Predicting the Health Insurance Premium in the United States

UNIVERSITY OF CALGARY

Fall 2023: DATA 603- L01

Group #19:

Prashant Dhungana-30080130
Prateek Kaushik-30229287
Mariya Mathews-30192182
Nisha Pillai-30158934

TABLE OF CONTENTS

INTRODUCTION

In Canada the health services are provided at no cost because it is covered and funded by the government. The system is called “Universal Health Care System” and was passed by legislation in 1984 under the Canada Health Act. However, the same does not apply for our neighboring country, the United States. The health services in the United States are mostly private and unlike Canada the people in the States have to pay for their health service fees. Due to the privatization of the healthcare system in the United States there is a competition within the industry to provide better services which makes it expensive for the people to afford. Therefore, health insurance companies play a huge role as they cover the cost of health services in exchange for an insurance premium. By purchasing health insurance people protect themselves from any future uncertainties and a huge debt. In order for the insurance companies to cover the cost of health services they must analyze how much premium to charge for each person. Every person has different day to day habits and can vary in their lifestyle choices which impacts how the insurance companies charge their premiums.

OBJECTIVE

The objective for this project is to develop the best model in predicting insurance premiums based on several factors using multiple linear regression. In order to predict the insurance premiums, we first must determine which predictor variables are significant and develop a model with predictors that explains a high variability in terms of the premiums charged. This topic is very important as people residing in Canada can use this model to predict the insurance premiums they will be charged according to their personal information if they ever plan on moving to the United States.

MOTIVATION

The motivation for our study is to find the factor that most influences insurance prices and help people plan accordingly to reduce their financial burden of paying high premiums. We will be using multiple linear regression to address this query.

METHODOLOGY

DATASET

In order to answer our research topic, we have acquired a dataset that has 7 variables and 1338 observations. The dataset contains both numerical and categorical values. The dataset is clean and no further cleaning was required. The variables are explained below according to the source of the dataset:

Response Variable:

Charges: Insured individual medical annual costs billed by the health insurance in USD ($). (Quantitative)

Predictor Variables:

1. Age: Age of the primary Insured. (Quantitative)
2. Sex: Insured gender “Male or Female”. (Qualitative)
3. BMI: Body Mass Index. (Quantitative)
4. Children: Insured number of dependents. (Quantitative)
5. Smoker: Insured smoking habits “Yes or No”. (Qualitative)
6. Region: Insured region of residence “Southwest”, “Southeast”, “Northwest”, “Northeast”. (Qualitative)

The dataset was acquired through Kaggle website and is an open-sourced data licensed under “Open Data Commons”.

MODEL PLANNING

In order to create a model that can successfully predict the insurance charges we will implement the multiple linear regression method. The first step is to create a linear model with all the predictors variables. The second step is to conduct stepwise, forward and backward regression methods to find the predictor variables that are significant to be included in the model. The third step is to conduct an all-possible regression procedure to compare if we get the same predictors results as the stepwise model. Following these steps, we will determine our best additive model and the next steps will be to see if there is any interaction and higher orders terms that are significant statistically. After finding the best model we will be checking for the regression assumptions in order:

1. Linearity Assumption
2. Independence Assumption
3. Equal Variance Assumption
4. Normality Assumption
5. Multicollinearity Assumption
6. Outliers

If all the assumptions hold, we then proceed to do our prediction on the insurance premium(charges). However, in the case if some assumptions are not met then we will proceed to do Log and Box-Cox transformations to improve our model. Once we check all the assumptions, we proceed to do predictions based on our best fit model.

WORKLOAD DISTRIBUTION

STEPS	METHODOLOGY	TEAM MEMBER
Step 1	Model building and Model selection including Stepwise, Forward, Backwards, All possible regressions selection, T-test & F-test	Prashant Dhungana
Step 2	Model building continued with Interaction model and Quadratic (higher) order model	Prateek Kaushik
Step 3	Checking the Regression Assumptions including Linearity, Independence, Equal variance, Normality, Multicollinearity, and identifying if any Outliers	Nisha Pillai
Step 4	Box-cox transformations in case of non-normality, Model evaluation by reiterating and refining process and Predicting the Insurance premium	Mariya Mathews
Step 5	Preparing Final report	All team members with equitable contribution

RESULTS

VARIABLE SELECTION PROCEDURE

For the first model we created the linear model with all the predictors variables and decided to conduct the stepwise, forward and backward regression procedure to find the predictors that are significant statistically.

The full model is as follows:

$$\widehat{Charges} = \beta_0 + \beta_1 Age + \beta_2 \text{Sexmale} + \beta_3 \text{BMI} + \beta_4 \text{Children} + \beta_5 \text{Smokeryes} + \beta 6 \text{Regionnorthwest} + \beta 7 \text{Regionsoutheast} + \beta 8 \text{Regionsouthwest}$$

Here is the summary for the full model:

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

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            -11938.5      987.8 -12.086  < 2e-16 ***
age                       256.9       11.9  21.587  < 2e-16 ***
factor(sex)male          -131.3      332.9  -0.394 0.693348
bmi                       339.2       28.6  11.860  < 2e-16 ***
children                  475.5      137.8   3.451 0.000577 ***
factor(smoker)yes       23848.5      413.1  57.723  < 2e-16 ***
factor(region)northwest   -353.0      476.3  -0.741 0.458769
factor(region)southeast  -1035.0      478.7  -2.162 0.030782 *
factor(region)southwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Figure 1. Output summary of full model

Individual Coefficient T-Test:

The Hypothesis:

Null Hypothesis H(0): βi=0
Alt. Hypothesis H(A): βi≠0 (i=age, bmi,.,region)

Based on the summary from the full model we observe the p-value of predictor variable sex is 0.6933 which is greater than the significance level at alpha=0.05 therefore we fail to reject the null hypothesis therefore should remove this variable from our model. However, all the other predictors have their p-value less than 0.05 therefore we reject the null hypothesis indicating these predictors (age, bmi, children, smoker and region) are significant statistically. Therefore after removing the predictor variable sex from our model the reduced model is as follows:

$$\widehat{Charges} = \beta_0 + \beta_1 Age + \beta_2 \text{BMI} + \beta_3 \text{Children} + \beta_4 \text{Smokeryes} + \beta 5 \text{Regionnorthwest} + \beta 6 \text{Regionsoutheast} + \beta 7 \text{Regionsouthwest}$$

Step-Wise, Forward, and Backward Selection Procedure

In order to determine if we get the same results whereas, only the predator variable sex is insignificant from on determining the charges we conduct all three regression procedures. Based on these procedures we were able to get the same results where the significant variables were (age, bmi, children, smoker and region). The outputs were Adjusted R2=0.7496 and RMSE=6060.178.

All Possible Selection Procedure

The next step is to conduct an all-possible-regression-selection procedure. We computed the AIC, Adjusted_R2 and Cp to determine if we get the same results as the procedures above. The table below provides the comparison:

	rsquare	AdjustedR	cp	AIC
[1,]	0.6197648	0.6194802	694.739482	27667.46
[2,]	0.7214008	0.7209834	154.461974	27253.32
[3,]	0.7474772	0.7469093	17.332067	27123.84
[4,]	0.7496945	0.7489434	7.501312	27114.04
[5,]	0.7508839	0.7495727	3.155553	27113.66
[6,]	0.7509130	0.7494136	5.000000	27115.51

Figure 2. Output summary of all-possible-regression-selection procedure

From the table above we observe that the model with the five predictors has the lowest AIC=27113.66, with the highest Adjusted R2=0.74957 and the lowest CP=3.1555. Based on this we can infer the model with the five predictor variables is the best model. Therefore, our best model after all the regression selection procedure is the same as above after removing the predictor variable sex which is as follows:

$$\widehat{Charges} = \beta_0 + \beta_1 Age + \beta_2 \text{BMI} + \beta_3 \text{Children} + \beta_4 \text{Smokeryes} + \beta 5 \text{Regionnorthwest} + \beta 6 \text{Regionsoutheast} + \beta 7 \text{Regionsouthwest}$$

Interaction Model

After finding the best additive model we now conduct an interaction regression procedure to find if any interaction terms are significant statistically. Based on the Individual T-test at significance level alpha=0.05.

