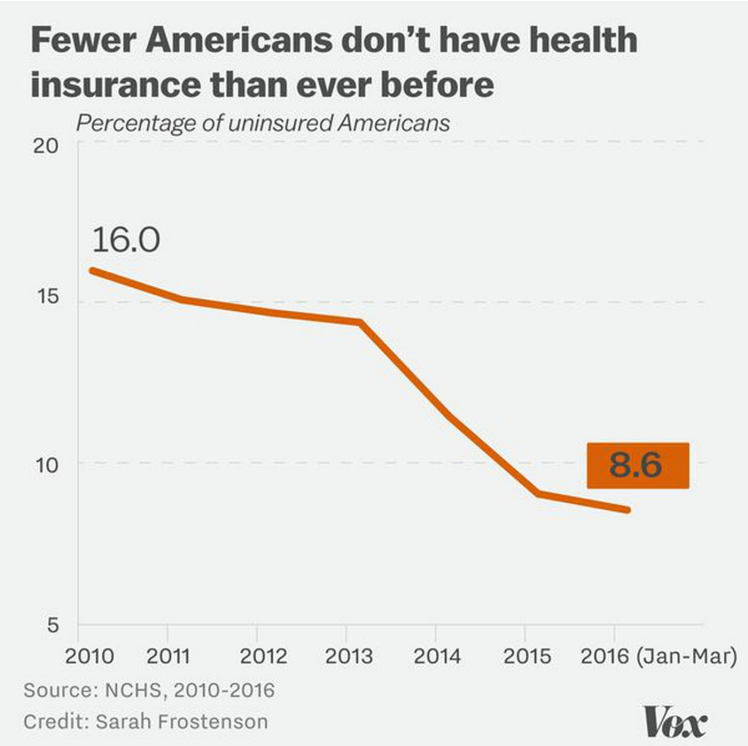

Figure 1: Graph displaying the decreasing number of Americans without health insurance.

Insurance companies increase rates depending on a variety of features. Variables such as age, indulgence in unhealthy activities (e.g., smoking), chronic health conditions and more, have an effect on how much the insurance company will charge an individual.[2] While companies often tell clients what factors will have an impact on charges increasing, analysis of collected data can tell us which factors are statistically significant. By running statistical tests on a data set of insurance charges and various factors, it can be determined what has the largest influence in increasing insurance charges. It is expected that factors related to the buyers overall health will have the most significant factors. Using the statistical software R, more information can be interpreted from previous insurance charges and potential buyers are more informed on how their identities and lifestyles will influence insurance rates.

II Methodology

A Data Source

The data was collected in a CSV format from Kaggle.com, which is an open-source data website. The data source does not specify the year that the data was acquired, only that there are 1,338 rows of insured data.

The data set does not have any missing or undefined values and was read into R using the *read.csv* function. The data required for the regression analysis included the dependent variable *Charges* in U.S. dollars ($), as a function of three quantitative and three qualitative variables: *Age, Sex, Smoker, Body Mass Index (BMI), Children, and Region.*

B Variable Explanations and Data Assumptions

The data set contains 1,338 rows of insured data, where the insurance charges (dependent variable) are given against the six independent variables. The three qualitative variables are *Sex* (Male/Female), *Smoker* (Yes/No), and *Region* (Southwest/Southeast/Northwest/Northeast). The three quantitative variables are *Age, Children, and BMI.* The *Age* variable is continuous and ranges from ages 18 to 64. The discrete variable, amount of *Children*, ranges from 0 to 5. Lastly, the *BMI* is a continuous variable that is a measure of body fat based on height and weight.

Below is a complete list of variables used in the modelling process:

1. *Charges*: The individual medical costs billed by health insurance in U.S. dollars ($) - **Dependent Variable**
2. *Age*: The age of primary beneficiary (Years) - **Independent Variable**
3. *Sex*: Insurance contractor gender (Male/Female) - **Independent Variable**
4. *BMI*: Body mass index of the individual, measured in kg/m^2 - **Independent Variable**
5. *Children*: Number of children covered by health insurance (or number of dependants) - **Independent Variable**
6. *Smoker*: Qualitative variable indicating whether the individual smokes or not (Yes/No) - **Independent Variable**
7. *Region*: The beneficiary's residential area in the U.S. (Southwest/Southeast/Northwest/Northeast) - **Independent Variable**

This data set can help in understanding risk underwriting in Health Insurance, the relationships of various attributes of the insured and their affect on the insurance premium. An insurer's profitability depends on how well it can reduce the costs associated with managing claims.[3] The amount charged for providing coverage is a critical aspect of the underwriting process and the premium must be sufficient to cover the expected claims. That being said, the underlying assumption of the study is that the variables most closely related to the overall health of the beneficiary will be most significant when forecasting the insurance charges.

