Applied Statistics Final Presentation: Insurance Rates

Nidhi Vadnere and Jahnavi Bonagiri

2022-12-02

Section 0: Loading the Dataset

The library functions we need for the insurance data set and load data set

```
library(readxl)
library(car)
```

```
## Loading required package: carData
```

```
library(MASS)
library(lmtest)
```

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

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

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

```
insurance <- read.csv('insurance.csv')
names(insurance)<-c('age','sex','bmi','children','smoker','region','charges')
insurance<-na.omit(insurance)
str(insurance)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr  "female" "male" "male" "male" ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

Section 1: Introduction and Purpose

Health care is one of the most costliest expenses when living in the U.S. costing families hundreds of dollars per month. According to the U.S Bureau of Labor, medical insurance is one the highest spending costs for many middle and low income families. As many families need to budget their daily expenses in an efficient manner, it is important to recognize the factors that may influence a family's rising costs so they can better manage their resources. Economic stress on medical costs can affect long-term monetary dreams of families. The solution to conquer this crisis is to help assist these families in predicting and assisting them with the rising costs. Health insurance is a product that covers fees associated with medicine, surgical procedures or hospital visits of an insured which could be an individual, family, or a collection of people. When people first hear of medical insurance costs, factors such as health history, age, gender, and children first come to mind.

The purpose of this project is to gain a deeper understanding of what factors play the most important role in identifying which families have the higher insurance expenses. By understanding the key factors that affect medical costs, we can predict in advance about the health insurance expenses, which could prove to be very beneficial for insurers and patients to manage their assets appropriately.

Initial assumptions we had going into the analysis was that smoker and bmi would have the highest effect on insurance costs due to medical conditions and medicines they may need.

Section 2: Data Source

The dataset we acquired for the project was provided by a verified data scientist on Kaggle on medical costs. The main data we will be using is the charges as the response variable and bmi, age, children, smoker history and region as the explanatory variables. There was prior work done using different models such as Lasso and Random regression, so we will use this opportunity to perform a thorough analysis to conclude a solid result to see if insurance costs are correlated with bmi, age, children, region and medical history using a Linear regression and various other transformations.

We will be using all of these variables to perform regression analysis of various models:

Numerical Variables:

age: age of primary beneficiary

bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ˆ 2) using the ratio of height to weight, ideally 18.5 to 24.9

charges: Individual medical costs billed by health insurance

Categorical Variables:

sex: insurance contractor gender, female, male 1=Male 0=Female

smoker: Smoker or non-smoker 1= yes 0= no

region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest. southwest=1 southeast=2 northwest=3 northeast=4

children: Number of children covered by health insurance / Number of dependents

Catagorical Variable Tables