The Hypothesis
Null Hypothesis H(0) βi=0
Alt. Hypothesis H(A): βi≠0 (i= all interaction terms)

The summary of the interaction terms is as follows:

```
Call:
lm(formula = charges ~ (age + bmi + children + factor(smoker) +
    factor(region))^2, data = insurance_data)

Residuals:
     Min       1Q   Median       3Q      Max
-11933.3  -2033.5  -1216.5   -205.3  30110.7

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -2413.784   2458.849  -0.982 0.326442
age                                       199.844     52.191   3.829 0.000135 ***
bmi                                        54.835     80.758   0.679 0.497254
children                                  712.551    654.593   1.089 0.276557
factor(smoker)yes                       -20760.944   1919.482 -10.816  < 2e-16 ***
factor(region)northwest                  -1332.395   2247.427  -0.593 0.553381
factor(region)southeast                   3325.053   2154.165   1.544 0.122939
factor(region)southwest                    -73.839   2162.551  -0.034 0.972767
age:bmi                                      1.191      1.628   0.732 0.464438
age:children                                -1.687      8.553  -0.197 0.843657
age:factor(smoker)yes                       -2.460     23.892  -0.103 0.918007
age:factor(region)northwest                 17.529     27.325   0.642 0.521309
age:factor(region)southeast                 49.037     27.420   1.788 0.073948 .
age:factor(region)southwest                 46.601     27.827   1.675 0.094238 .
bmi:children                                -0.117     19.114  -0.006 0.995117
bmi:factor(smoker)yes                     1476.322     55.822  26.447  < 2e-16 ***
bmi:factor(region)northwest                 -8.402     70.063  -0.120 0.904559
bmi:factor(region)southeast               -190.765     60.527  -3.152 0.001660 **
bmi:factor(region)southwest                -94.332     66.954  -1.409 0.159100
children:factor(smoker)yes                -409.109    284.473  -1.438 0.150636
children:factor(region)northwest           304.086    322.643   0.942 0.346118
children:factor(region)southeast          -174.003    321.830  -0.541 0.588829
children:factor(region)southwest          -352.934    308.972  -1.142 0.253544
factor(smoker)yes:factor(region)northwest  -178.139    967.473  -0.184 0.853941
factor(smoker)yes:factor(region)southeast -1078.716    921.117  -1.171 0.241773
factor(smoker)yes:factor(region)southwest   958.144    980.187   0.978 0.328496
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4831 on 1312 degrees of freedom
Multiple R-squared:  0.8438,    Adjusted R-squared:  0.8409
F-statistic: 283.6 on 25 and 1312 DF,  p-value: < 2.2e-16
```

Figure 3. Output summary of all interaction model

Based on the summary the only interaction terms that are statistically significant with the p-value less than alpha=0.05 are the interaction between (bmi: factor(smoker)) and (bmi: factor(region)). Now we include the interaction terms that are significant to our best additive model. Our best model with interaction terms is reduced to only include the interaction terms that are significant and is as follows:

$$\widehat{Charges} = \beta_0 + \beta_1 Age + \beta_2 \text{BMI} + \beta_3 \text{Children} + \beta_4 \text{Smokeryes} + \beta 5 \text{Regionnorthwest} + \beta 6 \text{Regionsoutheast} + \beta 7 \text{Regionsouthwest} + \beta 8 \text{BMI} * \text{Smokeryes} + \beta 9 \text{BMI} * \text{Regionnorthwest} + \beta 10 \text{BMI} * \text{Regionsoutheast} + \beta 11 \text{BMI} * \text{Region}_{\text{southwest}}$$

Based on this model we get the Adjusted_R2=0.8405 with RMSE=4836.0. After finding our best interaction model we need to compare this model with the best additive model. We compare these two models in the table below:

Model <chr>	Adjusted_R2 <dbl>	RMSE <dbl>
reduced_model	0.7495727	6060.178
interact_final_model	0.8405284	4836.000

Figure 4. Adjusted R Squared and RMSE for reduced_model and interact_final_model

We observe that the Adjusted_R2 is higher for the (interact_final_model) interaction model and which also has the lowest RMSE value in comparison to the (reduced_model) which is the best additive model. Therefore, we conclude that the interaction terms are more significant and the next step is to add higher order terms in this best interaction model.

HIGHER ORDER MODEL

In order to find if any of the quantitative predictors are significant to be added to the model we start with a quadratic term and if any of the quadratic terms are significant which is determined by the individual t-test. If the p-value of the higher order is less than alpha=0.05 we conclude it to be significant to be added to the model whereas, if p-value is greater than alpha=0.05 we do not add it to our model as it is insignificant.

Individual Coefficient t-test of Higher Order Model

The Hypothesis:

Null Hypothesis H(0): βi=0
Alt. Hypothesis H(A): βi≠0 (i= all higher order terms)
After adding multiple higher order terms, we were able to find this model as the best higher order model that is statistically significant. The summary of the best higher order model is provided below:

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

Residuals:
    Min      1Q  Median      3Q     Max
-8917.8 -1968.6 -1257.3  -440.1 30625.5

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  8.395e+04  2.829e+04   2.967  0.00306 **
age                         -2.071e+01  6.407e+01  -0.323  0.74658
I(age^2)                     3.579e+00  7.989e-01   4.480 8.12e-06 ***
bmi                         -1.029e+04  3.687e+03  -2.791  0.00533 **
I(bmi^2)                     4.592e+02  1.747e+02   2.628  0.00869 **
I(bmi^3)                    -8.556e+00  3.581e+00  -2.389  0.01702 *
I(bmi^4)                     5.640e-02  2.682e-02   2.103  0.03569 *
children                     6.677e+02  1.135e+02   5.881 5.15e-09 ***
factor(smoker)yes           -2.073e+04  1.635e+03 -12.677  < 2e-16 ***
factor(region)northwest      9.681e+00  2.039e+03   0.005  0.99621
factor(region)southeast      2.852e+03  2.026e+03   1.407  0.15953
factor(region)southwest      1.097e+03  1.987e+03   0.552  0.58117
bmi:factor(smoker)yes        1.452e+03  5.217e+01  27.840  < 2e-16 ***
bmi:factor(region)northwest -2.136e+01  6.867e+01  -0.311  0.75581
bmi:factor(region)southeast -1.264e+02  6.417e+01  -1.969  0.04917 *
bmi:factor(region)southwest -7.966e+01  6.514e+01  -1.223  0.22160
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4761 on 1322 degrees of freedom
Multiple R-squared:  0.8472,    Adjusted R-squared:  0.8455
F-statistic: 488.6 on 15 and 1322 DF,  p-value: < 2.2e-16
```

Figure 5. Output summary of best higher order model

When adding the I(bmi^5) the model becomes insignificant for the predictor bmi as the p-value increases over alpha=0.05. The same happens when we increase the power to three for the age variable it becomes insignificant. The higher order for the children was also insignificant at power of two as the p-value was higher than alpha=0.05.
The table below compares all the model that were significant to find the best higher order model:

Model <chr>	Adjusted_R2 <dbl>	RMSE <dbl>
quad_model	0.8449461	4768.545
cubic_model	0.8449461	4768.545
fourth_model	0.8454561	4760.697

Figure 6. Adjusted R Squared and RMSE for quad_model, cubic_model, fourth_model

From the table above we observe that the (fourth_model) is the best because it has the highest Adjusted R2 with the lowest RMSE value.

A brief overview of the other higher order models is that our (quad_model) had the higher order terms (age^2) in addition with (bmi^2) and (children^2) and the only significant variables were age and bmi. The (cubic_model) had the higher order terms (age^2), (age^3) and (bmi^2), (bmi^3) and the only significant variables were the bmi variables. For the (fourth_model) the higher order terms were (age^2) and (bmi^2), (bmi^3), (bmi^4) and all were significant at alpha=0.05. When we increase the power of bmi to five it becomes insignificant as its p-value is greater than alpha=0.05. Therefore, we conclude that (fourth_model) is the best higher order model. The higher order model is as follows:

$$\widehat{Charges} = \beta_0 + \beta_1 Age + \beta_2(\text{Age}^2) + \beta_3\text{BMI} + \beta_4(\text{BMI}^2) + \beta_5(\text{BMI}^3) + \beta_6(\text{BMI}^4) + \beta_7\text{Children} + \beta_8\text{Smokeryes} + \beta 9\text{Regionnorthwest} + \beta 10\text{Regionsoutheast} + \beta 11\text{Regionsouthwest} + \beta 12\text{BMI} * \text{Smokeryes} + \beta 13\text{BMI} * \text{Regionnorthwest} \beta 14\text{BMI} * \text{Regionsoutheast} + \beta 15\text{BMI} * \text{Region}_{\text{southwest}}$$

We now compare our best interaction model with the best higher order model to determine that all the added higher order terms are statistically significant by conducting a F-test.

F-Test (Interaction vs Higher Order) Model

The Hypothesis:

Null Hypothesis H(0): $\beta_{p-q+1}=\beta_{p-q+2}=...=\beta_p= 0$. (Higher order terms are not significant)
Alt. Hypothesis H(A): At Least one $\beta_p \neq 0$. (At least one higher order term is significant)

Anova Table:

```
Analysis of Variance Table

Model 1: charges ~ age + I(age^2) + bmi + I(bmi^2) + I(bmi^3) + I(bmi^4) +
    children + factor(smoker) + factor(region) + bmi:factor(smoker) +
    bmi:factor(region)
Model 2: charges ~ age + bmi + children + factor(smoker) + factor(region) +
    bmi:factor(smoker) + bmi:factor(region)
  Res.Df        RSS Df   Sum of Sq     F    Pr(>F)
1   1322 2.9962e+10
2   1326 3.1011e+10 -4 -1048901645 11.57 3.096e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7. Output summary of ANOVA table for Interaction vs Higher Order

From the Anova table above we observe that the p-value is less than the alpha=0.05 therefore we reject the null hypothesis which indicates that the higher order terms added to the model are statistically significant. Therefore, our best model so far, for predicting insurance charges, is the model with higher terms added.