C Modelling Plan

The strategy for selecting a model is to first identify the response y and the set of independent variables $x_1, x_2, ..., x_p$. This will include only the main effects of both quantitative (*Age, BMI, Children*) and qualitative (*Sex, Smoker, Region*) variables. Once that is entered into R, the **Stepwise Regression Procedure** will be applied to recommend a model of main effects.

To check for significant higher-order terms and interactions, individual t-tests will be conducted. If there are any significant higher-order or interaction terms, they will be added to the main effects model. A **Partial F-test** will then be conducted to compare the full model and reduced model. The final step will be to conduct diagnostics tests, which will include testing the following assumptions:

1. Linearity Assumption: Review the Residual vs Fitted Plot
2. Independence Assumption
3. Normality Assumption: Using Q-Q plot and Shapiro-Wilk test
4. Equal Variance Assumption (homoscedasticity) - Using Residual vs Fitted and Scale-Location plots, as well as Breusch-Pagan test
5. Multicollinearity: Using Variance Inflation Factors (VIF)
6. Outliers: Using Leverage and Cook's distance

Improvements will be made to the model if it does not meet any of these assumptions. Once the final model is acquired, it will be used to predict Insurance Charges in the U.S.

III Results

A Variable Selection Procedure

To begin building our model, a first-order model was created using including all possible variables.

$$\widehat{Charge} = \hat{\beta}_0 + \hat{\beta}_1 Age + \hat{\beta}_2 BMI + \hat{\beta}_3 Children + \hat{\beta}_4 Sex + \hat{\beta}_5 Smoker + \hat{\beta}_6 Region$$

Stepwise Regression was used to determine the best first-order model with addition and removal thresholds of $p = 0.05$ and $p = 0.1$, respectively. Although *Region* was significant at the $\alpha = 0.05$ level with an individual t-test, the resulting best model did not include this variable. The variables added in each stage of the stepwise regression are shown in **Figure 2**.

Step	Smoker p-value	Age p-value	BMI p-value	Children p-value
Stage 1	0.000			
Stage 2	0.000	0.000		
Stage 3	0.000	0.000	0.000	
Stage 4	0.000	0.000	0.000	0.001

Figure 2: Table of p-values as variables are added in each stage of Stepwise Regression.

i Individual Coefficients Test (t-test)

To make sure all the variables in the first-order model are significant, individual coefficients tests were performed with the following hypotheses:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq (i = 1, 2, ..., p)$$

Using a significance level of $\alpha = 0.05$, the results in **Table 1** indicate that the null hypothesis should be rejected for all four chosen variables. This provides evidence that the age, BMI, smoker status and number of children a person has are all significantly influential on the amount of an insurance charge. After checking the individual significance of each variable, the first-order model can be stated as follows:

$$\widehat{Charge} = \hat{\beta}_0 + \hat{\beta}_1 Age + \hat{\beta}_2 BMI + \hat{\beta}_3 Children + \hat{\beta}_4 Smoker$$

Variable	t-value	p-value
Smoker	57.904	<2e-16
Age	21.675	<2e-16
BMI	11.756	<2e-16
Children	3.436	0.000608

Table 1: Individual t-test results for the stepwise regression model.

ii Interaction Terms

Next, the variables in the first-order model were checked for interaction terms to see if any were significant.

Variable	t-value	p-value
Smoker*Age	-0.038	0.969
Smoker*BMI	27.029	$<$ 2e-16
Smoker*Children	-1.358	0.175
Age*BMI	1.225	0.221
Age*Children	0.141	0.888
BMI*Children	-0.286	0.775

Table 2: Individual t-test results for interaction terms.