```
table(insurance$children)
```

```
##
##   0   1   2   3   4   5
## 574 324 240 157  25  18
```

```
table(insurance$sex)
```

```
##
## female   male
##    662    676
```

```
table(insurance$smoker)
```

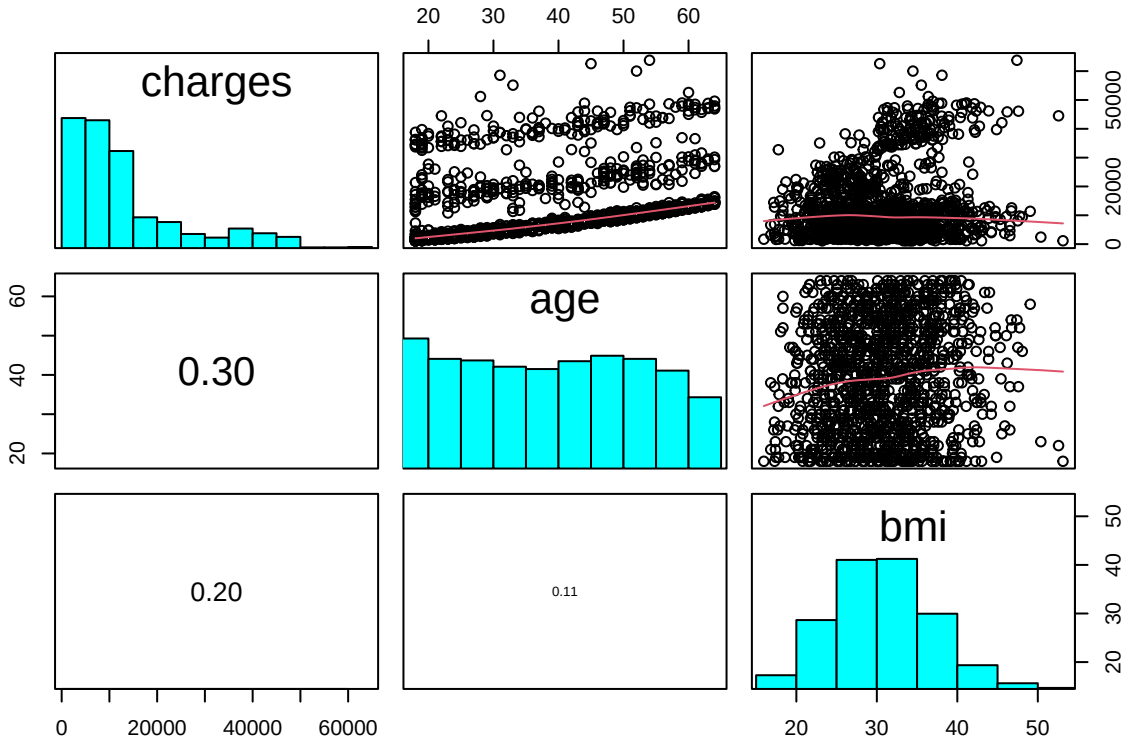
```
##
##   no  yes
## 1064  274
```

```
table(insurance$region)
```

```
##
## northeast northwest southeast southwest
##       324       325       364       325
```

Pairs on Numerical Variables

```
source("pairs.r")
pairs(insurance[c(7, 1,3)],panel=panel.smooth,diag.panel=panel.hist,lower.panel=panel.cor)
```



