Predicting Medical Insurance Costs in the United States

Phuong Nguyen, Nimra Aamir, Tobi Adelodun, Emma Lait

07/12/2021

Chapter 1: Introduction

Healthcare is a necessity that many are not able to afford. High medical costs with or without insurance is a problem that is quite apparent in the United States. In the United States there is a connection between healthcare, healthcare insurance costs, and poverty (Hoffman & Paradise, 2008, p. 149). In fact, findings of a survey conducted in 2013 across 11 countries by a team of researchers shows that many people in the United States will not seek medical treatment even if they have insurance due to medical costs (Schoen et al, 2013, p. 2205). Unlike Canada, there is not partially free coverage for healthcare in the US, so private sectors are responsible for covering 100% of medical costs. With medical costs comes the aspect of insurance. According to Riedel, approximately two-thirds of Americans under age 65 have health insurance coverage (2009, p. 439). It is important to note that there is a large number of people who do not have insurance coverage for health care costs because they are not eligible for certain programs or are not able to "afford nongroup coverage" (Riedel, 2009, p. 439), further providing an outlook into issues that individuals face regularly when it comes to medical costs and insurance. Insurance is one of the primary means used to cover medical costs but there are various factors that influence insurance costs which must be analyzed.

To further examine this problem, statistical research on the topic of medical expenses that impact healthcare insurance in the United States will be carried out. For this research, each individual is categorized by seven variables: age, sex, body mass index (BMI), how many kids are covered under the health insurance, region in the United States, and their charges billed by health insurance. We are looking to see how these specific variables impact insurance costs and medical charges as a whole. To complete this analysis, the Medical Cost Personal Dataset: Insurance Forecast by using Linear Regression was taken from Kaggle under an open database license. The dataset itself is taken and cleaned from the book Machine Learning with R by Brett Lantz.

Chapter 2: Methodology

Our dataset consists of seven variables which are named: age, sex, BMI, children, smoker, region, and charges. Age, BMI, children, and charges are quantitative variable whilst sex, smoker, and region are qualitative variables.

Age = how old the primary beneficiary of the insurance policy is (years)

Sex = the gender of the beneficiary and consists of two factors (male or female)

BMI = body mass index of the beneficiary (kg/m^2)

Children = amount of children that is covered by the policy or the amount of dependents that the primary beneficiary has, ranging from 0 to 5 (children)

Smoker = whether the primary beneficiary is a smoker (yes or no)

Region = the area of the United States of America that the beneficiary lives in (southwest, northwest, southeast, northeast)

The response variable is Charges = cost of health insurance that gets billed to the owner of the health insurance policy (dollars)

To begin the process of building our model, we will first build the full model with all the available variables so that we can have a base model to work with. Throughout this whole process of modelling, we will be using a significance level of 0.05. We will look at the summary of the full model and observe the T-statistic and the P-value of each variable to see which ones are considered significant. By having this base model, we are able to remove and add different variables that would improve the accuracy of the model. Once we remove insignificant variables, we will then compare the original model and our first-order model with an ANOVA test to see whether the variables we removed were worth removing or not by observing the F-statistic and P-value of the ANOVA. After, we will then check the interaction terms by building the interaction model from the first order model. The next step is to see if there are any interaction terms that should be added to the first order model. We will also perform a stepwise model selection which will add the variables one by one, and then remove variables if they are not improving the overall model. Another addition to the model is that we are going to check the higher order models, both squared and cubic, to see if there are any terms that we can add to improve the accuracy of our model. After we have confirmed the most accurate model we will then perform all the necessary assumption tests to ensure it meets the needed assumptions. We will use ggplot to graph the residuals against the fitted values of the linear model to check the linearity assumption, the bptest function for the equal variance assumption, shapiro.test for the normality assumption, the imcdiag function to test for multicollinearity and also test for outliers using the hatvalues function to ensure that there are no values skewing the dataset. Finally after we perform and confirm all the assumptions, we will perform a Box Cox transformation to see whether we can improve the model further or not.

Chapter 3: Main results of the analysis

```
## Rows: 1338 Columns: 7

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): sex, smoker, region
## dbl (4): age, bmi, children, charges

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## # A tibble: 6 x 7
##     age sex      bmi children smoker region    charges
##   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
## 1    19 female  27.9        0 yes    southwest  16885.
## 2    18 male    33.8        1 no     southeast   1726.
## 3    28 male    33          3 no     southeast   4449.
## 4    33 male    22.7        0 no     northwest  21984.
## 5    32 male    28.9        0 no     northwest   3867.
## 6    31 female  25.7        0 no     southeast   3757.
```

```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000
```

```
##     smoker             region            charges
##  Length:1338        Length:1338        Min.   : 1122
##  Class :character   Class :character   1st Qu.: 4740
##  Mode  :character   Mode  :character   Median : 9382
##                                        Mean   :13270
##                                        3rd Qu.:16640
##                                        Max.   :63770
```

```
## integer(0)
```

```
fullmodel <- lm(charges~age+factor(sex)+
                  bmi+children+factor(smoker)+
                  factor(region), data=insurance)
summary(fullmodel)
```

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

We built an initial first-order model with all of the independent variables and utilized the stepwise regression in order to select the important ones to be included. The maximum model was specified as below:

$$Y_{charges} = \beta_0 + \beta_1 X_{Age} + \beta_2 X_{BMI} + \beta_3 X_{Children} + \beta_4 X_{Smoker} + \beta_5 X_{Region} + \beta_6 X_{Sex}$$

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
stepmod=ols_step_both_p(fullmodel, pent=0.1, prem=0.3, details=TRUE)
```

```
## Stepwise Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. age
## 2. factor(sex)
## 3. bmi
## 4. children
## 5. factor(smoker)
## 6. factor(region)
##
## We are selecting variables based on p value...
##
##
## Stepwise Selection: Step 1
##
## - factor(smoker) added
##
##                            Model Summary
## ----------------------------------------------------------------------
## R                       0.787       RMSE                  7470.216
## R-Squared               0.620       Coef. Var               56.292
## Adj. R-Squared          0.619       MSE                 55804130.200
## Pred R-Squared          0.618       MAE                     5662.090
## ----------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                       ANOVA
## ---------------------------------------------------------------------------------------
##                        Sum of
##                       Squares          DF         Mean Square       F          Sig.
## ---------------------------------------------------------------------------------------
## Regression    121519903621.668           1    121519903621.668    2177.615    0.0000
## Residual       74554317946.700        1336        55804130.200
## Total         196074221568.368        1337
## ---------------------------------------------------------------------------------------
##
##                                    Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##             model         Beta    Std. Error    Std. Beta      t        Sig         lower        upp
## ----------------------------------------------------------------------------------------------------
##       (Intercept)     8434.268       229.014                 36.829    0.000     7985.002     8883.53
## factor(smoker)yes    23615.964       506.075        0.787    46.665    0.000    22623.175    24608.75
## ----------------------------------------------------------------------------------------------------
##
```

```
##
##
## Stepwise Selection: Step 2
##
## - age added
##
##                            Model Summary
## ------------------------------------------------------------------
## R                        0.849       RMSE                  6396.752
## R-Squared                0.721       Coef. Var               48.203
## Adj. R-Squared           0.721       MSE                 40918439.071
## Pred R-Squared           0.720       MAE                     4122.078
## ------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                      ANOVA
## ---------------------------------------------------------------------------------------
##                       Sum of
##                      Squares          DF        Mean Square       F          Sig.
## ---------------------------------------------------------------------------------------
## Regression    141448105408.046           2    70724052704.023    1728.415    0.0000
## Residual       54626116160.323        1335       40918439.071
## Total         196074221568.368        1337
## ---------------------------------------------------------------------------------------
##
##                                        Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##             model         Beta    Std. Error    Std. Beta      t        Sig         lower        upp
## ----------------------------------------------------------------------------------------------------
##       (Intercept)    -2391.626       528.302                -4.527    0.000    -3428.019    -1355.23
## factor(smoker)yes    23855.305       433.488        0.795    55.031    0.000    23004.912    24705.69
##               age      274.871        12.455        0.319    22.069    0.000      250.437      299.30
## ----------------------------------------------------------------------------------------------------
##
##
##
##                            Model Summary
## ------------------------------------------------------------------
## R                        0.849       RMSE                  6396.752
## R-Squared                0.721       Coef. Var               48.203
## Adj. R-Squared           0.721       MSE                 40918439.071
## Pred R-Squared           0.720       MAE                     4122.078
## ------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                      ANOVA
## ---------------------------------------------------------------------------------------
##                       Sum of
##                      Squares          DF        Mean Square       F          Sig.
## ---------------------------------------------------------------------------------------
```