Using the same individual t-test hypotheses as above, the only interaction term to be significant at the $\alpha = 0.05$ level is $Smoker * BMI$ (**Table 2**). This term is added to the first-order model, giving the model below:

$$\widehat{Charge} = \hat{\beta}_0 + \hat{\beta}_1 Age + \hat{\beta}_2 BMI + \hat{\beta}_3 Children + \hat{\beta}_4 Smoker + \hat{\beta}_5 Smoker * BMI$$

iii Higher-Order Terms

After determining the best first-order model, the higher-order terms can be checked for significance. First, pairwise plots were visually inspected to look for relationships between variables (**Figure 3**). This was done using the *ggpairs* function. The non-linear relationship between $Charges$ and Age suggests that a higher order term may need to be added to the model.

Adding a quadratic term for Age resulted in a better fitting model than the interaction model. The adjusted R-squared increased to 0.8407 and RMSE decreased to 4834 on 1331 degrees of freedom. The F-statistic for the quadratic model is 1177 on 6 and 1331 degrees of freedom. Adding a cubic term resulted in a slightly worse model with lower R^2_{adj} and higher RMSE therefore the following model is taken to be the best at this point:

$$\widehat{Charge} = \hat{\beta}_0 + \hat{\beta}_1 Age + \hat{\beta}_1 Age^2 + \hat{\beta}_2 BMI + \hat{\beta}_3 Children + \hat{\beta}_4 Smoker + \hat{\beta}_5 Smoker * BMI$$

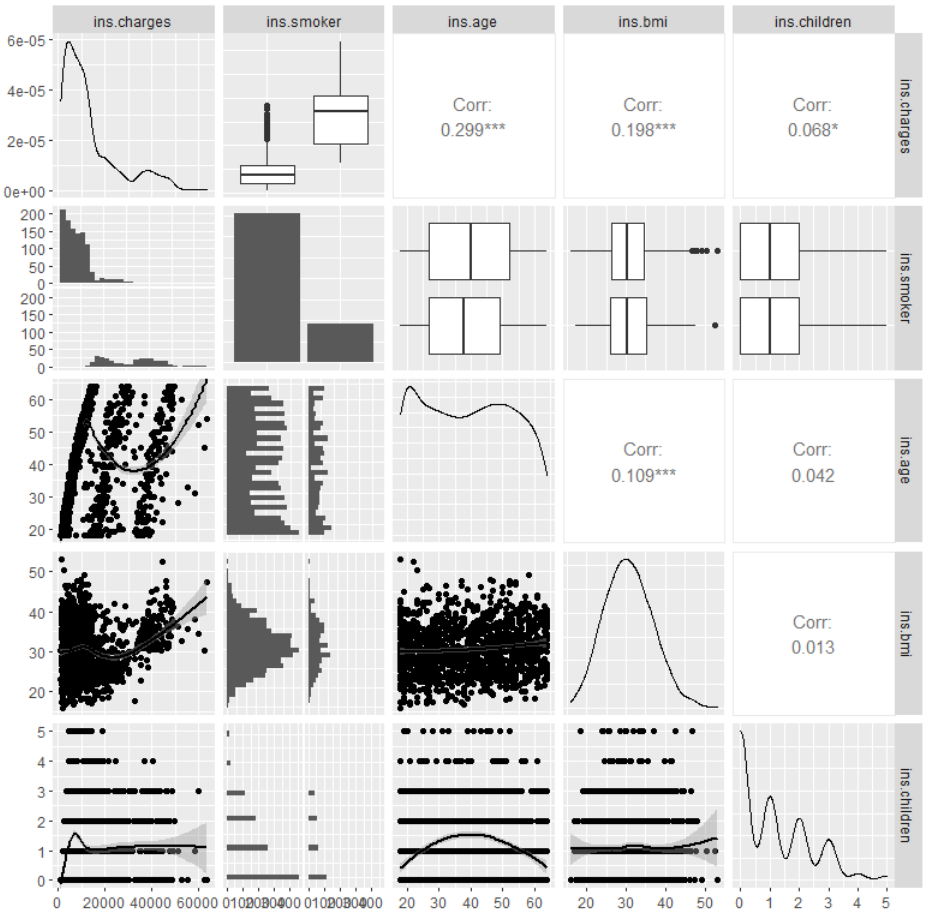

Figure 3: Pairwise plots of significant first order variables.