After running pairs.r, we saw that was an interaction present between charges vs bmi. There was also an interaction between age vs bmi. We also that the effect of age with charges is additive, producing 3 regression lines that are parallel, with only the intercept changing according to the model. We will need to perform more analysis of the regression lines to see if the correlation is stronger with or without the interaction effects.

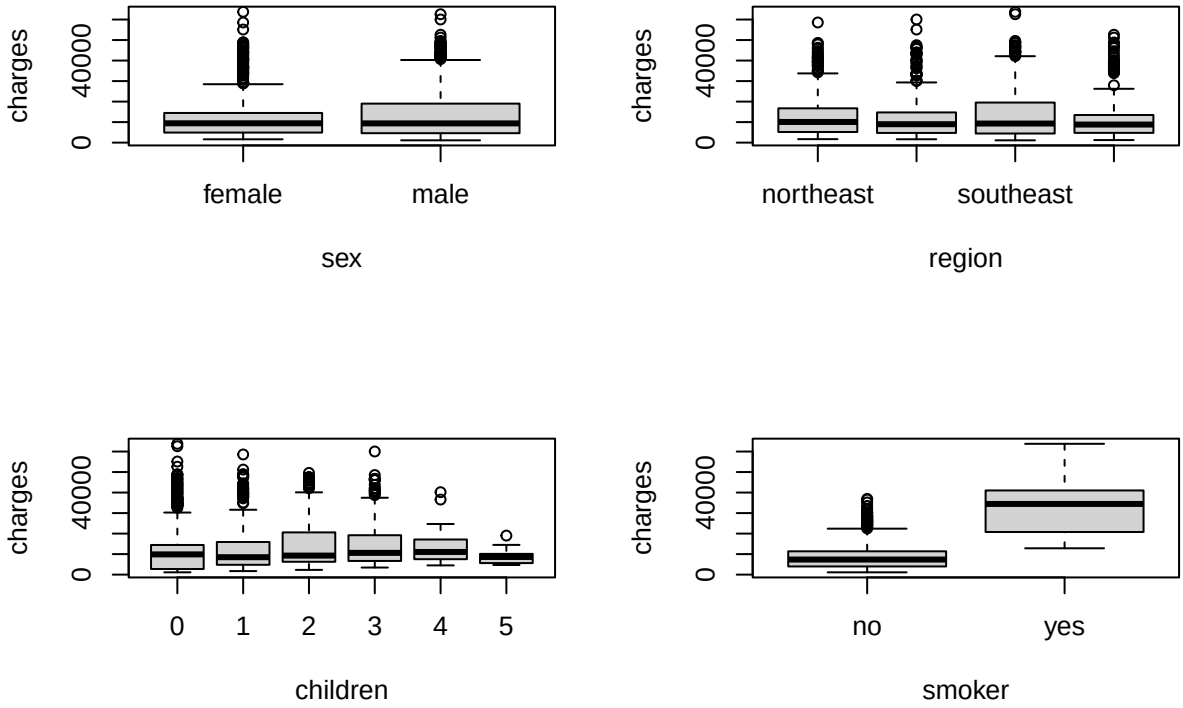
Storing the Categorical Variables

Stored all the categorical variables using as.factor so the code looks cleaner when using various models.