Our best fitted model:

$$\widehat{Charges}=\beta_0+\beta_1 Age+\beta_2(\text{Age}^2)+\beta_3\text{BMI}+\beta_4(\text{BMI}^2)+$$
$$\beta_5(\text{BMI}^3)+\beta_6(\text{BMI}^4)+\beta_7\text{Children}+\beta_8\text{Smokeryes}+\beta 9\text{Regionnorthwest}+$$
$$\beta 10\text{Regionsoutheast}+\beta 11\text{Regionsouthwest}+\beta 12\text{BMI}*\text{Smokeryes}+$$
$$\beta 13\text{BMI}*\text{Regionnorthwest}\beta 14\text{BMI}*\text{Regionsoutheast}+\beta 15\text{BMI}*\text{Region}_{\text{southwest}}$$

MULTIPLE REGRESSION ASSUMPTIONS

The sections below will address how we tested our model to meet various assumptions associated with running multiple regression. These assumptions must be tested, to ensure that our model results are, to an extent, trustworthy

The model we will be working with is as follows:

$$\widehat{Charges} = 83950 - 20.71Age + 3.579(\text{Age}^2) - 10290\text{BMI} + 459.2(\text{BMI}^2) - 8.556(\text{BMI}^3) + 0.0564(\text{BMI}^4) + 667.7\text{Children} - 20730\text{Smokeryes} + 9.681\text{Regionnorthwest} + 2852\text{Regionsoutheast} + 1097\text{Regionsouthwest} + 1452\text{BMI} * \text{Smokeryes} - 21.36\text{BMI} * \text{Regionnorthwest} - 126.4\text{BMI} * \text{Regionsoutheast} - 79.66\text{BMI} * \text{Regionsouthwest}$$

Interpretation of model coefficients:

This regression model aims to predict charges based on various predictors. Let's interpret the coefficients:

- **Intercept (Constant):** The intercept is significant (p-value = 0.00306) and suggests the expected charges when all other predictors are zero.
- **Age and its quadratic term (age, I(age^2)):** Age and its squared term have mixed significance. The squared term (p-value = 8.12e-06) indicates a significant non-linear relationship between age and charges.
- **BMI and its polynomial terms (bmi, I(bmi^2), I(bmi^3), I(bmi^4)):** BMI and its polynomial terms show significance, indicating a potentially non-linear relationship between BMI and charges. The coefficients suggest a complex relationship between BMI and charges, where higher-order terms contribute to the prediction.
- **Children:** Each additional child is associated with an increase of approximately 667.7 units in charges (p-value < 2e-16).
- **Smoker (factor(smoker)):** Being a smoker significantly decreases charges by approximately 20730 units (p-value < 2e-16).
- **Region (factor(region)):** The impact of regions on charges seems less significant, with p-values above typical significance levels (p > 0.05), except for the southeast region, which has a borderline significance (p-value = 0.04917).
- **Interaction Terms (bmi: factor(smoker), bmi: factor(region)):** These interaction terms show significance, suggesting that the relationship between BMI and charges might vary based on smoking status and region.
- **R-squared and F-statistic:** The model has a high R-squared value of 0.8472, indicating that the predictors explain approximately 84.72% of the variability in charges. The F-statistic (488.6) and its associated p-value (< 2.2e-16) suggest that the overall model is significant in predicting charges.

1. LINEARITY ASSUMPTION

The linearity assumption in statistics, particularly in regression analysis, assumes that the relationship between the independent variables (predictors) and the dependent variable (outcome) is linear. This means that changes in the predictors are associated with a constant change in the outcome variable, holding all other variables constant. When discussing linear regression specifically, it assumes that the relationship between the predictors and the response variable can be described by a straight line in a multi-dimensional space (for multiple predictors).

a. Residuals vs. Fitted Values plot for best-fitted model

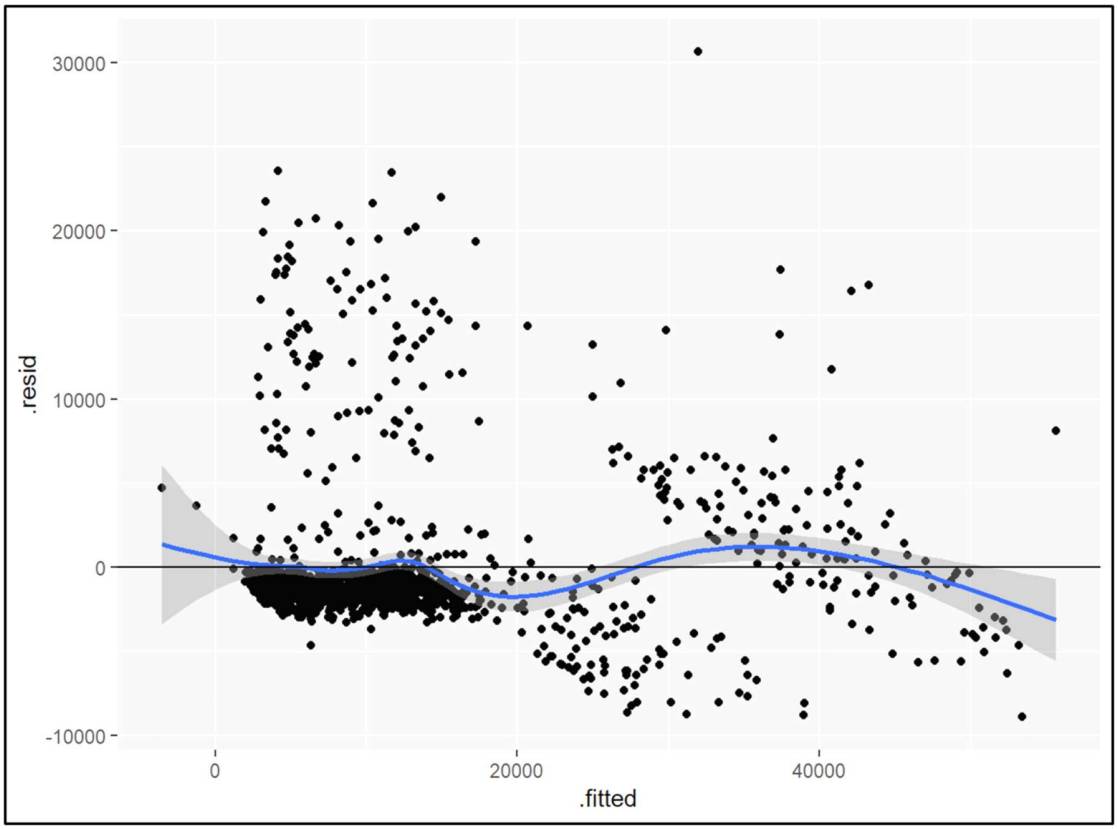

Figure 8. Plot to check linearity

Interpretation:

The Residual vs. Fitted plot shown in Figure 8., for the fourth higher-order model, identified as the best fitted model, exhibits no discernible pattern among the residuals that are non-linear. Upon reviewing this plot, there are no apparent concerns regarding the residual patterns. Therefore, it is reasonable to assert that the linearity assumption has been satisfied.

2. INDEPENDENCE ASSUMPTION

The independence assumption in statistics, particularly in regression analysis, refers to the assumption that the errors (or residuals) of the model are independent of each other. It specifically means that knowing the value of one error does not provide any information about the value of another error.

a. Residuals vs. Region plot for best-fitted model

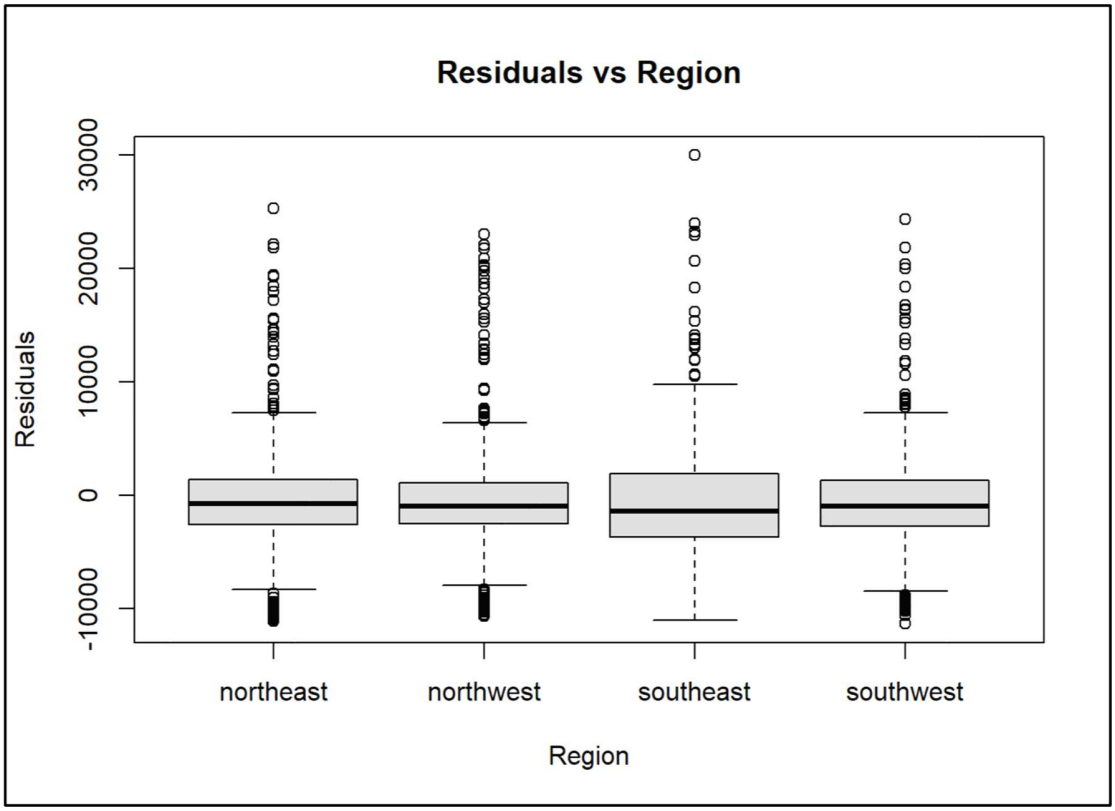

Figure 9. Plot to check for independence of error terms

Interpretation:

In this dataset, the observations were not related to time but were associated with spatial data represented by the region variable, indicating a potential grouping effect. While spatial association exists among the measurements, the plot of residuals against regions (Figure 9.) reveals no notable clustering of residuals. This lack of substantial grouping implies that the independence assumption holds, suggesting that the data satisfies the requirement of independence.