iv Partial F-test

A partial F-test can be performed to ensure that the interaction and higher order terms should be included in the full model. Two tests were performed, one dropping the interaction term and one dropping the higher order term, to ensure the full model is the best. The hypotheses for a partial F-test are as follows:

$$H_0 : \beta_{p-1+1} = \beta_{p-2+1} = \beta_p = 0$$
$$H_a : \text{at least one } \beta_i \neq 0$$

When the interaction term is dropped from the model, the output gives Fcal=745.82 with df=1,1331 (p-value<2.2e-16< α=0.05), indicating that we should clearly reject the null hypothesis (**Figure 4**). The null hypothesis is also rejected when the quadratic term is dropped from the model, with an output of Fcal=21.407 with df=1,1331 (p-value<2.2e-16< α=0.05) (**Figure 5**).

Source of Variation	Df	Sum of Squares	Mean Square	F-Statistic
Regression	1	1.7426e10	1.7426e10	745.82
Residual	1331	3.11e10	23365890	
Total	1332	4.8526e10		

Figure 4: ANOVA table for full model vs. model without interaction term.

Source of Variation	Df	Sum of Squares	Mean Square	F-Statistic
Regression	1	500194672	500194672	21.407
Residual	1331	3.11e10	23365890	
Total	1332	3.16e10		

Figure 5: ANOVA table for full model vs. model without quadratic term.

The results of the partial F-test confirm that neither the chosen interaction term or quadratic term should be dropped and the full model is the best.

B Multiple Regression Assumptions

i Linearity Assumption

The linearity assumption is checked by analyzing the residuals vs fitted plot. Ideally, the residual vs fitted plot will have no discernible pattern. If a pattern is present there may be a problem with the linearity assumption. The residual vs fitted plot in **Figure 6** shows clumping of data points however the best fit line is relatively flat with no obvious pattern to it. This indicates that the linearity assumption is met.

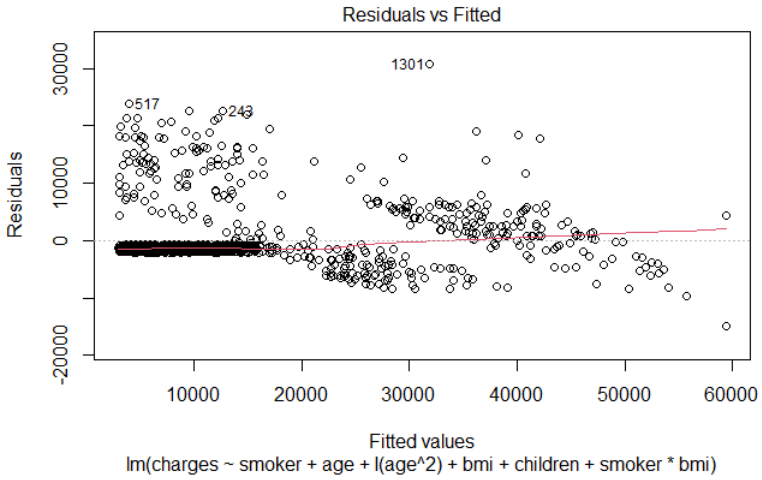

Figure 6: Residuals vs Fitted plot for the full model.

ii Independence Assumption

It is very important when making a linear model that the error terms are uncorrelated. Another way to say this is that the error terms must be mutually independent; if the errors are correlated, the independence assumption is not met. It is most common for the independence assumption to be violated with time series data where you have multiple data points for the same subject. The insurance data set used is this project is not time series data and can be assured as independent.

iii Normality Assumption

A multiple linear regression analysis requires that the errors between the observed and predicted values be normally distributed. This assumption can be checked either visually, using a Q-Q plot (**Figure 7**), normal probability plot or a histogram of the residuals (**Figure 8**), or mathematically, using the Shapiro-Wilk test.