```
insurance$children<-as.factor(insurance$children)
insurance$sex<-as.factor(insurance$sex)
insurance$smoker<-as.factor(insurance$smoker)
insurance$region<-as.factor(insurance$region)
```

Box Plot on Catagorical Variables

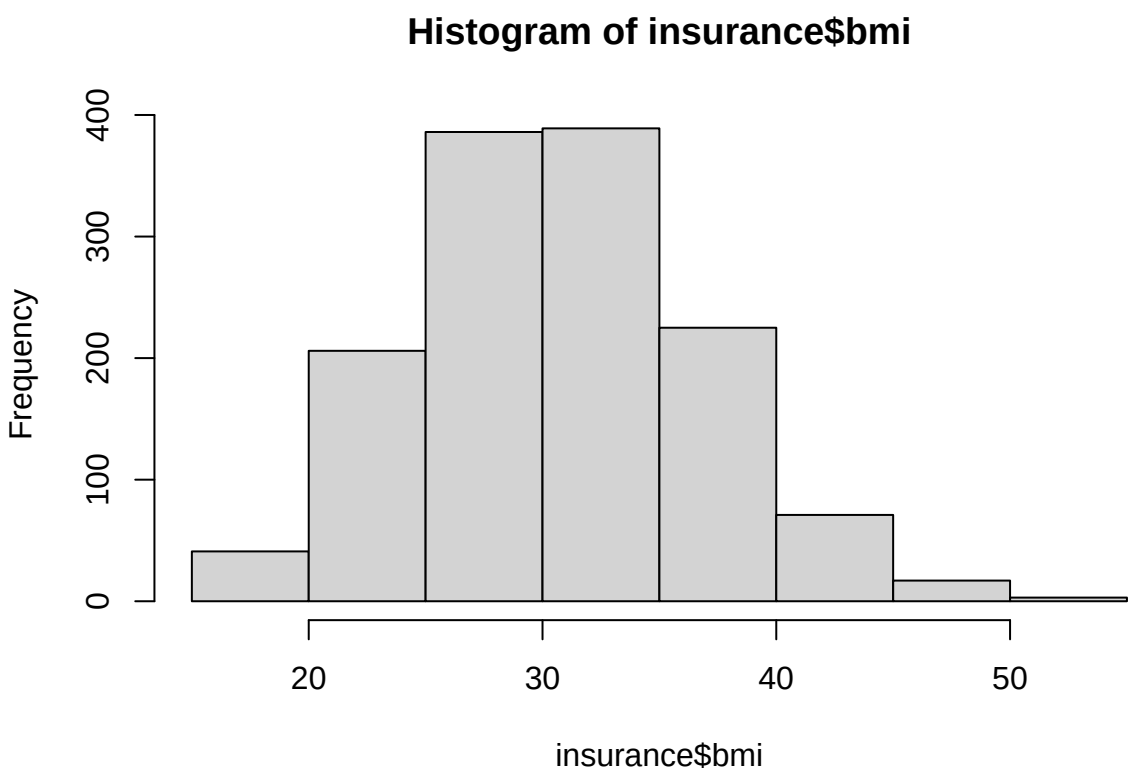
```
par(mfrow=c(2,2))
boxplot(charges~sex,data=insurance)
boxplot(charges~region,data=insurance)
boxplot(charges~children,data=insurance)
boxplot(charges~smoker,data=insurance)
```



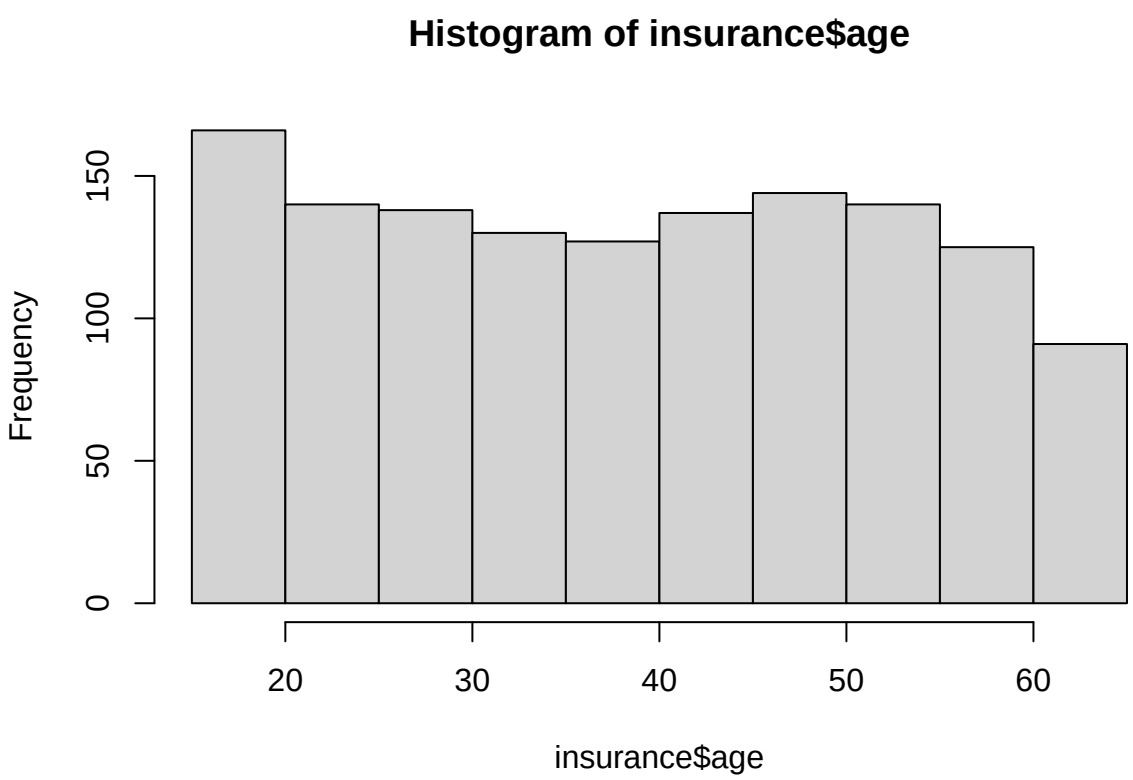
We produced box plots for all the categorical variables to compare the mean and see the differences between all the variables.Based on the results, we noticed that there were a few outliers present for the children, sex and region variables. It was interesting to see how non smoker had a lower median insurance charge rate compared to the smoker median insurance charge rate. We will use this information later when looking for interactions

Histogram Representation of the Numerical Variables