```
## Regression    141448105408.046          2    70724052704.023    1728.415    0.0000
## Residual       54626116160.323       1335       40918439.071
## Total         196074221568.368       1337
## ----------------------------------------------------------------------------------
##
##                                        Parameter Estimates
## ---------------------------------------------------------------------------------------------------
##             model         Beta    Std. Error    Std. Beta      t        Sig         lower        upp
## ---------------------------------------------------------------------------------------------------
##       (Intercept)    -2391.626       528.302                -4.527    0.000    -3428.019    -1355.2
## factor(smoker)yes    23855.305       433.488        0.795    55.031    0.000    23004.912    24705.6
##               age      274.871        12.455        0.319    22.069    0.000      250.437      299.3
## ---------------------------------------------------------------------------------------------------
##
##
##
## Stepwise Selection: Step 3
##
## - bmi added
##
##                            Model Summary
## ------------------------------------------------------------------
## R                        0.865       RMSE                 6092.319
## R-Squared                0.747       Coef. Var              45.909
## Adj. R-Squared           0.747       MSE                37116356.457
## Pred R-Squared           0.746       MAE                    4216.776
## ------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                      ANOVA
## ----------------------------------------------------------------------------------
##                        Sum of
##                       Squares          DF        Mean Square       F          Sig.
## ----------------------------------------------------------------------------------
## Regression         1.46561e+11           3    48853667351.396     1316.23    0.0000
## Residual       49513219514.179        1334       37116356.457
## Total         196074221568.368        1337
## ----------------------------------------------------------------------------------
##
##                                        Parameter Estimates
## ---------------------------------------------------------------------------------------------------
##             model          Beta    Std. Error    Std. Beta       t        Sig          lower        u
## ---------------------------------------------------------------------------------------------------
##       (Intercept)    -11676.830       937.569               -12.454    0.000    -13516.100    -983
## factor(smoker)yes     23823.684       412.867        0.794     57.703    0.000     23013.746    2463
##               age       259.547        11.934        0.301     21.748    0.000       236.136      28
##               bmi       322.615        27.487        0.162     11.737    0.000       268.692      37
## ---------------------------------------------------------------------------------------------------
##
##
##
##                            Model Summary
```

```
## ----------------------------------------------------------------------
## R                         0.865       RMSE                   6092.319
## R-Squared                 0.747       Coef. Var                45.909
## Adj. R-Squared            0.747       MSE                  37116356.457
## Pred R-Squared            0.746       MAE                      4216.776
## ----------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                      ANOVA
## ---------------------------------------------------------------------------------------
##                         Sum of
##                        Squares          DF        Mean Square       F          Sig.
## ---------------------------------------------------------------------------------------
## Regression         1.46561e+11           3    48853667351.396     1316.23    0.0000
## Residual       49513219514.179        1334       37116356.457
## Total         196074221568.368        1337
## ---------------------------------------------------------------------------------------
##
##                                          Parameter Estimates
## ----------------------------------------------------------------------------------------------------------
##             model           Beta    Std. Error    Std. Beta       t        Sig          lower        u
## ----------------------------------------------------------------------------------------------------------
##       (Intercept)    -11676.830       937.569                -12.454    0.000    -13516.100    -983
## factor(smoker)yes     23823.684       412.867        0.794    57.703    0.000     23013.746    2463
##               age       259.547        11.934        0.301    21.748    0.000       236.136      28
##               bmi       322.615        27.487        0.162    11.737    0.000       268.692      37
## ----------------------------------------------------------------------------------------------------------
##
##
##
## Stepwise Selection: Step 4
##
## - children added
##
##                             Model Summary
## ----------------------------------------------------------------------
## R                         0.866       RMSE                   6067.787
## R-Squared                 0.750       Coef. Var                45.724
## Adj. R-Squared            0.749       MSE                  36818042.098
## Pred R-Squared            0.747       MAE                      4178.681
## ----------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                      ANOVA
## ---------------------------------------------------------------------------------------
##                         Sum of
##                        Squares          DF        Mean Square       F          Sig.
## ---------------------------------------------------------------------------------------
## Regression    146995771451.750           4    36748942862.937    998.123    0.0000
## Residual       49078450116.619        1333       36818042.098
```

```
## Total         196074221568.368        1337
## ----------------------------------------------------------------------------
##
##                                        Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##             model          Beta    Std. Error    Std. Beta       t        Sig          lower
## ----------------------------------------------------------------------------------------------------
##       (Intercept)    -12102.769       941.984                -12.848    0.000    -13950.702    -1025
## factor(smoker)yes     23811.400       411.220        0.794     57.904    0.000     23004.692     246
##               age       257.850        11.896        0.299     21.675    0.000       234.512       28
##               bmi       321.851        27.378        0.162     11.756    0.000       268.143       3
##          children       473.502       137.792        0.047      3.436    0.001       203.190       74
## ----------------------------------------------------------------------------------------------------
##
##
##
##                            Model Summary
## ----------------------------------------------------------------
## R                       0.866       RMSE                 6067.787
## R-Squared               0.750       Coef. Var              45.724
## Adj. R-Squared          0.749       MSE                 36818042.098
## Pred R-Squared          0.747       MAE                     4178.681
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                     ANOVA
## ----------------------------------------------------------------------------
##                       Sum of
##                      Squares          DF        Mean Square       F         Sig.
## ----------------------------------------------------------------------------
## Regression    146995771451.750           4    36748942862.937    998.123    0.0000
## Residual       49078450116.619        1333       36818042.098
## Total         196074221568.368        1337
## ----------------------------------------------------------------------------
##
##                                        Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##             model          Beta    Std. Error    Std. Beta       t        Sig          lower
## ----------------------------------------------------------------------------------------------------
##       (Intercept)    -12102.769       941.984                -12.848    0.000    -13950.702    -1025
## factor(smoker)yes     23811.400       411.220        0.794     57.904    0.000     23004.692     246
##               age       257.850        11.896        0.299     21.675    0.000       234.512       28
##               bmi       321.851        27.378        0.162     11.756    0.000       268.143       3
##          children       473.502       137.792        0.047      3.436    0.001       203.190       74
## ----------------------------------------------------------------------------------------------------
##
##
##
## Stepwise Selection: Step 5
##
## - factor(region) added
##
```

```
##                             Model Summary
## ------------------------------------------------------------------
## R                       0.867       RMSE                 6060.178
## R-Squared               0.751       Coef. Var              45.667
## Adj. R-Squared          0.750       MSE                36725751.333
## Pred R-Squared          0.747       MAE                    4171.710
## ------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                    ANOVA
## --------------------------------------------------------------------------------
##                         Sum of
##                        Squares          DF        Mean Square       F         Sig.
## --------------------------------------------------------------------------------
## Regression         1.47229e+11           7    21032710327.911    572.697    0.0000
## Residual       48845249272.988        1330       36725751.333
## Total         196074221568.368        1337
## --------------------------------------------------------------------------------
##
##                                          Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##                    model          Beta    Std. Error    Std. Beta       t        Sig          lower
## ----------------------------------------------------------------------------------------------------
##              (Intercept)    -11990.270       978.762                -12.250    0.000    -13910.355
##       factor(smoker)yes     23836.301       411.856        0.795     57.875    0.000     23028.341
##                      age       256.974        11.891        0.298     21.610    0.000       233.646
##                      bmi       338.665        28.559        0.171     11.858    0.000       282.639
##                 children       474.566       137.740        0.047      3.445    0.001       204.355
## factor(region)northwest      -352.182       476.120       -0.012     -0.740    0.460     -1286.211
## factor(region)southeast     -1034.360       478.537       -0.038     -2.162    0.031     -1973.130
## factor(region)southwest      -959.375       477.778       -0.034     -2.008    0.045     -1896.656
## ----------------------------------------------------------------------------------------------------
##
##
##
##                             Model Summary
## ------------------------------------------------------------------
## R                       0.867       RMSE                 6060.178
## R-Squared               0.751       Coef. Var              45.667
## Adj. R-Squared          0.750       MSE                36725751.333
## Pred R-Squared          0.747       MAE                    4171.710
## ------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                    ANOVA
## --------------------------------------------------------------------------------
##                         Sum of
##                        Squares          DF        Mean Square       F         Sig.
## --------------------------------------------------------------------------------
## Regression         1.47229e+11           7    21032710327.911    572.697    0.0000
```