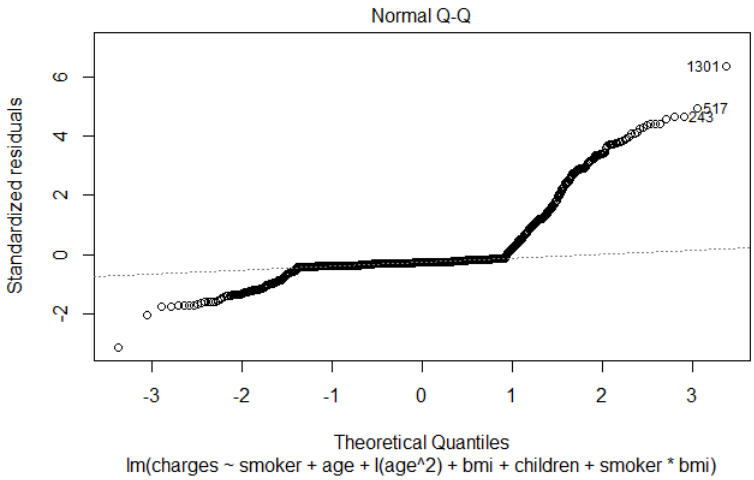

Figure 7: Q-Q plot of the full model. Noticeable departures from the diagonal line suggest non-normality.

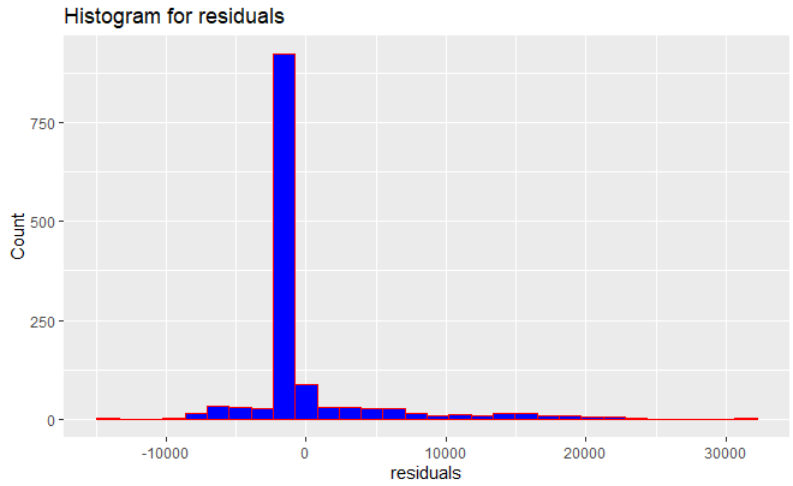

Figure 8: Histogram of the residuals from the full model

The selected model appears to violate the normality assumption. The normal Q-Q plot deviates substantially from the diagonal line. The histogram of residuals is very concentrated around one point and does not resemble a normal bell curve. The Shapiro-Wilk test for normality can be used to confirm our suspicions. The hypotheses for the Shapiro-Wilk test are:

H_0: the sample data are significantly normally distributed
H_a: the sample data are not significantly normally distributed

The output from the *shapiro.test* function provides a p-value of 2.2e-16 which is extremely low relative to $\alpha = 0.05$ (W=0.62486). This provides significant evidence to reject the null hypothesis and conclude the data are not normally distributed.

In an attempt to normalize the data, both the Box-Cox Transformation and the Log Transformation were applied. For the Box-Cox transformation, the best lambda was found to be 0.2879. After transforming the model, neither method was able to help the model pass the Shapiro-Wilk test, and both transformations made the model fail the Breusch-Pagan test.

iv Equal Variance Assumption

It is important for the error terms of a linear regression model have constant variance. This is also known as homoscedasticity. Non-constant variance (heteroscedasticity) can be identified either visually on a plot of residual vs fitted values or mathematically using the Breusch-Pagan test. A widening-to-the-right or widening-to-the-left pattern of the points on a residual vs fitted plot is a quick method to visually identify heteroscedasticity in a dataset. The Residual vs Fitted plot in **Figure 6** does not show either of those patterns. The *bptest* function was used to see if the chosen model passes the Breusch-Pagan test. The hypotheses of the Breusch-Pagan test are as follows:

$$H_0\text{: } \sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2$$
$$H_a\text{: at least one } \sigma_i^2 \text{ is different from the others; } i = 1, 2, ..., n$$

The output from the Breusch-Pagan test gives a p-value of 0.2797 (BP=7.4688), which is much greater than $\alpha = 0.05$, indicating that we fail to reject the null hypothesis. The quadratic model does not violate the equal variance assumption.

v Multicollinearity

The multicollinearity assumption was investigated using the *imcdiag* function. For the full model, multicollinearity was found for multiple vairables (*Smoker, Age, Age*^{2}, and *Smoker*BMI*) (**Table 3**). This was expected due to the interaction and quadratic terms. The test was repeated using the first-order model and no collinearity was detected (**Table 4**).