```
hist(insurance$bmi)
```




```
hist(insurance$age)
```

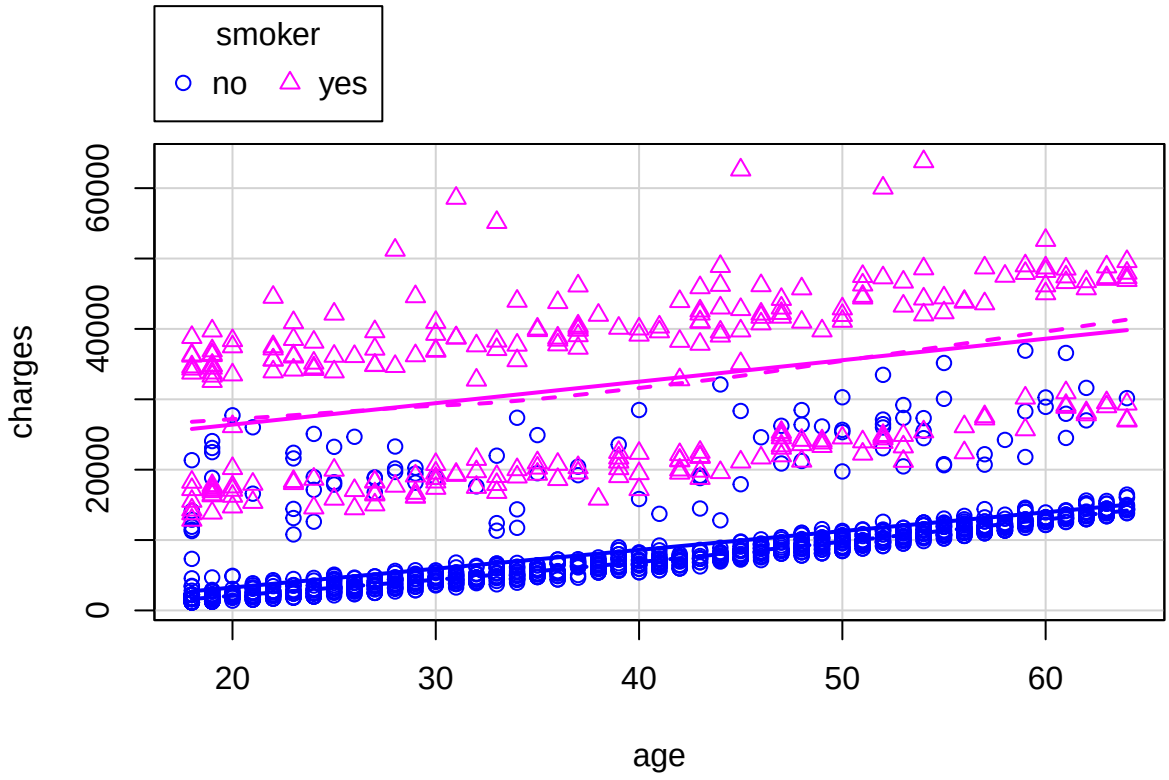


Based on the histogram for the numerical variables, we noticed that the age histogram all had high frequency values for every age. Whereas, the bmi histogram had more of a bell-shaped plot. The values were centered around the ranges of 30-40. As mentioned previously, the normal bmi values range from 18-24, so it made sense that the people with the higher bmi tend to have higher insurance costs.

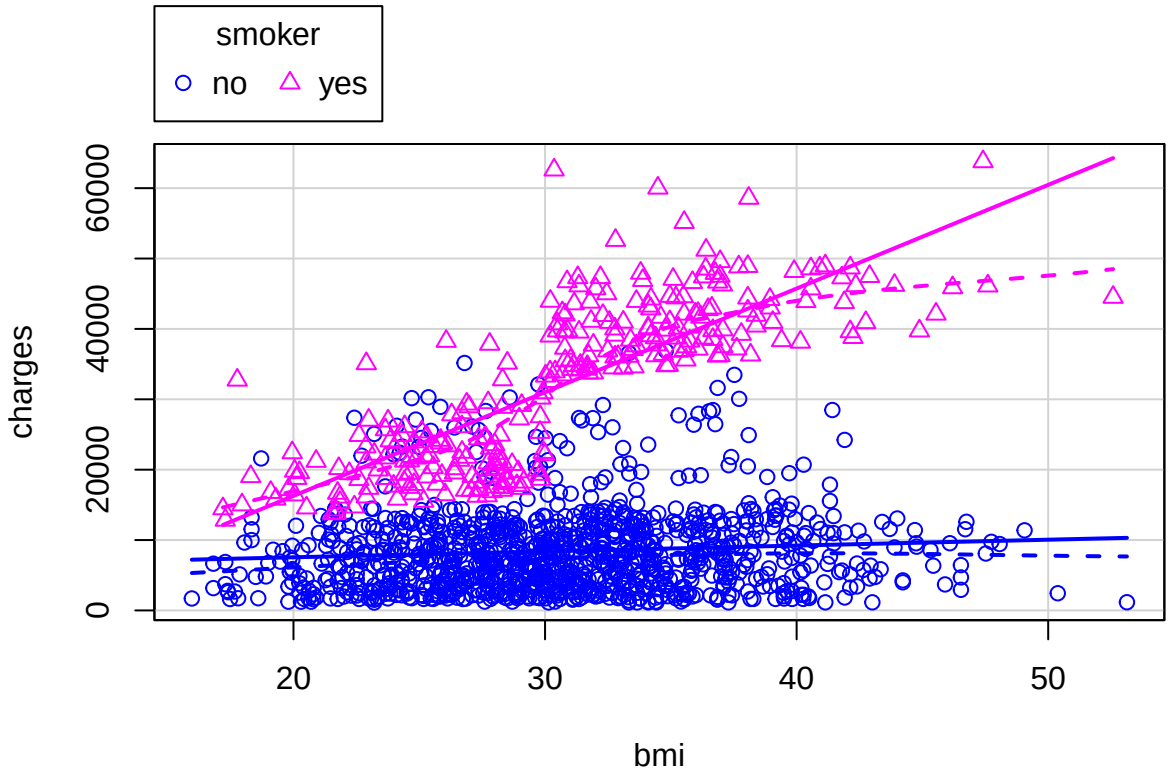
Interactions:

Based on pairs.r and the box plot, we plotted the interactions plots below. As we see in the interaction including smoker with Charges vs Age, there appeared to be a clump of blue dots mixed with pink triangles Additionally, there are two pink lines, which also shows us that we might potentially need another variable to explain this interaction.