```
## Residual       48845249272.988        1330       36725751.333
## Total         196074221568.368        1337
## ----------------------------------------------------------------------------
##
##                                          Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##                    model          Beta    Std. Error    Std. Beta       t        Sig          lower
## ----------------------------------------------------------------------------------------------------
##              (Intercept)    -11990.270       978.762                -12.250    0.000    -13910.355
##       factor(smoker)yes     23836.301       411.856        0.795     57.875    0.000     23028.341
##                     age       256.974        11.891        0.298     21.610    0.000       233.646
##                     bmi       338.665        28.559        0.171     11.858    0.000       282.639
##                children       474.566       137.740        0.047      3.445    0.001       204.355
## factor(region)northwest      -352.182       476.120       -0.012     -0.740    0.460     -1286.211
## factor(region)southeast     -1034.360       478.537       -0.038     -2.162    0.031     -1973.130
## factor(region)southwest      -959.375       477.778       -0.034     -2.008    0.045     -1896.656
## ----------------------------------------------------------------------------------------------------
##
##
##
## No more variables to be added/removed.
##
##
## Final Model Output
## ------------------
##
##                            Model Summary
## ------------------------------------------------------------------
## R                       0.867       RMSE                 6060.178
## R-Squared               0.751       Coef. Var              45.667
## Adj. R-Squared          0.750       MSE                36725751.333
## Pred R-Squared          0.747       MAE                    4171.710
## ------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                      ANOVA
## ----------------------------------------------------------------------------
##                        Sum of
##                       Squares          DF        Mean Square       F         Sig.
## ----------------------------------------------------------------------------
## Regression         1.47229e+11           7    21032710327.911    572.697    0.0000
## Residual       48845249272.988        1330       36725751.333
## Total         196074221568.368        1337
## ----------------------------------------------------------------------------
##
##                                          Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##                    model          Beta    Std. Error    Std. Beta       t        Sig          lower
## ----------------------------------------------------------------------------------------------------
##              (Intercept)    -11990.270       978.762                -12.250    0.000    -13910.355
##       factor(smoker)yes     23836.301       411.856        0.795     57.875    0.000     23028.341
##                     age       256.974        11.891        0.298     21.610    0.000       233.646
```

```
##                     bmi       338.665        28.559        0.171     11.858    0.000       282.639
##                children       474.566       137.740        0.047      3.445    0.001       204.355
## factor(region)northwest      -352.182       476.120       -0.012     -0.740    0.460     -1286.211
## factor(region)southeast     -1034.360       478.537       -0.038     -2.162    0.031     -1973.130
## factor(region)southwest      -959.375       477.778       -0.034     -2.008    0.045     -1896.656
## ------------------------------------------------------------------------------------------------------
```

```
summary(stepmod$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11367.2  -2835.4   -979.7   1361.9  29935.5
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -11990.27     978.76 -12.250  < 2e-16 ***
## factor(smoker)yes        23836.30     411.86  57.875  < 2e-16 ***
## age                        256.97      11.89  21.610  < 2e-16 ***
## bmi                        338.66      28.56  11.858  < 2e-16 ***
## children                   474.57     137.74   3.445 0.000588 ***
## factor(region)northwest   -352.18     476.12  -0.740 0.459618
## factor(region)southeast  -1034.36     478.54  -2.162 0.030834 *
## factor(region)southwest   -959.37     477.78  -2.008 0.044846 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1330 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7496
## F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

From the selection, the independent variables that produce the largest absolute t-values were declared. We used the default p-values such as any variables with a p-value lower than 0.1 - the entering threshold will enter the model and higher than p-value = 0.3 will be removed. The output from this procedure suggests the variable Sex to be dropped from the model. Therefore, we decided to include the main effects of both quantitative variables and dummy variables as such: Smoker, Age, BMI, Children, Region in our first-order model.

Individual T-tests:

Hypothesis statement: (the model without Sex variable) $H_0 : \beta_i = 0$ $H_a : \beta_i \neq 0$ (i = Age, BMI, Children, Smoker, Region)

We used the individual T-tests to determine what the best predictors are on the significance level $\alpha = 0.05$. From the output, the p-values of Age, BMI, Children, Smoker were less than $\alpha = 0.05$ which indicate that we should reject the null hypothesis and these variables are significant in the model. The Region variable had one category above our specified α among 4 categories. Therefore, we decided to keep Region as one of the predictors from this step for further comparison with interaction and higher order terms. The model from the individual T-tests is:

$Y_{Charges} = \beta_0 + \beta_1 X_{Age} + \beta_2 X_{BMI} + \beta_3 X_{Children} + \beta_4 X_{Smoker} + \beta_5 X_{Region}$

```r
fullmodel <-lm(charges~age+bmi+children+
                 factor(smoker)+factor(region),
               data=insurance)  #full model without Sex variable
summary(fullmodel)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + factor(smoker) +
##     factor(region), data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11367.2  -2835.4   -979.7   1361.9  29935.5
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -11990.27     978.76 -12.250  < 2e-16 ***
## age                         256.97      11.89  21.610  < 2e-16 ***
## bmi                         338.66      28.56  11.858  < 2e-16 ***
## children                    474.57     137.74   3.445 0.000588 ***
## factor(smoker)yes         23836.30     411.86  57.875  < 2e-16 ***
## factor(region)northwest    -352.18     476.12  -0.740 0.459618
## factor(region)southeast   -1034.36     478.54  -2.162 0.030834 *
## factor(region)southwest    -959.37     477.78  -2.008 0.044846 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1330 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7496
## F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

Partial F-test:

Hypothesis statement: (for Region variable) $H_0 : \beta_{Region} = 0$ $H_a : \beta_{Region} \neq 0$

```r
#F-test for model with or without region
firstordermodel <- lm(charges~age+bmi+children+factor(smoker),
                      data=insurance)  #without Region variable
summary(firstordermodel)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + factor(smoker),
##     data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11897.9  -2920.8   -986.6   1392.2  29509.6
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -12102.77     941.98 -12.848  < 2e-16 ***
## age                  257.85      11.90  21.675  < 2e-16 ***
```

```
## bmi                  321.85      27.38  11.756  < 2e-16 ***
## children             473.50     137.79   3.436 0.000608 ***
## factor(smoker)yes  23811.40     411.22  57.904  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6068 on 1333 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7489
## F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

```
firstordermodel1 <- lm(charges~age+bmi+children+factor(smoker)+factor(region),
                       data=insurance)   #with Region variable
summary(firstordermodel1)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + factor(smoker) +
##     factor(region), data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11367.2  -2835.4   -979.7   1361.9  29935.5
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -11990.27     978.76 -12.250  < 2e-16 ***
## age                        256.97      11.89  21.610  < 2e-16 ***
## bmi                        338.66      28.56  11.858  < 2e-16 ***
## children                   474.57     137.74   3.445 0.000588 ***
## factor(smoker)yes        23836.30     411.86  57.875  < 2e-16 ***
## factor(region)northwest   -352.18     476.12  -0.740 0.459618
## factor(region)southeast  -1034.36     478.54  -2.162 0.030834 *
## factor(region)southwest   -959.37     477.78  -2.008 0.044846 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1330 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7496
## F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

```
anova(firstordermodel, firstordermodel1)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ age + bmi + children + factor(smoker)
## Model 2: charges ~ age + bmi + children + factor(smoker) + factor(region)
##   Res.Df        RSS Df Sum of Sq      F  Pr(>F)
## 1   1333 4.9078e+10
## 2   1330 4.8845e+10  3 233200844 2.1166 0.09631 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The partial F-test was used in order to check the significance of the Region variable. The goal of this step is to investigate the contribution of this predictor individually. We defined the full model with all the

predictors and the reduced model with the whole set of predictors less the Region one. From the analysis of variance for comparing between these two models, the output shows that $F - value = 2.1166$ with df $= 1330$ ($p - value = 0.09631 > \alpha = 0.05$), indicating that we should clearly not to reject the null hypothesis. We should definitely drop the Region variable off the model. At this point, the model we have is:

$Y_{Charges} = \beta_0 + \beta_1 X_{Age} + \beta_2 X_{BMI} + \beta_3 X_{Children} + \beta_4 X_{Smoker}$

Interaction terms individual T-tests:

Hypothesis statement: $H_0 : \beta_i = 0$ $H_a : \beta_i \neq 0$ (i = Age*BMI,* Age*Children*, Age*Smoker,* BMI*Children*, BMI*Smoker,* Children*Smoker*)

```
itrmodel <- lm(charges~(age+bmi+children+factor(smoker))^2, data=insurance)
summary(itrmodel)
```

```
##
## Call:
## lm(formula = charges ~ (age + bmi + children + factor(smoker))^2,
##     data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13996.3  -1947.6  -1331.5   -406.4  29570.2
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -6.745e+02  2.106e+03  -0.320    0.749
## age                          2.055e+02  4.963e+01   4.140  3.7e-05 ***
## bmi                         -6.339e+01  6.756e+01  -0.938    0.348
## children                     6.957e+02  6.341e+02   1.097    0.273
## factor(smoker)yes           -1.983e+04  1.861e+03 -10.651  < 2e-16 ***
## age:bmi                      1.912e+00  1.561e+00   1.225    0.221
## age:children                 1.201e+00  8.527e+00   0.141    0.888
## age:factor(smoker)yes       -9.141e-01  2.384e+01  -0.038    0.969
## bmi:children                -5.334e+00  1.863e+01  -0.286    0.775
## bmi:factor(smoker)yes        1.437e+03  5.317e+01  27.029  < 2e-16 ***
## children:factor(smoker)yes -3.858e+02  2.841e+02  -1.358    0.175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4874 on 1327 degrees of freedom
## Multiple R-squared:  0.8392, Adjusted R-squared:  0.838
## F-statistic: 692.8 on 10 and 1327 DF,  p-value: < 2.2e-16
```

From the output of the T-tests for the interaction terms, there is only one term that is significant for the charge of medical insurance which is between BMI and Smoker (p-value < 0.05). The other interaction terms have small t-value and high p-value compared to our significance level. Therefore, we fail to reject the null hypothesis. The model with the interaction term is shown below:

```
# best interaction term
bestitrmodel <- lm(charges~age+bmi+factor(smoker)+
                     bmi*factor(smoker)+children,
                   data=insurance)
summary(bestitrmodel)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + factor(smoker) + bmi * factor(smoker) +
##     children, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14598.6  -1924.4  -1321.4   -465.6  29892.4
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -2729.002    831.270  -3.283  0.00105 **
## age                      264.948      9.553  27.735  < 2e-16 ***
## bmi                        5.656     24.873   0.227  0.82014
## factor(smoker)yes     -20194.709   1654.505 -12.206  < 2e-16 ***
## children                 508.924    110.615   4.601 4.61e-06 ***
## bmi:factor(smoker)yes   1433.788     52.823  27.143  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4871 on 1332 degrees of freedom
## Multiple R-squared:  0.8388, Adjusted R-squared:  0.8382
## F-statistic:  1387 on 5 and 1332 DF,  p-value: < 2.2e-16
```

$$Y_{Charges} = \beta_0 + \beta_1 X_{Age} + \beta_2 X_{BMI} + \beta_3 X_{Children} + \beta_4 X_{Smoker} + \beta_5 X_{BMI} * X_{Smoker}$$

Higher order model:

Hypothesis statement: $H_0 : \beta_i = 0$ $H_a : \beta_i \neq 0$ (i = Age^2, BMI^2, Children^2) (i = Age^2, BMI^2, Age^3, BMI^3) (i = Age^2, BMI^2, BMI^3)

```
hm <- lm(charges~age+bmi+factor(smoker)+children +
           I(age^2) + I(bmi^2) +
           I(children^2), data=insurance)
hm1 <- lm(charges~age+bmi+factor(smoker)+children +
            I(age^2) + I(bmi^2) + I(age^3) +
            I(bmi^3), data=insurance)
hm2 <- lm(charges~age+bmi+factor(smoker)+children +
            I(age^2) + I(bmi^2) +
            I(bmi^3), data=insurance)
```

We want to check for the significance of the higher order terms. Firstly, we included the quadratic terms of three quantitative variables which are Age, BMI and Children. All of the values are significant except the higher order of the Children variable (p-value = 0.0661). Therefore, we removed the higher order of Children and increased the order of Age and BMI to cubic terms. The cubic term of Age was insignificant at p-value = 0.7437. Therefore, we finalized our higher order terms as such: Age^2 (t = 3.838, p-value = 0.00013) BMI^2 (t = 2.020, p-value = 0.04361) BMI^3 (t = -2.253, p-value = 0.0241) The model with higher order terms:

$$Y_{Charges} = \beta_0 + \beta_1 X_{Age} + \beta_2 X_{BMI} + \beta_3 X_{Children} + \beta_4 X_{Smoker} + \beta_5 X_{Age^2} + \beta_6 X_{BMI^2} + \beta_7 X_{BMI^3}$$

Interaction terms and higher order model:

Hypothesis statement: $H_0 : \beta_i = 0$ $H_a : \beta_i \neq 0$ (i = BMI*factor(Smoker))

```r
highermodel <- lm(charges~age+bmi+factor(smoker)+
                    children + I(age^2) + I(bmi^2) +
                    I(bmi^3) + bmi*factor(smoker),
                  data=insurance)
summary(highermodel)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + factor(smoker) + children +
##     I(age^2) + I(bmi^2) + I(bmi^3) + bmi * factor(smoker), data = insurance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8511.4 -1849.2 -1318.2  -635.3 30538.3
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.752e+04  8.667e+03   3.175 0.001531 **
## age                   -2.621e+01  6.435e+01  -0.407 0.683833
## bmi                   -2.774e+03  8.335e+02  -3.328 0.000898 ***
## factor(smoker)yes     -2.048e+04  1.628e+03 -12.577  < 2e-16 ***
## children               6.683e+02  1.139e+02   5.867 5.60e-09 ***
## I(age^2)               3.645e+00  8.027e-01   4.541 6.11e-06 ***
## I(bmi^2)               9.714e+01  2.614e+01   3.716 0.000211 ***
## I(bmi^3)              -1.085e+00  2.666e-01  -4.068 5.01e-05 ***
## bmi:factor(smoker)yes  1.444e+03  5.198e+01  27.781  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4788 on 1329 degrees of freedom
## Multiple R-squared:  0.8446, Adjusted R-squared:  0.8437
## F-statistic: 903.2 on 8 and 1329 DF,  p-value: < 2.2e-16
```

```r
anova(hm2, highermodel)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ age + bmi + factor(smoker) + children + I(age^2) +
##     I(bmi^2) + I(bmi^3)
## Model 2: charges ~ age + bmi + factor(smoker) + children + I(age^2) +
##     I(bmi^2) + I(bmi^3) + bmi * factor(smoker)
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1   1330 4.8151e+10
## 2   1329 3.0462e+10  1 1.769e+10 771.78 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We conducted the ANOVA test to confirm the contribution of the interaction term with the higher order model. The reduced model is the one with the main effects and without the interaction term. Meanwhile, the full model includes the main effects, the interaction term and higher order terms. The result of ANOVA with $F = 771.78$ and p-value < 0.05 indicates that we should reject the hypothesis. We finalized our model with higher order and interaction terms as shown below:

$Y_{Charges} = \beta_0 + \beta_1 X_{Age} + \beta_2 X_{BMI} + \beta_3 X_{Children} + \beta_4 X_{Smoker} + \beta_5 X_{Age^2} + \beta_6 X_{BMI^2} + \beta_7 X_{BMI^3} + \beta_8 X_{BMI} * X_{Smoker}$

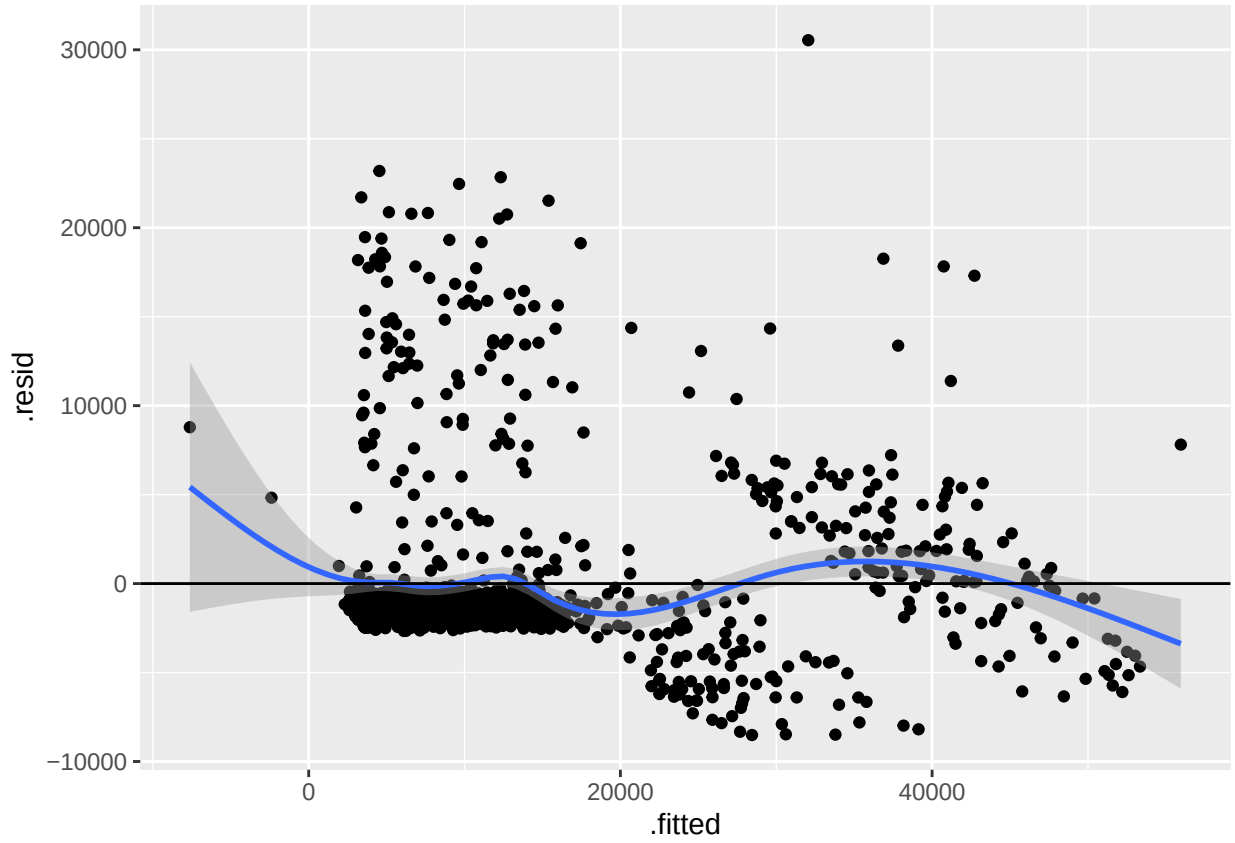
Multiple regression assumptions

1. Linearity assumption:

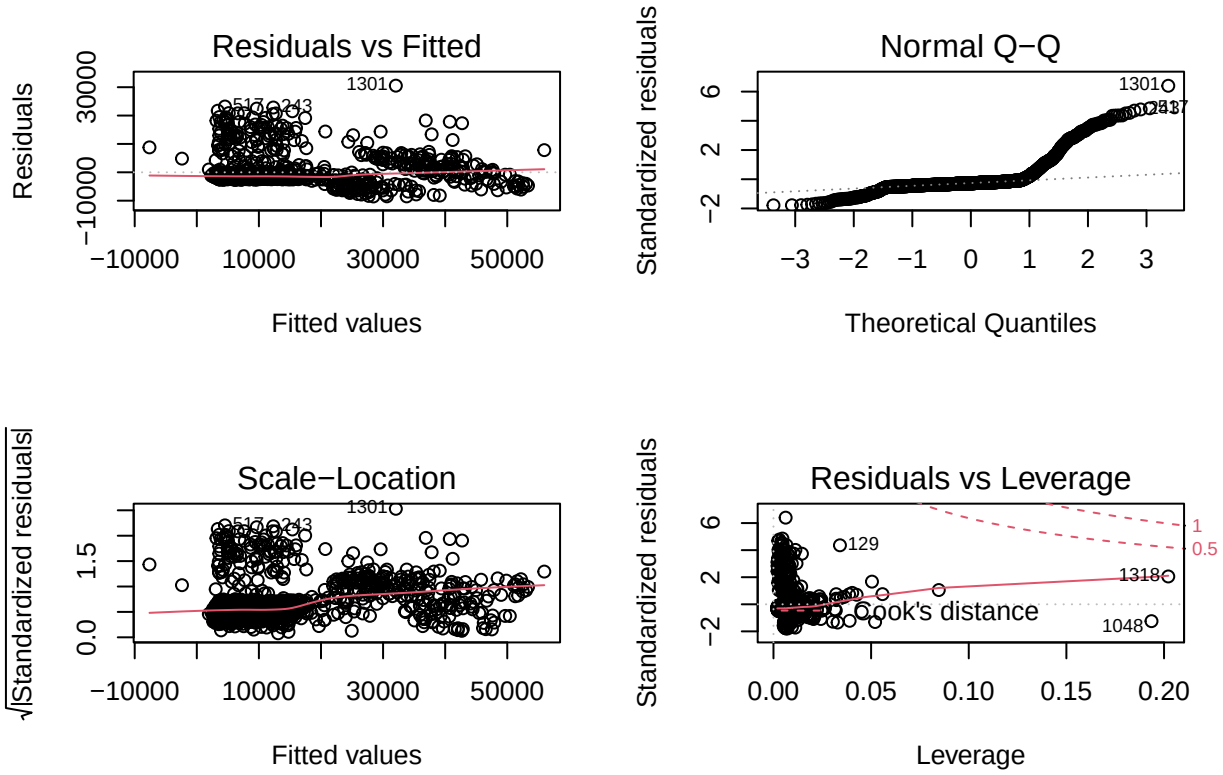
The best predicted model assumes that there is a straight-line relationship between all the predictors and the response. We expect to see the linear pattern when plotting the residuals and fitted values from the model. Using the residual plot as shown below, there are no discernible patterns detected. Therefore, this model passes the linearity assumption.

```
library(ggplot2)
ggplot(highermodel, aes(x=.fitted, y=.resid))+geom_point()+geom_smooth()+geom_hline(yintercept = 0)
```

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




```
par(mfrow=c(2,2))
plot(highermodel)
```



2. Equal variance assumption:

H_0 : Heteroscedasticity is not present H_a : Heteroscedasticity is present

The model is also assumed to have the error terms that have a constant variance. In order to verify whether our model is homoscedastic, a scale-location between fitted value and standardized residuals as well as the studentized Breusch-Pagan test were utilized. From the plot, we can see the horizontal line with equally spread points. The output of the Breusch-Pagan (BP = 6.8338 and p-value = 0.5547) indicates that we should not reject the null hypothesis. Therefore, it suggests that the predicted model passes the assumption and is homoscedastic.

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.1.2
```

```
## Loading required package: zoo
```

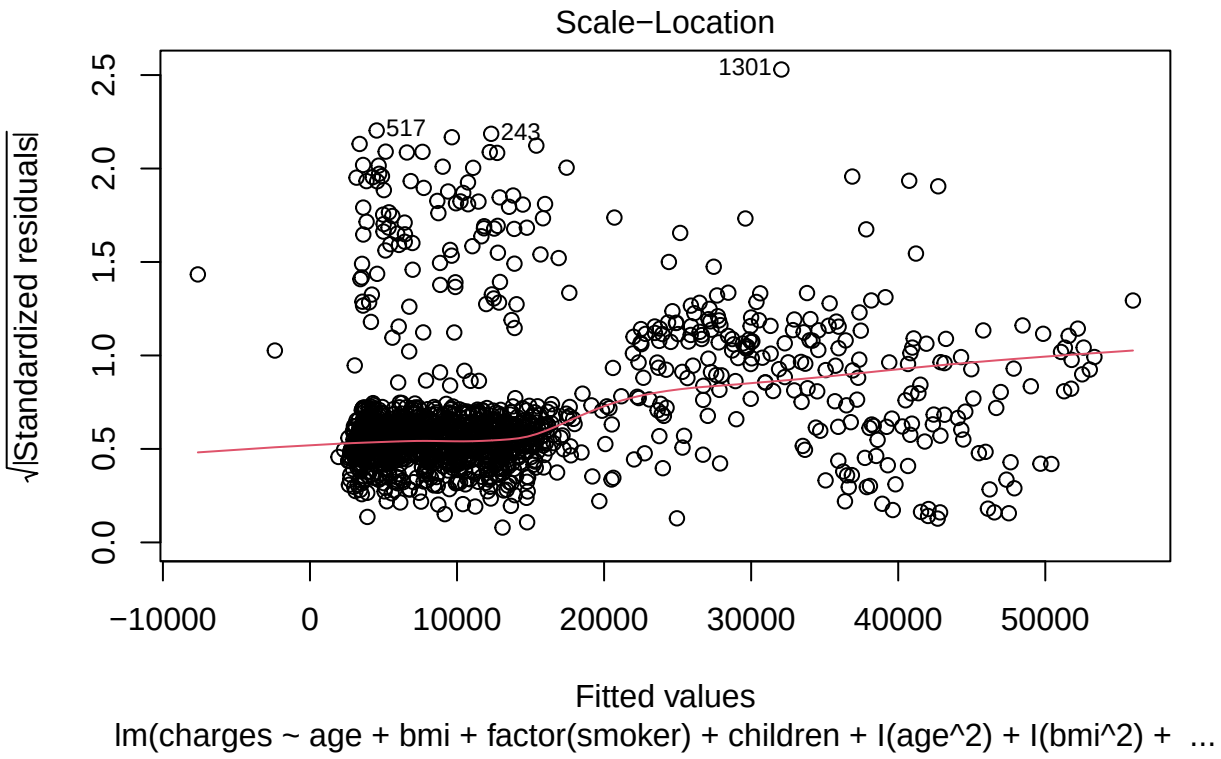
```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(highermodel)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  highermodel
## BP = 6.8338, df = 8, p-value = 0.5547
```

```
plot(highermodel, which=3)
```