Variable	VIF	Collinearity
Smoker	25.1440	Yes
Age	47.6081	Yes
Age^2	47.5636	Yes
BMI	1.2976	No
Children	1.0998	No
Smoker*BMI	25.4502	Yes

Table 3: Variance inflation factors of the best model to check for multicollinearity.

Variable	VIF	Collinearity
Smoker	1.0007	No
Age	1.0145	No
BMI	1.0122	No
Children	1.0019	No

Table 4: Variance inflation factors of the main effects model to check for multicollinearity.

vi Outliers

Using visualization, it can be seen that there are several points beyond Cook's distance on the residuals vs leverage plot (**Figure 9**) and the Cook's Distance plot (**Figure 10**).

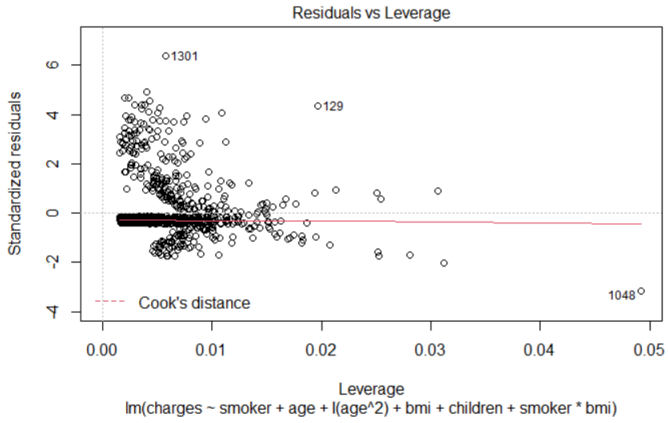

Figure 9: Residuals vs Leverage plot for detecting outliers.

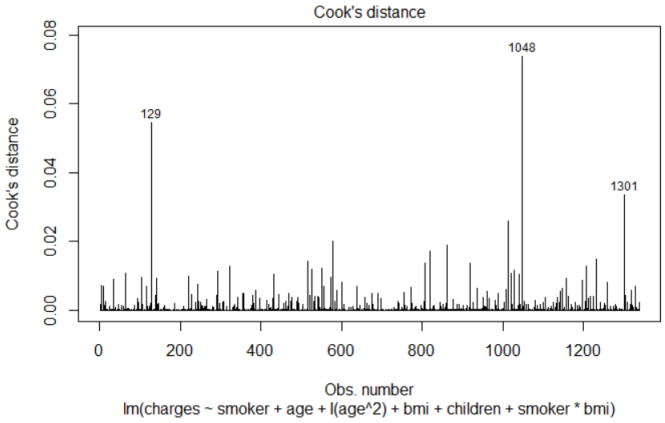

Figure 10: Plot showing the Cook's distance of all points.

By clearly defining the observation number and the magnitude of its influence, this plot aids us in determining the overall impact the outlier points have on our regression. The less conservative threshold for leverage value h_i was used as defined below:

$$h_i > \frac{3p}{n}$$

where h_i is the leverage for the ith observation, p is the number of predictors and n is the number of the sample size. This results in 29 outliers, displayed above the red line in **Figure 11**.

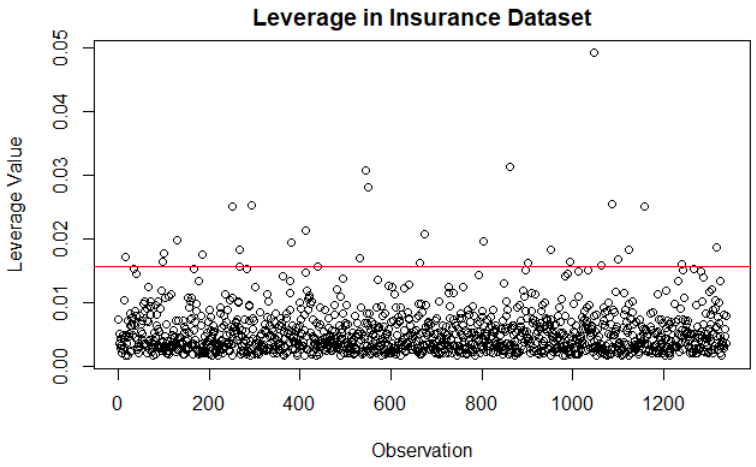

Figure 11: Leverage points above h=0.01569 in the insurance dataset.