```
#Interaction with Smoker for Charges vs Age and Charges vs BMI
library(car)
scatterplot(charges ~ age | smoker, data=insurance)
```




```
scatterplot(charges ~ bmi | smoker, data=insurance)
```



Cross Validation

Train and Test To perform cross validation we used a 80:20 ratio to split our data into a training and testing data set.

```
names(insurance)
```

```
## [1] "age"      "sex"      "bmi"      "children" "smoker"   "region"   "charges"
```

```
n<- length(insurance$charges)
cvindex<- sample(1:n,.8*n,replace=FALSE)
train<-insurance[cvindex,]
test<-insurance[-cvindex,]

View(test)
View(train)
```

Running transformations to find the best full model:

We ran different models with the suspected interactions that would give us the best results. Upon running these models, modE, which had an interaction with smoker and poly of 3, gave us the highest R^2 and adjR^2 out of all the models. Multiple R-squared: 0.8496, Adjusted R-squared: 0.8451

```r
#Full Model with interactions using BMI
modA<-lm(charges~bmi*sex+bmi*children+bmi*smoker+bmi*region+bmi*age,data=train)
#Full Model with interactions using Age
modB<-lm(charges~age*sex+age*children+age*smoker+age*region+age*bmi,data=train)
#Full Model with interactions using BMI with Poly
modC<-lm(charges~bmi*sex+bmi*children+bmi*smoker+bmi*region+bmi*poly(age,3),data=train)
#Full Model with interactions using Age with Poly
modD<-lm(charges~age*sex+age*children+age*smoker+age*region+age*poly(bmi,3),data=train)
#Full Model with interactions using Smoker with Poly on bmi and age
modE<-lm(charges~smoker*sex+smoker*poly(bmi,3)+smoker*children+smoker*region+smoker*poly(age,3),data=tra