3. EQUAL VARIANCE ASSUMPTION

The equal variance assumption, also known as homoscedasticity, is a fundamental assumption in many statistical models, particularly in linear regression. It refers to the assumption that the variance of the residuals (or errors) across all levels of the predictors remains constant or uniform.

We visually evaluate homoscedasticity using diagnostic tools such as residual plots and scale-location plots, focusing on the relationship between fitted values and standardized residuals. Additionally, we use the Breusch-Pagan test, a formal statistical method, to detect potential heteroscedasticity in our regression analysis.

a. Residual plot and Scale-location plot for best-fitted model

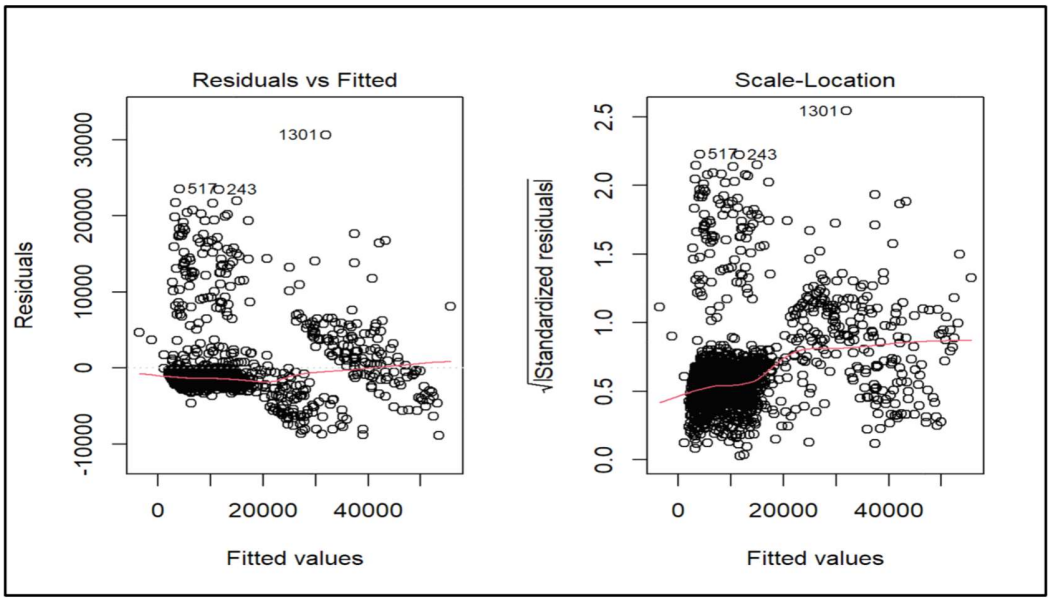

Figure 10. Residual plot and Scale-location plot to check for homoscedasticity

Interpretation:

The residual plot and scale-location plot from Figure 10. show no evidence of funneling, suggesting the possible presence of homoscedasticity. These visual assessments indicate no apparent issue with the homoscedasticity assumption. Nonetheless, to further confirm homoscedasticity, we'll utilize the Breusch-Pagan test as part of our assessment.

b. Breusch-Pagan test

The Hypothesis:

Null Hypothesis H(0): Heteroscedasticity is not present (homoscedasticity)
Alt. Hypothesis H(A): Heteroscedasticity is present

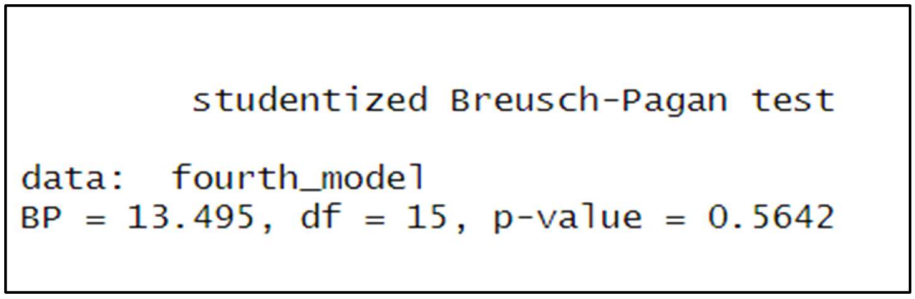

Figure 11. Output of Breusch-Pagan test

Interpretation:

The null hypothesis is that we have homoscedasticity. From Figure 11. we can see that the p-value is 0.5642 which is greater than the level of significance of 0.05 hence we clearly fail to reject the null hypothesis and conclude that heteroscedasticity is not present. This finding indicates the presence of homoscedasticity, affirming the validity of the equal variance assumption without any apparent issues

4. NORMALITY ASSUMPTION

The Normality Assumption in statistics pertains to the assumption that the residuals (or errors) of a statistical model follow a normal distribution. This assumption is crucial in various statistical techniques, particularly in linear regression and other parametric methods.

We visually evaluate normality using diagnostic tools such as histograms and Q-Q plots, focusing on the relationship between theoretical quantiles and standardized residuals. Additionally, we use the Shapiro-Wilk normality test, a formal statistical method, to confirm normality in our regression analysis.

a. Histogram and Q-Q plot of residuals values for best-fitted model

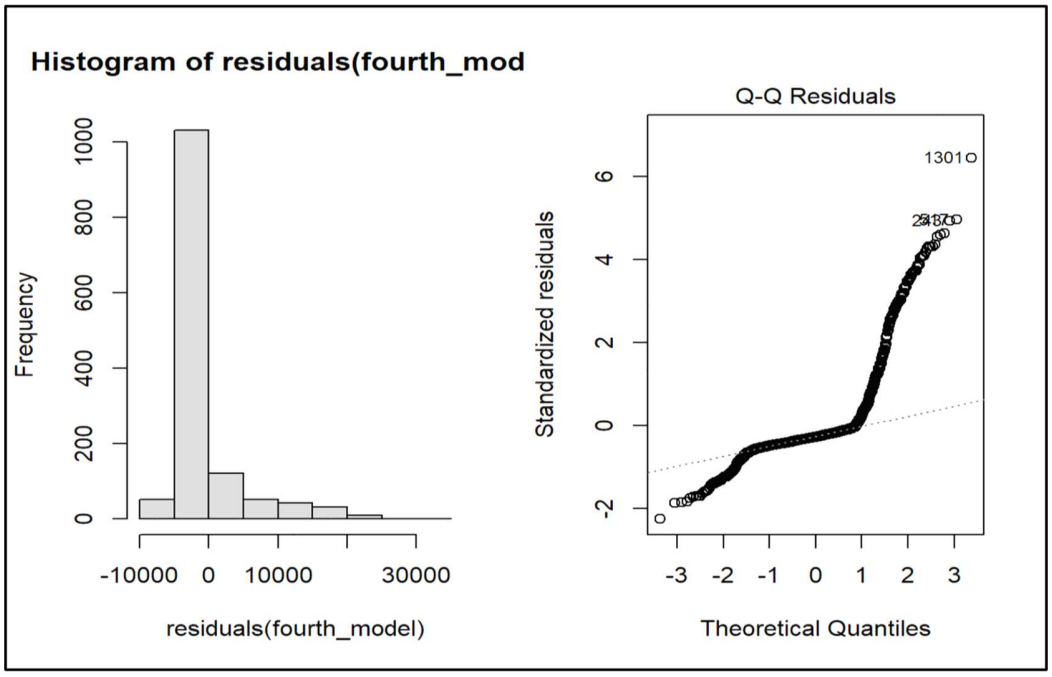

Figure 12. Histogram and Q-Q plot to check for normal distribution

Interpretation:

The Histogram plot and Q-Q plot from Figure 12. shows that the distribution is not symmetrical. From the Q-Q plot it can be seen that the plot follows a S-shaped pattern of deviations indicating that the residuals have excessive kurtosis. These visual assessments indicate an issue with the normality assumption. To confirm if the data significantly deviates from a normal distribution we conduct the Shapiro-Wilk normality test.

b. Shapiro-Wilk normality test

The Hypothesis:

Null Hypothesis H(0): The sample data are significantly normally distributed
Alt. Hypothesis H(A): The sample data are not significantly normally distributed

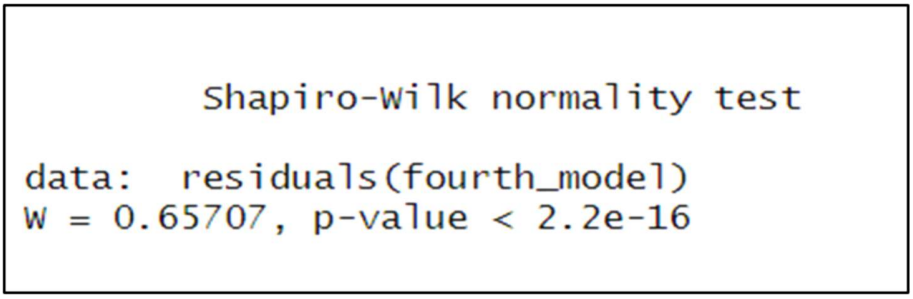

Figure 13. Output of Shapiro-Wilk test

Interpretation:

The Shapiro-Wilk normality test (Figure 13.) shows p-value < 2.2e-16 which is significantly less than α=0.05. Hence, we clearly reject the null hypothesis of normality. This finding indicates the absence of normality; hence we can confirm that the normality assumption is not met by our dataset.