3. Multicollinearity test:

There is a chance that independent variables are correlated with each other. To test for multicollinearity, we performed the test with the variance inflation factor (VIF). Since our model has 3 higher-order terms and 1 interaction term, the high values of VIF are more likely to happen. The diagnostic shows that the multicollinearity may be due to the variables with higher order, the categorical variable Smoker and the interaction term between BMI and Smoker. Therefore, we are safe to ignore the high VIF and conclude that there are no extreme multicollinearities detected.

```
library(mctest)
imcdiag(highermodel, method="VIF")
```

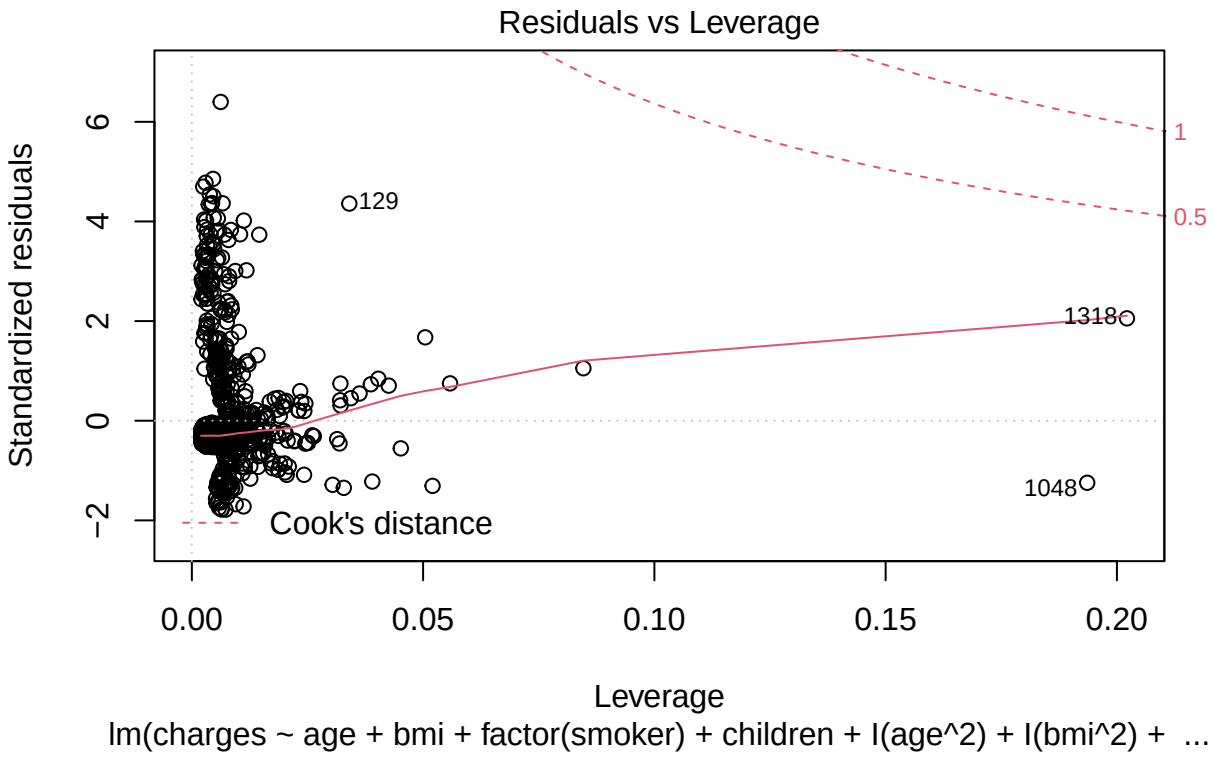
```
##
```

```
## Call:
## imcdiag(mod = highermodel, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                              VIF detection
## age                      47.6852         1
## bmi                    1506.8357         1
## factor(smoker)yes        25.1962         1
## children                  1.0999         0
## I(age^2)                 47.6361         1
## I(bmi^2)               5996.5695         1
## I(bmi^3)               1574.5771         1
## bmi:factor(smoker)yes    25.5082         1
##
## Multicollinearity may be due to age bmi factor(smoker)yes I(age^2) I(bmi^2) I(bmi^3) bmi:factor(smok
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ====================================
```

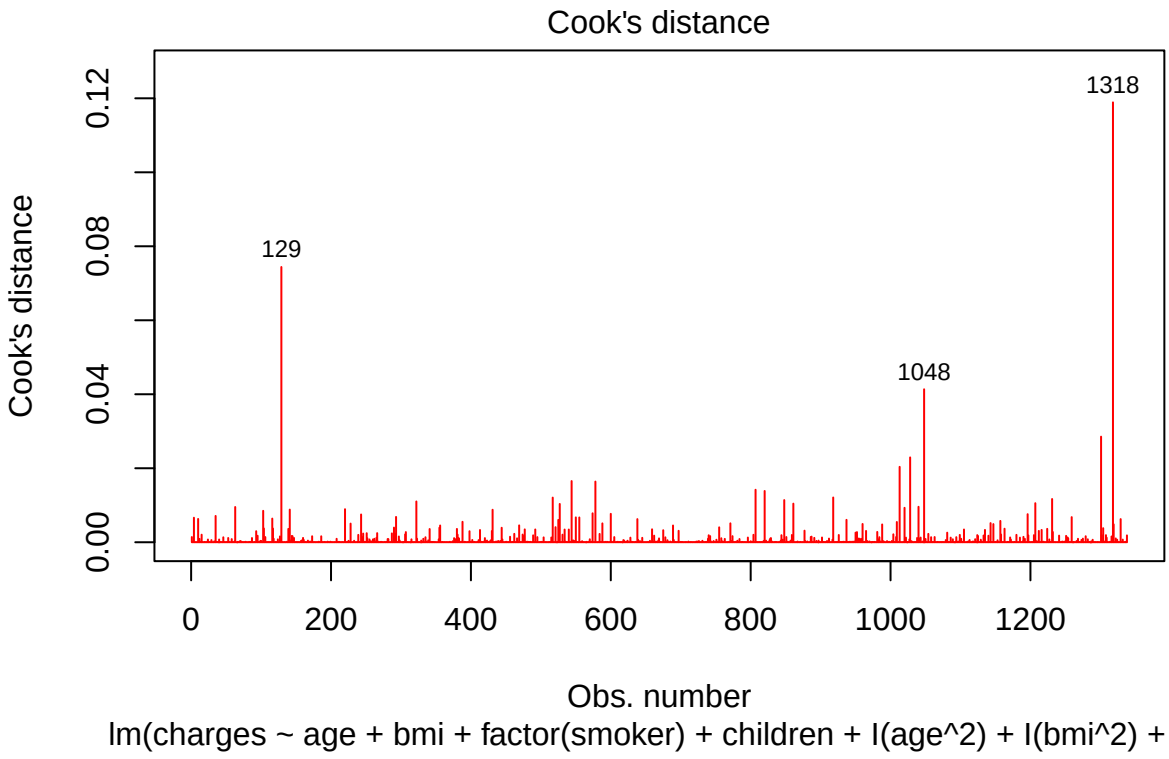
4. Influential points and outliers:

We used residuals and leverage plot to detect outliers and influential points. We observed no points beyond Cook's distance, which means all of the points do not have high Cook's distance scores. In the plot showing the Cook's distance D_i of each observation, there are no points that have a distance greater than 0.5. Therefore, they are not influential. The leverage plot shows multiple influential points that are beyond the 2p/n threshold. However, when we removed these points from the data, the adjusted R square decreased and there are no reasons to delete these points. Therefore, we kept the original data and the predicted model.

```
plot(highermodel, which=5)
```

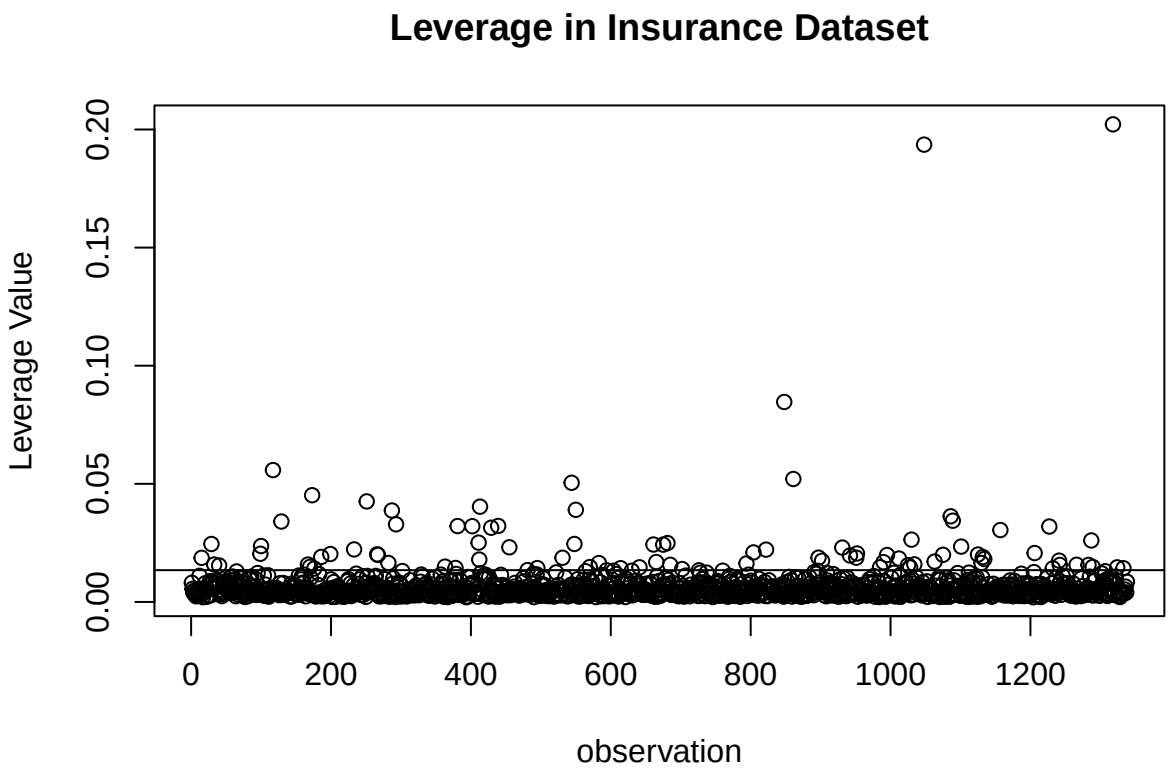



```
plot(highermodel, pch=18, col="red", which=c(4))
```




```
lev=hatvalues(highermodel)
p = length(coef(highermodel))
n = nrow(insurance)
outlier = lev[lev>(2*p/n)]

plot(rownames(insurance),lev, main = "Leverage in Insurance Dataset", xlab="observation",
    ylab = "Leverage Value")
abline(h = 2 *p/n, lty = 1)
```




```
influential <- as.numeric(names(outlier))
data <- insurance[-influential, ]
testmodel <- lm(charges~age+bmi+I(age^2)+I(bmi^2)+I(bmi^3)+children+factor(smoker)+bmi*factor(smoker), d
summary(testmodel)
```

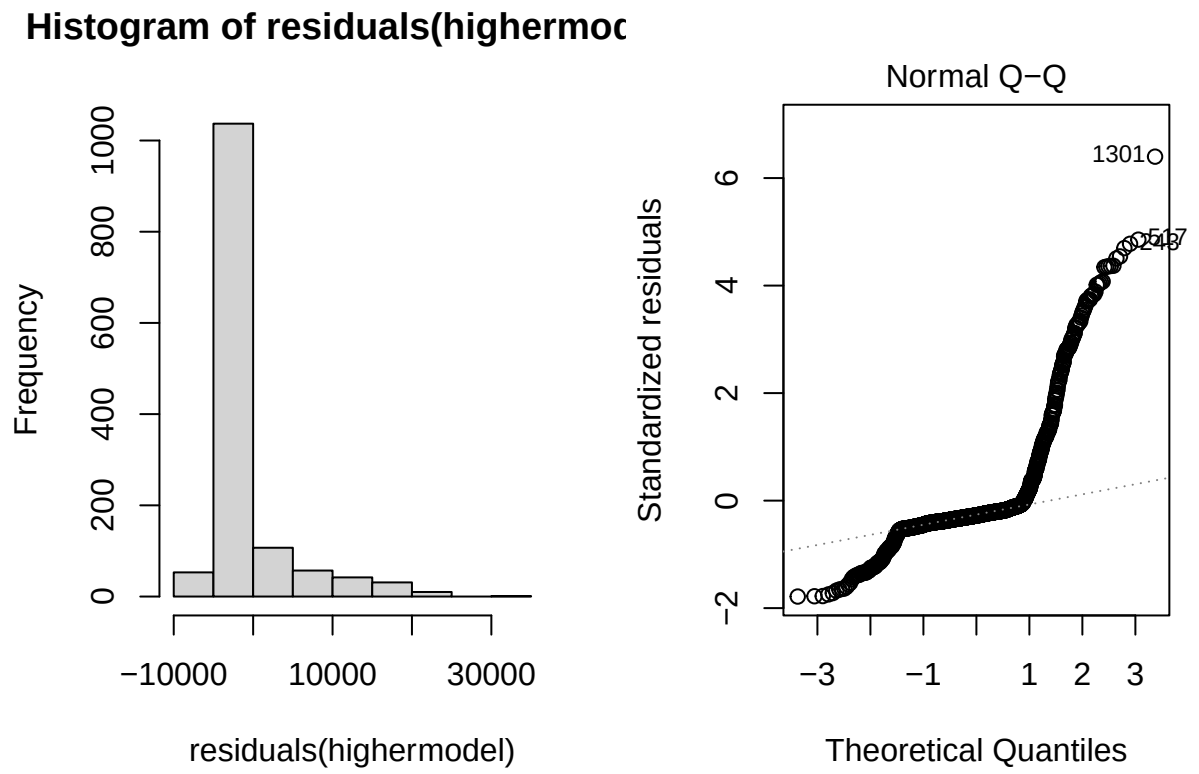
```
##
## Call:
## lm(formula = charges ~ age + bmi + I(age^2) + I(bmi^2) + I(bmi^3) +
##     children + factor(smoker) + bmi * factor(smoker), data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8540  -1741  -1332   -795  30391
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.592e+04  1.872e+04   0.850    0.395
## age                    -3.878e+01  6.675e+01  -0.581    0.561
## bmi                    -1.560e+03  1.866e+03  -0.836    0.403
## I(age^2)                3.831e+00  8.311e-01   4.609 4.46e-06 ***
## I(bmi^2)                5.727e+01  6.070e+01   0.944    0.346
## I(bmi^3)               -6.680e-01  6.453e-01  -1.035    0.301
## children                7.022e+02  1.197e+02   5.866 5.73e-09 ***
## factor(smoker)yes      -3.314e+04  2.383e+03 -13.907  < 2e-16 ***
## bmi:factor(smoker)yes   1.867e+03  7.758e+01  24.060  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4761 on 1240 degrees of freedom
## Multiple R-squared:  0.8299, Adjusted R-squared:  0.8288
## F-statistic: 756.2 on 8 and 1240 DF,  p-value: < 2.2e-16
```

5. Normality assumption:

H_0 : The sample data are significantly normally distributed H_a : The sample data are not significantly normally distributed Another assumption of multiple linear regression is that the errors between observed and predicted values should be normally distributed. Looking at the histogram of residuals and the Q-Q plot, we fail to observe the normal trend of the distribution of the residuals and the points falling close to the diagonal reference line respectively. We used the Shapiro-Wilk test to confirm the normality assumption. On the significance level $\alpha = 0.05$, the result from the test (W = 0.6393, p-value < 2.2e-16), we reject the null hypothesis. Overall, our data does not meet the normality assumption.

```
par(mfrow=c(1,2))
hist(residuals(highermodel))
plot(highermodel, which=2)
```




```
shapiro.test(residuals(highermodel))
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  residuals(highermodel)
## W = 0.63931, p-value < 2.2e-16
```

6. Independence assumption: Our data for both dependent and independent variables are not observed sequentially over a period of time (time-series data). The response Y - charges is not related to time. Therefore, we do not check the model with independence assumption for this dataset.