summary(modA)
```

```
##
## Call:
## lm(formula = charges ~ bmi * sex + bmi * children + bmi * smoker +
##     bmi * region + bmi * age, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13145.6  -2116.9  -1257.0   -171.7  25242.9
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -3474.338   2788.853  -1.246  0.21312
## bmi                     71.506     91.042   0.785  0.43239
## sexmale              -1585.153   1553.988  -1.020  0.30794
## children1              124.237   1961.315   0.063  0.94950
## children2             5525.725   2084.817   2.650  0.00816 **
## children3              290.964   2684.983   0.108  0.91373
## children4            -2322.788   7417.662  -0.313  0.75423
## children5             4636.742   6132.844   0.756  0.44979
## smokeryes           -21445.570   1907.234 -11.244  < 2e-16 ***
## regionnorthwest       -507.922   2299.139  -0.221  0.82520
## regionsoutheast       4709.568   2194.090   2.146  0.03206 *
## regionsouthwest       1881.483   2308.602   0.815  0.41526
## age                    230.449     54.329   4.242 2.41e-05 ***
## bmi:sexmale             28.240     49.823   0.567  0.57096
## bmi:children1           17.623     62.633   0.281  0.77848
## bmi:children2         -114.332     66.225  -1.726  0.08457 .
## bmi:children3           31.844     85.663   0.372  0.71017
## bmi:children4          194.740    230.609   0.844  0.39861
## bmi:children5          -86.593    208.909  -0.415  0.67859
## bmi:smokeryes         1463.670     61.036  23.981  < 2e-16 ***
## bmi:regionnorthwest    -13.143     77.366  -0.170  0.86514
## bmi:regionsoutheast   -194.179     69.139  -2.809  0.00507 **
## bmi:regionsouthwest   -113.932     75.516  -1.509  0.13168
## bmi:age                 1.147      1.729   0.664  0.50705
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4811 on 1046 degrees of freedom
## Multiple R-squared:  0.8397, Adjusted R-squared:  0.8362
```

```
## F-statistic: 238.2 on 23 and 1046 DF,  p-value: < 2.2e-16
```

```
summary(modB)
```

```
##
## Call:
## lm(formula = charges ~ age * sex + age * children + age * smoker +
##     age * region + age * bmi, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12124  -2936  -1057   1425  25591
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.132e+04  2.983e+03  -3.795 0.000156 ***
## age                  2.556e+02  7.349e+01   3.478 0.000526 ***
## sexmale             -1.540e+01  1.098e+03  -0.014 0.988806
## children1            6.384e+02  1.462e+03   0.437 0.662384
## children2            3.809e+03  1.669e+03   2.282 0.022703 *
## children3            4.888e+02  2.112e+03   0.231 0.817048
## children4            1.919e+03  4.245e+03   0.452 0.651345
## children5            4.571e+03  6.440e+03   0.710 0.478013
## smokeryes            2.235e+04  1.387e+03  16.111  < 2e-16 ***
## regionnorthwest     -7.474e+02  1.559e+03  -0.479 0.631768
## regionsoutheast     -1.599e+03  1.595e+03  -1.002 0.316413
## regionsouthwest     -1.725e+03  1.605e+03  -1.075 0.282518
## bmi                  3.199e+02  9.381e+01   3.410 0.000675 ***
## age:sexmale         -1.067e+01  2.634e+01  -0.405 0.685425
## age:children1        8.013e+00  3.503e+01   0.229 0.819090
## age:children2       -4.492e+01  4.080e+01  -1.101 0.271108
## age:children3        2.145e+01  4.919e+01   0.436 0.662828
## age:children4        4.092e+01  1.061e+02   0.386 0.699902
## age:children5       -8.417e+01  1.674e+02  -0.503 0.615139
## age:smokeryes        2.823e+01  3.371e+01   0.838 0.402432
## age:regionnorthwest  8.419e-01  3.737e+01   0.023 0.982031
## age:regionsoutheast  1.109e+01  3.796e+01   0.292 0.770292
## age:regionsouthwest  1.446e+01  3.853e+01   0.375 0.707588
## age:bmi              6.991e-02  2.274e+00   0.031 0.975478
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5998 on 1046 degrees of freedom
## Multiple R-squared:  0.7508, Adjusted R-squared:  0.7454
## F-statistic:   137 on 23 and 1046 DF,  p-value: < 2.2e-16
```