5. MULTICOLLINEARITY TESTS

Multicollinearity refers to a scenario in regression analysis where two or more predictor variables in a model are highly correlated with each other. This high correlation creates redundancy in the information provided by these variables, leading to issues in interpreting the model and estimating the coefficients accurately.

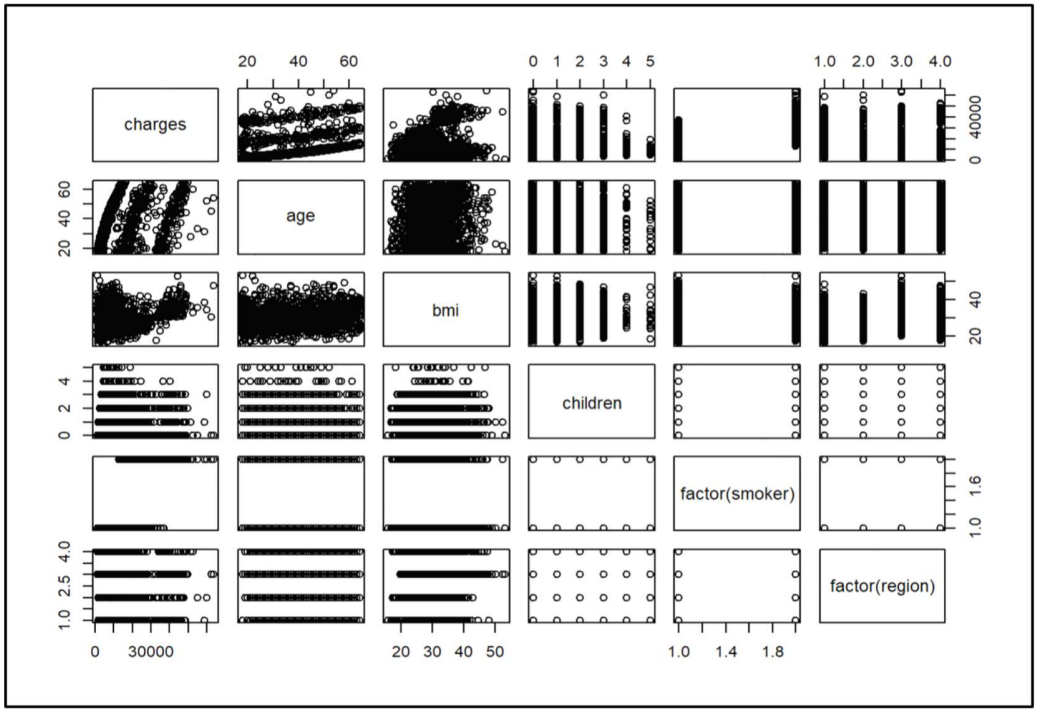

Figure 14. Plots to check for multicollinearity

Interpretation:

From Figure 14. we do not detect any correlation between the independent predictor variables.

```
Call:
imcdiag(mod = insurance_vif_model, method = "VIF")


 VIF Multicollinearity Diagnostics

                             VIF detection
age                       1.0162         0
bmi                       1.1042         0
children                  1.0037         0
factor(smoker)yes         1.0064         0
factor(region)northwest 1.5188         0
factor(region)southeast 1.6522         0
factor(region)southwest 1.5294         0

NOTE:  VIF Method Failed to detect multicollinearity


0 --> COLLINEARITY is not detected by the test

===================================
```

Figure 15. Output from VIF test

Interpretation:

The diagnostic assessments from Figure 15. reveal no signs of multicollinearity among any variables incorporated in the model. With all Variance Inflation Factor (VIF) values well below the threshold of 10, there's no indication that collinearity is impacting the model's parameter estimates. According to the VIF test, we can infer that no evidence of multicollinearity exists among the independent predictors within the linear regression model. Hence, we can conclude that each coefficient distinctly contributes to the variance in insurance expenses without excessive influence from other variables in the model.

6. OUTLIERS

An outlier is an observation or data point that significantly differs from other observations in a dataset. These data points are notably distant from the rest of the data and can skew statistical analyses, affecting the overall interpretation and reliability of the results.

a. Residual vs leverage plot for the best fitted model

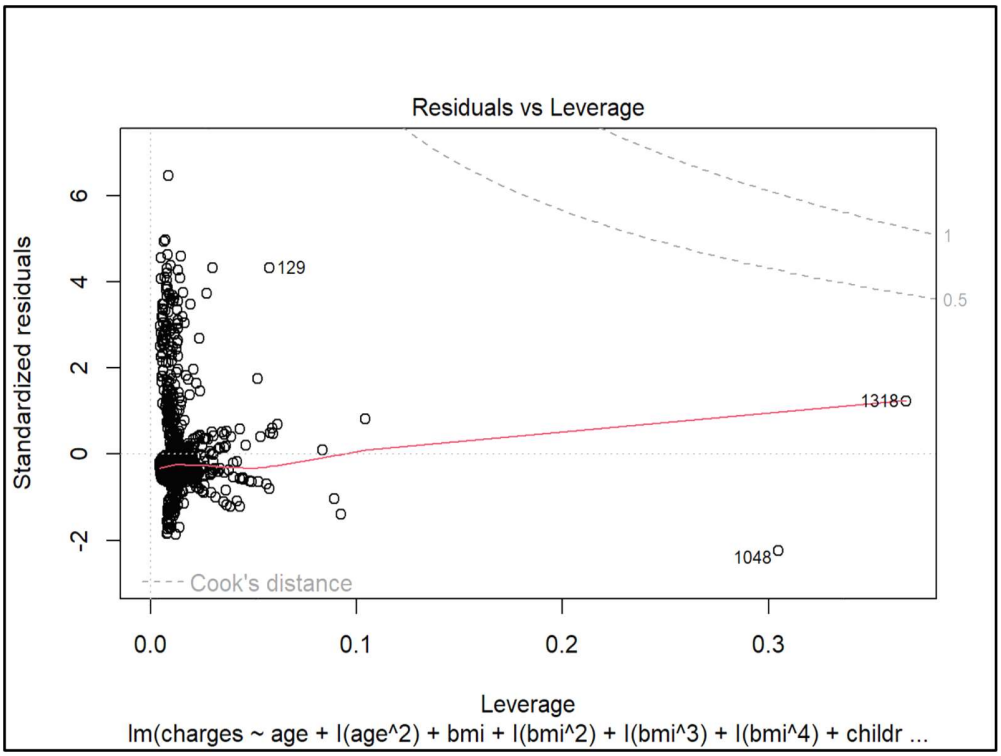

Figure 16. Plot to check for influential points

Interpretation:

Influential observations possess significant impact on our model's outcomes. To examine this, we generate a Residual vs Leverage plot (depicted in Figure 16.) Upon reviewing the plot, we observe no data points exceeding Cook's distance. This indicates the absence of influential points that could disproportionately affect our regression results.

b. Cook's Distance

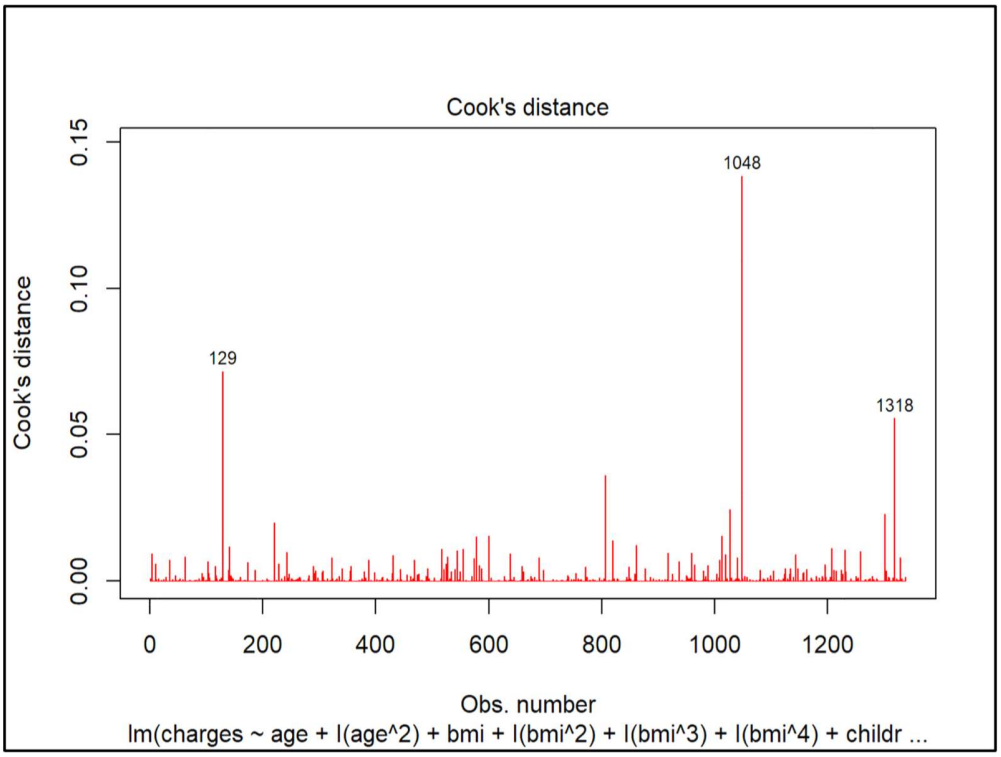

Figure 17. Plots to check for outliers -Cook's Distance

Interpretation:

Figure 17. displays the Cook's distance plotted against each observation, aiding in assessing the collective influence of outliers on our regression analysis. This plot effectively identifies each observation number along with the magnitude of its impact. Notably, observations 129, 1318, and 1048 exhibit the greatest Cook's distance values. Despite this, their Cook's Distance values, all below 0.5, indicate that they do not exert influential effects on the analysis.

c. Leverage points

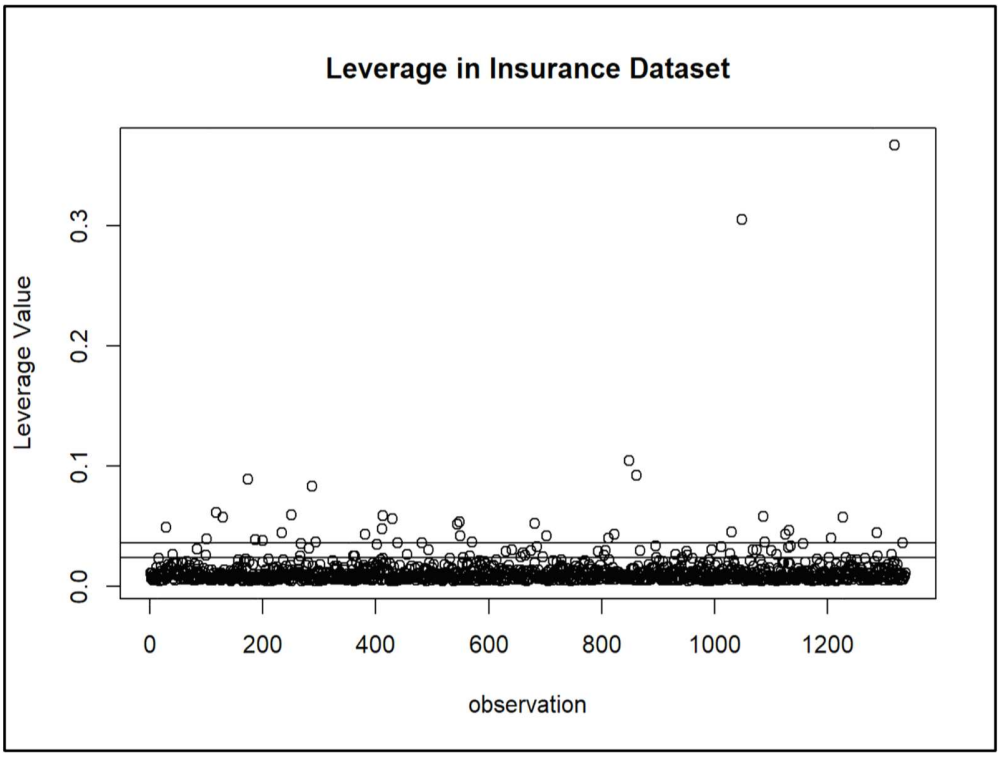

Figure 18. Plots to check for outliers -Leverage values vs observation Plot

Interpretation:

The plot in Figure 18. illustrating leverage values against observations, indicates the presence of leverage points. However, upon examination, none of these points are influential. Hence we do not detect any concerning influential outliers.

BOX-COX TRANSFORMATION

To remedy the nonnormality of the selected multiple linear regression model, transformation on Y is required, since the shapes and spreads of the distributions of Y need to be changed. Such a transformation on Y can be achieved by box-cox transformation. Note that the regression model includes an additional parameter, λ, which needs to be estimated. The Box-Cox procedure uses the method of maximum likelihood to estimate λ

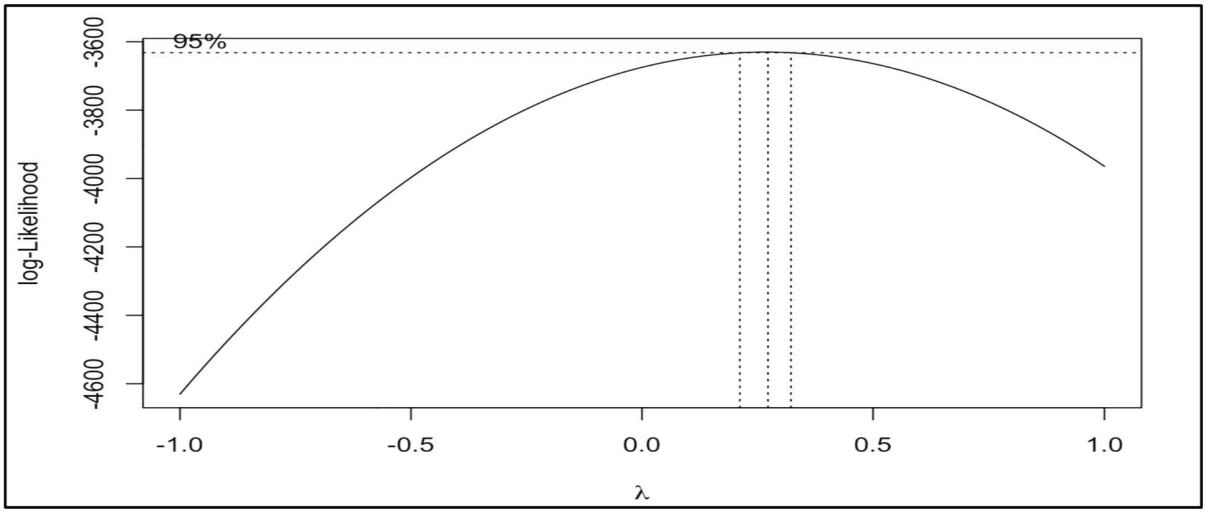

Figure 19. Log Likelihood Plot

Best lambda turns out to be 0.2727273.

Let’s apply the transformation using the best lambda and observe the result.

```
lm(formula = (((charges^bestlambda) - 1)/bestlambda) ~ age +
    I(age^2) + bmi + I(bmi^2) + I(bmi^3) + I(bmi^4) + children +
    factor(smoker) + factor(region) + bmi:factor(smoker) + bmi:factor(region),
    data = insurance_data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2696  -2.2587  -1.0274   0.2482  26.1985

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   5.571e+01  2.927e+01   1.904   0.0572 .
age                           3.946e-01  6.628e-02   5.954 3.36e-09 ***
I(age^2)                     -2.247e-04  8.264e-04  -0.272   0.7857
bmi                          -4.623e+00  3.814e+00  -1.212   0.2256
I(bmi^2)                      2.156e-01  1.808e-01   1.193   0.2332
I(bmi^3)                     -4.090e-03  3.704e-03  -1.104   0.2698
I(bmi^4)                      2.696e-05  2.775e-05   0.972   0.3313
children                      9.832e-01  1.174e-01   8.372  < 2e-16 ***
factor(smoker)yes            -3.187e+00  1.691e+00  -1.885   0.0597 .
factor(region)northwest       7.121e-01  2.109e+00   0.338   0.7356
factor(region)southeast       1.604e+00  2.096e+00   0.765   0.4443
factor(region)southwest      -3.103e-01  2.056e+00  -0.151   0.8800
bmi:factor(smoker)yes         7.714e-01  5.397e-02  14.294  < 2e-16 ***
bmi:factor(region)northwest -5.277e-02  7.103e-02  -0.743   0.4577
bmi:factor(region)southeast -1.042e-01  6.638e-02  -1.569   0.1168
bmi:factor(region)southwest -4.363e-02  6.739e-02  -0.647   0.5175
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.925 on 1322 degrees of freedom
Multiple R-squared:  0.8108,    Adjusted R-squared:  0.8087
F-statistic: 377.7 on 15 and 1322 DF,  p-value: < 2.2e-16
```

Figure 20. Summary of Box-cox transformation model

Here, we observe that the Adjusted R-squared is decreased to 0.8087.

However, let's check if the normality assumption is met after applying the transformation.

Shapiro-Wilk Test

The Test Hypothesis:

Null Hypothesis H(0): The sample data are significantly normally distributed
Alt. Hypothesis H(A): The sample data are not significantly normally distributed

The result is as follows:

```
        Shapiro-Wilk normality test

data:  residuals(bcmodel1)
W = 0.73715, p-value < 2.2e-16
```

Figure 21. Output of Shapiro-Wilk test for Box-cox transformation model

Interpretation:

The Shapiro-Wilk normality test results indicate a lack of normal distribution in the residuals, with a p-value approximately approaching 0, falling below the 0.05 significance level. Consequently, we reject the null hypothesis that assumes normality, suggesting a deviation from the normality assumption even after implementing the Box-Cox transformation. It is evident that the transformed data does not exhibit a normal distribution.