Transformation for normality

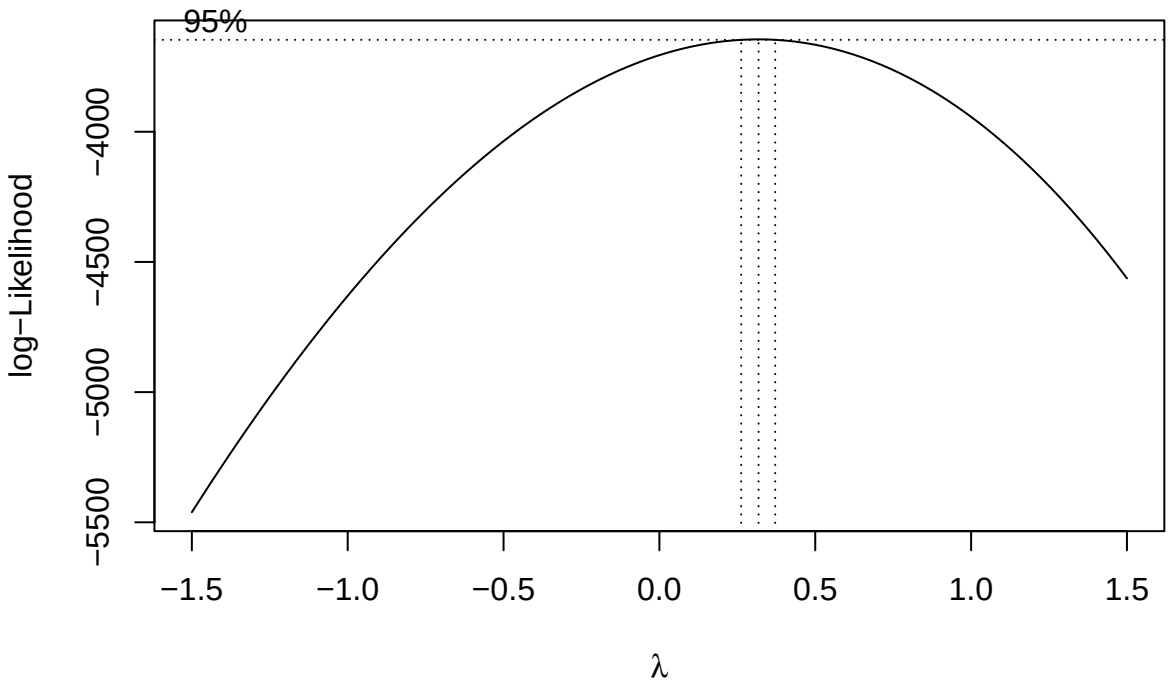
We made a transformation on Y - the response variable for nonnormality of the error terms by using a Box Cox transformation. The method determines the λ in order to transform Y to a replacement response variable Y^(lambda) with the expectation that the regression residuals become normally distributed. The optimal transformation is $\lambda = 0.31818$. However, after the transformation, we then run the Shapiro-Wilk normality test again where the result (W = 0.6393 and p-value < 2.2e-16) indicates that the null hypothesis is rejected. The transformed data still does not distribute normally and the model fails the normality assumption.

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:olsrr':
##
##     cement
```

```
bc=boxcox(highermodel,lambda=seq(-1.5,1.5))
```




```
bestlambda=bc$x[which(bc$y==max(bc$y))]
bestlambda
```

```
## [1] 0.3181818
```

```
bcmodel = lm((((charges^0.31818188)-1)/0.3181818)~age+
               bmi+I(bmi^2)+I(age^2)+I(bmi^3)+
               children+factor(smoker)+bmi*factor(smoker),
             data =insurance)
```

Conclusion

The best fit model for our dataset is:

$Y_{Charges} = \beta_0 + \beta_1 X_{Age} + \beta_2 X_{BMI} + \beta_3 X_{Children} + \beta_4 X_{Smoker} + \beta_5 X_{Age^2} + \beta_6 X_{BMI^2} + \beta_7 X_{BMI^3} + \beta_8 X_{BMI} * X_{Smoker}$

This model includes the main effects that were shown to have a significant impact on insurance charges and the interaction & higher order terms that also significantly affect insurance charges. The main effects included in this model were all determined to be significant through individual t-tests, stepwise selection, and a partial F-test. Both the interaction term and the higher order terms were kept in the model based on the results of their individual t-tests, and insignificant interaction terms and higher order terms were left out of the final model.

Independent Variable effects:

Intercept = 27520.39: When all other variables equal zero, the predicted insurance cost would be $27,520.39.

Age = -26.21: When all other predictor variables are held constant, the insurance cost decreases by $26.21 when the age of the insurance holder increases by one year.

BMI = -2773.974: When all other predictor variables are held constant, insurance cost decreases by $2773.97 when BMI of the insurance holder increases by one 1 kg/m^2.

Smoker (Yes) = -20476.48: When all other predictor variables are held constant, insurance cost decreases by $20,476.48 when the insurance holder is a smoker.

Children = 668.2903: When all other predictor variables are held constant, insurance cost increases by $668.29 with every additional child covered by the health insurance.

BMI & Smoker interaction = 1444.106: When all the other predictors are held constant and the insurance is a smoker, then the insurance cost will increase by $1444.11 when the BMI of the insurance holder increases by 1 kg/m^2.

The R^2 adjusted value for the best fit model obtained is 0.8437, which indicates that 84.37% of the variation in insurance cost is explained by this model. The RMSE is 4788 on 1329 degrees of freedom, which is the lowest RMSE value obtained in all the models that were tested. The minimized RMSE value indicates that this model is the best fit to the data.

Discussion

Our model had both expected and unexpected results. Region did not affect medical insurance costs, which isn't surprising considering that the regional variable is very broad and only considers four regions across the United States. For the region variable to have a noticeable impact on insurance costs, it would be effective to look at costs on a state or county level. Sex was also not a significant predictor, and the presence of influential points in the data made it difficult to draw conclusions with this variable. The results of the children variable were straightforward, the more children covered by the insurance policy, the higher the insurance costs.

The coefficient obtained in the model for the smoker variable was unusual, having a $20,476 decrease in insurance costs when the beneficiary is a smoker. This result was strange because when medical costs are considered with only smoker status as a variable, medical costs were higher when the beneficiary was a smoker. BMI in the model also returned a strange result, decreasing insurance cost by $2774 for every 1 kg/m^2 increase in BMI. Like smoker status, when medical costs are only analyzed with the BMI variable, medical costs increase as BMI increases. Age had the same effect as BMI and smoker status, having a negative effect in the model but a positive correlation when modeled on its own.

An aspect of this model that could be changed is changing the children variable from a quantitative variable to a qualitative variable where children are present or absent in the beneficiary's health insurance. This would likely yield a different result and would likely make predicting insurance costs with children more accurate. Another aspect that could be improved is working with the violation of the normality assumption. Despite transformations, the assumption was not met, and another type of modelling may be more appropriate for this dataset.

References

Choi, M. (2018). Medical Cost Personal Dataset: Insurance Forecast by using Linear Regression. Kaggle. https://www.kaggle.com/mirichoi0218/insurance.

Schoen, C., Osborn, R., Squires, D., & Doty, M. M. (2013). Access, Affordability, And Insurance Complexity Are Often Worse In The United States Compared To Ten Other Countries. Health Affairs, 32(12), 2205-15. https://ezproxy.lib.ucalgary.ca/login?url=https://www.proquest.com/scholarly-journals/access-affordability-insurance-complexity-are/docview/1467749977/se-2?accountid=9838

Riedel, L. M. (2009). Health insurance in the United States. AANA journal, 77(6), 439-444.

Hoffman, C., & Paradise, J. (2008). Health insurance and access to health care in the United States. Annals of the New York Academy of Sciences, 1136(1), 149-160.