To test whether removing the outliers improved the model, the Shapiro-Wilk and the Breusch-Pagan (BP) test were ran. The BP test yielded a p-value of 0.6643 which passed the test. The Shapiro Wilk test resulted in a highly significant p-value, indicating the model still does not meet the normality assumption. The diagnostic plots after removing the outliers are shown in **Figure 12**. It can be seen that they remain similar to the plots prior to removal of outliers. However, removing the outliers changed the quality the model slightly to $R^2_{adj} = 0.8331$ and $RMSE = 4738$. With the outliers in the model these values were $R^2_{adj} = 0.8407$ and $RMSE = 4834$.

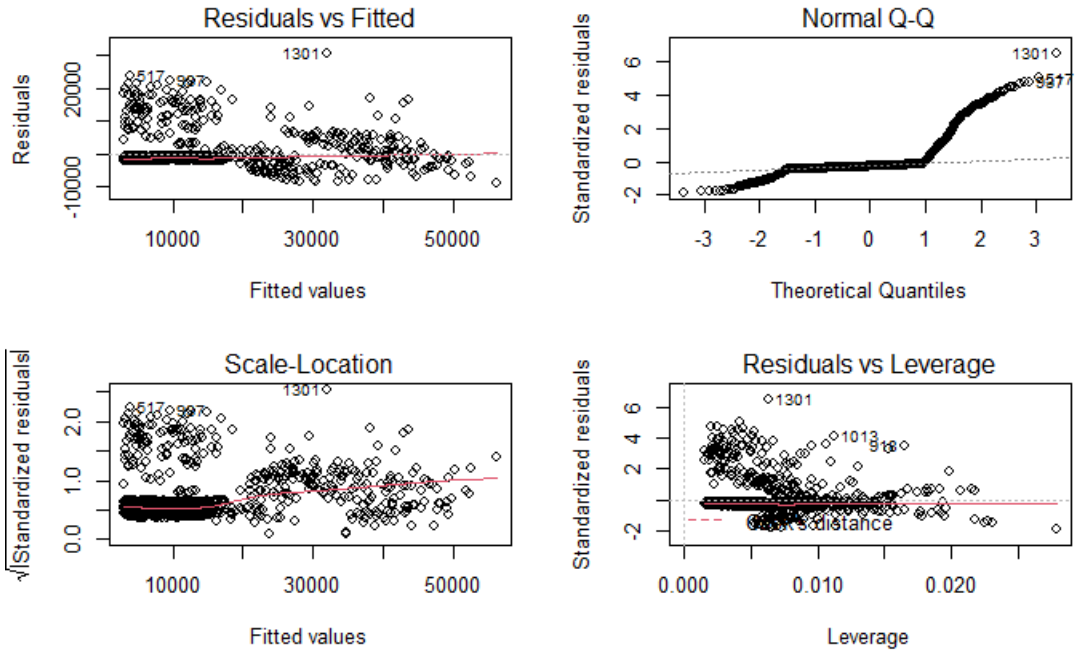

Figure 12: Diagnostic plots after removing outliers.

C Best Fitted Model and Interpreting Coefficients

i Final Model Statistics

Model	R^2_{adj}	RMSE
First Order	0.7489	6068
Interaction	0.8382	4871
Quadratic	0.7516	6036
Removed Outliers	0.8407	4834

Table 5: Comparison of Model Statistics

The final model with the interaction and higher-order term added and outliers removed yields the best result in terms of R^2_{adj} and RMSE, with 0.8407 and 4834, respectively.

ii Interpreting Coefficients

After performing the step wise regression and investigating possible interaction terms and higher order models the best fitting model is found to be:

$$\widehat{Charge} = \beta_0 - \beta_1 Age + \beta_2 Age^2 + \beta_3 BMI + \beta_4 Children -$$
$$\beta_5 Smoker + \beta_6 Smoker * BMI$$

$$\widehat{Charge} = \begin{cases} 22463.46 - 32.21 Age + 3.75 Age^2 + 1434.19 BMI + 667.56 Children, & \text{if Smoker=yes} \\ 2343.70 - 32.21 Age + 3.75 Age^2 + 2.49 BMI + 667.56 Children, & \text{if Smoker=no} \end{cases}$$

The final model with higher order and interaction terms has seven coefficients that affect the outcome of the dependent variable. They are interpreted below.