LOG TRANSFORMATION

Let's now conduct a log transformation and assess the normality distribution of the model following this transformation.

```
Call:
lm(formula = log(charges) ~ age + I(age^2) + bmi + I(bmi^2) +
    I(bmi^3) + I(bmi^4) + children + factor(smoker) + factor(region) +
    bmi:factor(smoker) + bmi:factor(region), data = insurance_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90598 -0.18858 -0.06255  0.05281  2.19848

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   8.299e+00  2.548e+00   3.257 0.001154 **
age                           5.379e-02  5.770e-03   9.323  < 2e-16 ***
I(age^2)                     -2.408e-04  7.195e-05  -3.347 0.000839 ***
bmi                          -1.900e-01  3.320e-01  -0.572 0.567158
I(bmi^2)                      9.279e-03  1.574e-02   0.590 0.555498
I(bmi^3)                     -1.760e-04  3.225e-04  -0.546 0.585343
I(bmi^4)                      1.117e-06  2.416e-06   0.462 0.643809
children                      9.246e-02  1.022e-02   9.044  < 2e-16 ***
factor(smoker)yes             1.463e-01  1.472e-01   0.993 0.320704
factor(region)northwest       6.721e-02  1.836e-01   0.366 0.714365
factor(region)southeast       6.586e-02  1.825e-01   0.361 0.718228
factor(region)southwest      -9.844e-02  1.790e-01  -0.550 0.582353
bmi:factor(smoker)yes         4.565e-02  4.698e-03   9.717  < 2e-16 ***
bmi:factor(region)northwest -4.830e-03  6.184e-03  -0.781 0.434901
bmi:factor(region)southeast -7.165e-03  5.779e-03  -1.240 0.215302
bmi:factor(region)southwest -1.554e-03  5.867e-03  -0.265 0.791138
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4287 on 1322 degrees of freedom
Multiple R-squared:  0.7851,    Adjusted R-squared:  0.7826
F-statistic: 321.9 on 15 and 1322 DF,  p-value: < 2.2e-16
```

Figure 22. Summary of model after log transformation

However, let's check if the normality assumption is met after applying the log transformation.

Shapiro-Wilk Test

The Test Hypothesis:

Null Hypothesis H(0): The sample data are significantly normally distributed
Alt. Hypothesis H(A): The sample data are not significantly normally distributed

The result is as follows:

```
        Shapiro-Wilk normality test

data:  residuals(bcmodel2)
W = 0.81538, p-value < 2.2e-16
```

Figure 23. Output of Shapiro-Wilk test for log transformation model

Interpretation:

The Shapiro-Wilk normality test results indicate a lack of normal distribution in the residuals, with a p-value approximately approaching 0, falling below the 0.05 significance level. Consequently, we reject the null hypothesis that assumes normality, suggesting a deviation from the normality assumption even after implementing the log transformation. It is evident that the transformed data does not exhibit a normal distribution.

Despite implementing both Box-Cox and log transformations on the chosen model, the assumption of normality remains unmet. This presents an opportunity for future exploration into alternative techniques to address the persisting issue of non-normality.

PREDICTION

Now, let's utilize the optimal model to predict insurance charges. Predicting the insurance cost for two individuals with similar characteristics, differing only in their smoking status as smoking is one of the most influential variables in the selected model.

Individual 1: Age 55, BMI 39, No Children, resides in Southwestern region, Smoker

```
     fit      lwr      upr
47565.31 38126.78 57003.83
```

Figure 24. Output of Prediction 1

Individual 2: Age 55, BMI 39, No Children, resides in Southwestern region, Non-Smoker

```
     fit      lwr      upr
11646.55 2248.299 21044.79
```

Figure 25. Output of Prediction 2

Interpretation:

For an individual who is 55 years old, has a BMI of 39, no children, and resides in the southwestern region, the anticipated insurance cost is estimated to be $47,565.31 if the person is a smoker ("yes"). On the other hand, if the individual is a non-smoker ("no"), the estimated insurance cost is lower at $11,646.55. The significant difference in these costs highlights the impact of smoking status on the anticipated insurance expenses, with smokers generally incurring higher insurance costs compared to non-smokers.

CONCLUSION

In summary of our analysis results, we were able to determine the predictor variables age, bmi, children, smoker and region were the most significant variables in determining insurance charges. The only predictor variable that was not significant in determining the insurance charges was the sex. The interaction between the (bmi: smoker) and (bmi: region) were the only statistically significant interaction terms as their p-value was less than the significance level of alpha=0.05.

Based on the higher order model we were able to determine our fourth model with the (age^2) and bmi up to the power of four was the best fit model. The best fit model is as follows:

$$\widehat{Charges} = 83950 - 20.71Age + 3.579(\text{Age}^2) - 10290\text{BMI} + 459.2(\text{BMI}^2) - 8.556(\text{BMI}^3) + 0.0564(\text{BMI}^4) + 667.7\text{Children} - 20730\text{Smokeryes} + 9.681\text{Regionnorthwest} + 2852\text{Regionsoutheast} + 1097\text{Regionsouthwest} + 1452\text{BMI} * \text{Smokeryes} - 21.36\text{BMI} * \text{Regionnorthwest} - 126.4\text{BMI} * \text{Regionsoutheast} - 79.66\text{BMI} * \text{Regionsouthwest}$$

The best fit model had an Adjusted R2=0.8454 which was the highest out of all the models and the RMSE=4760.67 which is lowest out of all the models.

We then checked the regression assumptions on our best fit model where all the assumptions were met except for the normality. In order to make this model better we conducted a log and box-cox transformation however, the transformations did not improve the normality assumptions. Therefore, this presents an opportunity for future exploration into alternative techniques to address the persisting issue of non-normality.

DISCUSSION

In figuring out how much insurance might cost, we looked closely at things like age, body mass index (BMI), and whether someone smokes. Turns out, whether you smoke or not makes a big difference in how much you might pay for insurance. Our model, which is like a smart tool that predicts costs, and from our final model where 84.45% of variation in the response variable is explained by the predictors.

However, we faced some challenges, especially when dealing with normality assumption. Even after applying box-cox and log transformations, the normality assumption failed. As we look ahead, we think our model can get even better. We might need more data or consider adding more details to make our predictions more accurate. Thinking about how things change over time, like lifestyle or health habits, could also make our predictions even smarter.

Insurance costs are tricky because they depend on many things. While our model does a good job, we know there's more to it. Future improvements might involve bringing in extra information from different sources to make our predictions even more complete and reliable.

In wrapping up, our project made a solid step forward in figuring out how to predict insurance costs using simple details about people. We're happy with what we've achieved, but we're also excited about making our model even better as we learn more about the complex world of insurance pricing.

REFERENCES

6 reasons healthcare is so expensive in the u. S. (n.d.). Investopedia. Retrieved December 8, 2023, from https://www.investopedia.com/articles/personal-finance/080615/6-reasons-healthcare-so-expensive-us.asp

Canada, S. of. (n.d.). Standing senate committee on social affairs, science and technology(37th parliament, 2nd session). SenCanada. Retrieved December 8, 2023, from https://sencanada.ca/en/committees/soci/

Health insurance statistics and facts – Forbes advisor. (n.d.). Retrieved December 8, 2023, from https://www.forbes.com/advisor/health-insurance/health-insurance-statistics-and-facts/

Medical cost personal datasets. (n.d.). Retrieved December 8, 2023, from https://www.kaggle.com/datasets/mirichoi0218/insurance

U. S. Health care from a global perspective, 2019: Higher spending, worse outcomes? (2020, January 30). https://doi.org/10.26099/7avy-fc29

APPENDIX

R-code:

```
#LIBRARY USED FOR THE PROJECT
library(mctest)
library(lmtest)
library(ggplot2)
library(GGally)
library(MASS)
library(olsrr)
library(leaps)
insurance_data=read.csv("insurance.csv")
head(insurance_data,4)
#FULL ADDITIVE MODEL
full_model=lm(charges~age+factor(sex)+bmi+children+factor(smoker)+factor(region),data=insurance_data)
summary(full_model)
f_test_model=lm(charges~1,data=insurance_data)
anova(f_test_model,full_model)
pairs(~charges+age+factor(sex)+bmi+children+factor(smoker)+factor(region),data=insurance_data)
imcdiag(full_model,method="VIF")
#STEPWISE
step_model=ols_step_both_p(full_model,pent = 0.1, prem = 0.3, details=FALSE)
summary(step_model$model)
#FORWARD
forward_model=ols_step_forward_p(full_model, penter =0.1,details=FALSE)
summary(forward_model$model)
#BACKWARD
backward_model=ols_step_backward_p(full_model, prem = 0.3, details=FALSE)
summary(backward_model$model)
full_model=lm(charges~age+factor(sex)+bmi+children+factor(smoker)+factor(region),data=insurance_data)
reduced_model=lm(charges~age+bmi+children+factor(smoker)+factor(region),data=insurance_data)
```

```
anova(reduced_model,full_model)

summary(reduced_model)

full_model=lm(charges~age+factor(sex)+bmi+children+factor(smoker)+factor(region),data=insurance_dat
a)

best.subset=regsubsets(charges~age+factor(sex)+bmi+children+factor(smoker)+factor(region),data=insu
rance_data)

bestsubset=summary(best.subset)

bestsubset

cp=c(bestsubset$cp)

AdjustedR=c(bestsubset$adjr)

BIC=c(bestsubset$bic)

RMSE=c(bestsubset$rss)

cbind(cp,RMSE,AdjustedR,BIC)

par(mfrow=c(3,2)) # split the plotting panel into a 3 x 2 grid

plot(bestsubset$cp,type = "o",pch=10, xlab="Number of Variables",ylab= "Cp")

plot(bestsubset$rss,type = "o",pch=10, xlab="Number of Variables",ylab= "RMSE")

plot(bestsubset$adjr2,type = "o",pch=10, xlab="Number of Variables",ylab= "Adjusted R^2")

plot(bestsubset$bic,type = "o",pch=10, xlab="Number of Variables",ylab= "BIC")

summary(reduced_model)

#Select the subset of predictors that do the best at meeting some well-defined objective criterion, such as
having the largest R2 value or the smallest MSE, Mallow's Cp or AIC.

ExecSubsets=ols_step_best_subset(full_model, details=TRUE,nv = 7)

summary(ExecSubsets)

# for the output interpretation

rsquare=c(ExecSubsets$rsquare)

AdjustedR=c(ExecSubsets$adjr)

cp=c(ExecSubsets$cp)

AIC=c(ExecSubsets$aic)

cbind(rsquare,AdjustedR,cp,AIC)

par(mfrow=c(2,2)) # split the plotting panel into a 2 x 2 grid

plot(ExecSubsets$cp,type = "o",pch=10, xlab="Number of Variables",ylab= "Cp")

plot(ExecSubsets$rsquare,type = "o",pch=10, xlab="Number of Variables",ylab= "R^2")

plot(ExecSubsets$aic,type = "o",pch=10, xlab="Number of Variables",ylab= "AIC")

plot(ExecSubsets$adjr,type = "o",pch=10, xlab="Number of Variables",ylab= "Adjusted R^2")
```

```
ExecSubsets$predictors

#Full Model

full_model=lm(charges~age+factor(sex)+bmi+children+factor(smoker)+factor(region),data=insurance_data)

#The best additive model is the reduced_model

reduced_model=lm(charges~age+bmi+children+factor(smoker)+factor(region),data=insurance_data)

#Interaction model

interact_model=lm(charges~(age+bmi+children+factor(smoker)+factor(region))^2,data=insurance_data)

summary(interact_model)

interact_final_model=lm(charges~age+bmi+children+factor(smoker)+factor(region)+bmi:factor(smoker)+bmi:factor(region),data=insurance_data)

summary(interact_final_model)

data.frame(Model = c("reduced_model","interact_final_model"),Adjusted_R2 =c(summary(reduced_model)$adj.r.squared,summary(interact_final_model)$adj.r.squared),RMSE=c(summary(reduced_model)$sigma,summary(interact_final_model)$sigma))

#Best Model Thus FAR is the interact_final_model

quad_model=lm(charges~age+I(age^2)+bmi+I(bmi^2)+children+I(children^2)+factor(smoker)+factor(region)+bmi:factor(smoker)+bmi:factor(region),data=insurance_data)

summary(quad_model)

cubic_model=quad_model=lm(charges~age+I(age^2)+I(age^3)+bmi+I(bmi^2)+I(bmi^3)+children+factor(smoker)+factor(region)+bmi:factor(smoker)+bmi:factor(region),data=insurance_data)

summary(cubic_model)

fourth_model=lm(charges~age+I(age^2)+bmi+I(bmi^2)+I(bmi^3)+I(bmi^4)+children+factor(smoker)+factor(region)+bmi:factor(smoker)+bmi:factor(region),data=insurance_data)

summary(fourth_model)

fifth_model=lm(charges~age+I(age^2)+bmi+I(bmi^2)+I(bmi^3)+I(bmi^4)+I(bmi^5)+children+factor(smoker)+factor(region)+bmi:factor(smoker)+bmi:factor(region),data=insurance_data)

summary(fifth_model)

data.frame(Model = c("quad_model","cubic_model","fourth_model"),Adjusted_R2 =c(summary(quad_model)$adj.r.squared,summary(cubic_model)$adj.r.squared,summary(fourth_model)$adj.r.squared),RMSE=c(summary(quad_model)$sigma,summary(cubic_model)$sigma,summary(fourth_model)$sigma))

interact_final_model=lm(charges~age+bmi+children+factor(smoker)+factor(region)+bmi:factor(smoker)+bmi:factor(region),data=insurance_data)

fourth_model=lm(charges~age+I(age^2)+bmi+I(bmi^2)+I(bmi^3)+I(bmi^4)+children+factor(smoker)+factor(region)+bmi:factor(smoker)+bmi:factor(region),data=insurance_data)

anova(fourth_model,interact_final_model)

#BEST MODEL THUS FAR
```

```
fourth_model=lm(charges~age+I(age^2)+bmi+I(bmi^2)+I(bmi^3)+I(bmi^4)+children+factor(smoker)+factor(region)+bmi:factor(smoker)+bmi:factor(region),data=insurance_data)

summary(fourth_model)

#Checking linearity assumptions

#Plotting residuals vs predicted value

ggplot(fourth_model, aes(x=.fitted, y=.resid)) +

  geom_point() +geom_smooth()+

  geom_hline(yintercept = 0)

#Checking Independence association

#Plotting residual vs spatial variable(region)

full_model=lm(charges~age+factor(sex)+bmi+children+factor(smoker)+factor(region),data=insurance_data)

residuals = residuals(full_model)

boxplot(residuals ~ region, data = insurance_data, xlab = "Region", ylab = "Residuals", main = "Residuals vs Region")

#Checking equal Variance (homoscedasticity)

#Plotting a residual plot and scale location

par(mfrow=c(1,2))

plot(fourth_model, which=1)

plot(fourth_model, which=3)

library(lmtest)

bptest(fourth_model)

#Checking Normality assumption

#Plotting histogram and Q-Q plot

par(mfrow=c(1,2))

hist(residuals(fourth_model))

plot(fourth_model, which=2)

#Testing for Normality

shapiro.test(residuals(fourth_model))

#Checking Multicollinearity

library(mctest) #for VIF

# We are only using the main effect independent predictors from the above fourth model for scatterplot and VIF

#From the fourth model the independent predictors are age, bmi, children, smoker and region which will be used to check multicollinearity
```

```
pairs(charges~age+bmi+children+factor(smoker)+factor(region), data=insurance_data)

insurance_vif_model= lm(charges~age+bmi+children+factor(smoker)+factor(region),
data=insurance_data)

imcdiag(insurance_vif_model, method="VIF")

#Checking outliers

# 1.Residuals vs Leverage Plot

plot(fourth_model,which=5)

# 2.Cooks distance

plot(fourth_model,pch=18,col="red",which=c(4))

# 3.Leverage points

lev=hatvalues(fourth_model)

p = length(coef(fourth_model))

n = nrow(insurance_data)

outlier2p = lev[lev>(2*p/n)]

outlier3p = lev[lev>(3*p/n)]

print("h_I>2p/n, outliers are")

print(outlier2p)

print("h_I>3p/n, outliers are")

print(outlier3p)

plot(rownames(insurance_data),lev, main = "Leverage in Insurance Dataset", xlab="observation",
   ylab = "Leverage Value")

abline(h = 2 *p/n, lty = 1)

abline(h = 3 *p/n, lty = 1)

#BOX-COX TRANSFORMATIONS

bc_fourthmodel=boxcox(fourth_model,lambda=seq(-1,1))

#extract best lambda

bestlambda=bc_fourthmodel$x[which(bc_fourthmodel$y==max(bc_fourthmodel$y))]

bestlambda

#BOXCOX TRANSFORMATION MODEL

bcmodel1=lm((((charges^bestlambda)-1)/bestlambda) ~ age + I(age^2) + bmi + I(bmi^2) + I(bmi^3) +
   I(bmi^4) + children + factor(smoker) + factor(region) + bmi:factor(smoker) +
   bmi:factor(region),data=insurance_data)

summary(bcmodel1)

#testing for Normality
```

```
shapiro.test(residuals(bcmodel1))
#LOG TRANSFORMATION MODEL
bcmodel2=lm(log(charges) ~ age + I(age^2) + bmi + I(bmi^2) + I(bmi^3) +
   I(bmi^4) + children + factor(smoker) + factor(region) + bmi:factor(smoker) +
   bmi:factor(region),data=insurance_data)
summary(bcmodel2)
#testing for Normality
shapiro.test(residuals(bcmodel2))
#PREDICTION
#Use the final model to estimate the anticipated insurance cost for an individual who is 55 years old, has a BMI of 39, does not have children, is a smoker, and resides in the southwestern region
smoker_data = data.frame(age=55, bmi=39,children = 0, smoker = "yes", region = "southwest")
predict(fourth_model,smoker_data,interval="predict")
#Use the final model to estimate the anticipated insurance cost for an individual who is 55 years old, has a BMI of 39, does not have children, is a non-smoker, and resides in the southwestern region
non_smoker_data = data.frame(age=55, bmi=39,children = 0, smoker = "no", region = "southwest")
predict(fourth_model,non_smoker_data,interval="predict")
```

ANNEXURES

List of all Figures