β_0: 2343.70, the price of policy premiums in dollars if all of the independent variables are held equal to zero. As Age and BMI cannot equal zero, this value cannot be meaningfully interpreted.

$\beta_0 + \beta_5$: 22463.46, the price of policy premiums in dollars if the person is a smoker and all other independent variables are held equal to zero. As Age and BMI cannot equal zero, this value cannot be meaningfully interpreted.

β_1: -32.21, has no meaningful interpretation when in the presence of a quadratic term.

β_2: 3.75, a positive number, indicates that an increase in age will cause an increase in the insurance charge.

β_3: 2.49, if the person is not a Smoker, for each one unit increase in a persons BMI, their premiums will increase 2.49 dollars.

$\beta_3 + \beta_6$: 1434.19, if the person is a Smoker, for each one unit increase in a persons BMI, their premiums will increase 1434.19 dollars.

β_4: 667.5636, for each additional child a person has their premium will increase by 667.56 dollars.

IV Conclusion and Discussion

To summarise the analysis, the Stepwise Regression Procedure chose the following independent variables for the model: Smoker, Age, BMI, and Children. Individual t-tests confirmed that these variables were significant at $\alpha = 0.05$ level. Based on the individual coefficient test, only the interaction between the variables Smoker and BMI were significant. The addition of a quadratic term for Age yielded improved results in terms of adjusted R-squared and RMSE values. The final step after combining main effects, interactions and higher order terms together was to conduct a partial F-test. The partial F-tests confirmed that the higher-order term and the interaction term should be kept in the model. The best fitted model from our results was:

$$\widehat{Charge} = \hat{\beta}_0 + \hat{\beta}_1 Age + \hat{\beta}_2 Age^2 + \hat{\beta}_3 BMI + \hat{\beta}_4 Children + \hat{\beta}_5 Smoker + \hat{\beta}_6 Smoker * BMI$$

$$\widehat{Charge} = \begin{cases} 22463.46 - 32.21 Age + 3.75 Age^2 + 1434.19 BMI + 667.56 Children, & \text{if Smoker=yes} \\ 2343.70 - 32.21 Age + 3.75 Age^2 + 2.49 BMI + 667.56 Children, & \text{if Smoker=no} \end{cases}$$

The results of the analysis provided some confirmations about how insurance premiums are priced. Unsurprisingly, the variables that best reflected an individual's overall health were the most significant in terms of predicting Charges. An older individual generally has more health issues than a younger one, and their premium will be higher. An individual that smokes, or has a high BMI, is on average less healthy. Overall, the higher the potential of an individual to get sick, the higher their premium charge. The interaction between an individual smoking and their BMI also has an effect on the premium charges.

The biggest challenge encountered was meeting the normality assumption. While most of the assumptions were met with the full model, it consistently failed the normality assumption. Even after performing a Box-Cox transformation, the residuals were not normally distributed. A possible reason for the lack of normal distribution in the residuals is the fact that the data was not a random sample, rather a simulated dataset created for a textbook. This may indicate the importance of true random data over false creations. However, it is likely that the biggest reason the model does not meet the normality assumption is that it did not contain the all of the influential independent variables. There are a lot of different factors that affect premium charges for insurance, some obvious ones that were not included in the dataset are: level of coverage, deductibles, medical history, and family history, to name a few. These are very important factors that were not part of the analysis and this may have been the underlying reason for the problems with our model.

While the R^2_{adj} and RMSE values for the final model were satisfactory, it ultimately cannot be said that the model is sufficient enough to make accurate predictions. The model could have been greatly improved by realizing the problem with the data from the beginning and searching for a dataset with more complete variables. This may have improved the predicting ability of the model, as well as fixed the non-normality issues.

References

[1] H. Levy and D. Meltzer. The impact of health insurance on health. *Annual Reviews*, 2008.

[2] Trisha Torrey. Why Health Insurance Premiums Increase. https://www.verywellhealth.com/health-insurance-premium-increase-2615099. Accessed: 2021.03.26.

[3] Julia Kagan. Underwriting risk. https://www.investopedia.com/terms/u/underwriting-risk.asp. Accessed: 2021.03.26.