```
summary(modC)
```

```
##
## Call:
## lm(formula = charges ~ bmi * sex + bmi * children + bmi * smoker +
##     bmi * region + bmi * poly(age, 3), data = train)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14220.8  -1937.7  -1126.2   -407.9  24051.3
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5580.749   1921.050   2.905  0.00375 **
## bmi                    106.405     63.069   1.687  0.09188 .
## sexmale              -1608.332   1547.534  -1.039  0.29891
## children1              139.071   2015.611   0.069  0.94501
## children2             5665.641   2235.790   2.534  0.01142 *
## children3              392.503   2790.448   0.141  0.88817
## children4            -2977.901   7403.183  -0.402  0.68759
## children5             5289.836   6159.727   0.859  0.39066
## smokeryes           -21212.704   1894.794 -11.195  < 2e-16 ***
## regionnorthwest       -653.916   2279.403  -0.287  0.77426
## regionsoutheast       4781.345   2182.855   2.190  0.02872 *
## regionsouthwest       1850.482   2308.132   0.802  0.42290
## poly(age, 3)1       116437.738  25062.869   4.646 3.82e-06 ***
## poly(age, 3)2        17766.067  28443.406   0.625  0.53236
## poly(age, 3)3         3708.847  25210.831   0.147  0.88307
## bmi:sexmale             28.575     49.575   0.576  0.56447
## bmi:children1           34.845     64.623   0.539  0.58986
## bmi:children2          -98.784     71.322  -1.385  0.16634
## bmi:children3           46.654     89.259   0.523  0.60131
## bmi:children4          232.570    230.374   1.010  0.31295
## bmi:children5          -84.971    209.409  -0.406  0.68500
## bmi:smokeryes         1457.366     60.618  24.042  < 2e-16 ***
## bmi:regionnorthwest     -9.831     76.687  -0.128  0.89801
## bmi:regionsoutheast   -197.675     68.776  -2.874  0.00413 **
## bmi:regionsouthwest   -112.612     75.519  -1.491  0.13622
## bmi:poly(age, 3)1      186.552    796.687   0.234  0.81491
## bmi:poly(age, 3)2      247.165    903.007   0.274  0.78436
## bmi:poly(age, 3)3     -126.166    797.056  -0.158  0.87426
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4765 on 1042 degrees of freedom
## Multiple R-squared:  0.8433, Adjusted R-squared:  0.8393
## F-statistic: 207.7 on 27 and 1042 DF,  p-value: < 2.2e-16
```

```
summary(modD)
```

```
##
## Call:
## lm(formula = charges ~ age * sex + age * children + age * smoker +
##     age * region + age * poly(bmi, 3), data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12360  -3210  -1081   1774  25180
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)         -1356.447   1355.139  -1.001 0.317076
## age                   253.871     32.011   7.931 5.59e-15 ***
## sexmale                 6.151   1095.985   0.006 0.995523
## children1             786.457   1459.392   0.539 0.590075
## children2            4078.511   1666.097   2.448 0.014532 *
## children3             708.989   2109.237   0.336 0.736836
## children4            2145.759   4240.446   0.506 0.612948
## children5            3045.990   6442.975   0.473 0.636482
## smokeryes           22398.556   1383.639  16.188  < 2e-16 ***
## regionnorthwest      -865.741   1554.568  -0.557 0.577715
## regionsoutheast     -1791.317   1591.590  -1.125 0.260641
## regionsouthwest     -1915.396   1600.493  -1.197 0.231675
## poly(bmi, 3)1       69761.674  19070.022   3.658 0.000267 ***
## poly(bmi, 3)2        3662.582  16785.561   0.218 0.827317
## poly(bmi, 3)3        4969.805  15616.852   0.318 0.750372
## age:sexmale           -11.187     26.307  -0.425 0.670735
## age:children1           4.515     34.953   0.129 0.897249
## age:children2         -52.295     40.727  -1.284 0.199414
## age:children3          15.195     49.177   0.309 0.757385
## age:children4          31.682    105.983   0.299 0.765049
## age:children5         -31.823    167.697  -0.190 0.849530
## age:smokeryes          26.784     33.627   0.797 0.425908
## age:regionnorthwest     3.161     37.261   0.085 0.932409
## age:regionsoutheast    17.898     37.927   0.472 0.637100
## age:regionsouthwest    18.713     38.421   0.487 0.626324
## age:poly(bmi, 3)1    -185.488    464.843  -0.399 0.689950
## age:poly(bmi, 3)2    -531.202    443.211  -1.199 0.230983
## age:poly(bmi, 3)3    -569.981    450.852  -1.264 0.206430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5972 on 1042 degrees of freedom
## Multiple R-squared:  0.7539, Adjusted R-squared:  0.7475
## F-statistic: 118.2 on 27 and 1042 DF,  p-value: < 2.2e-16
```

```
summary(modE)
```

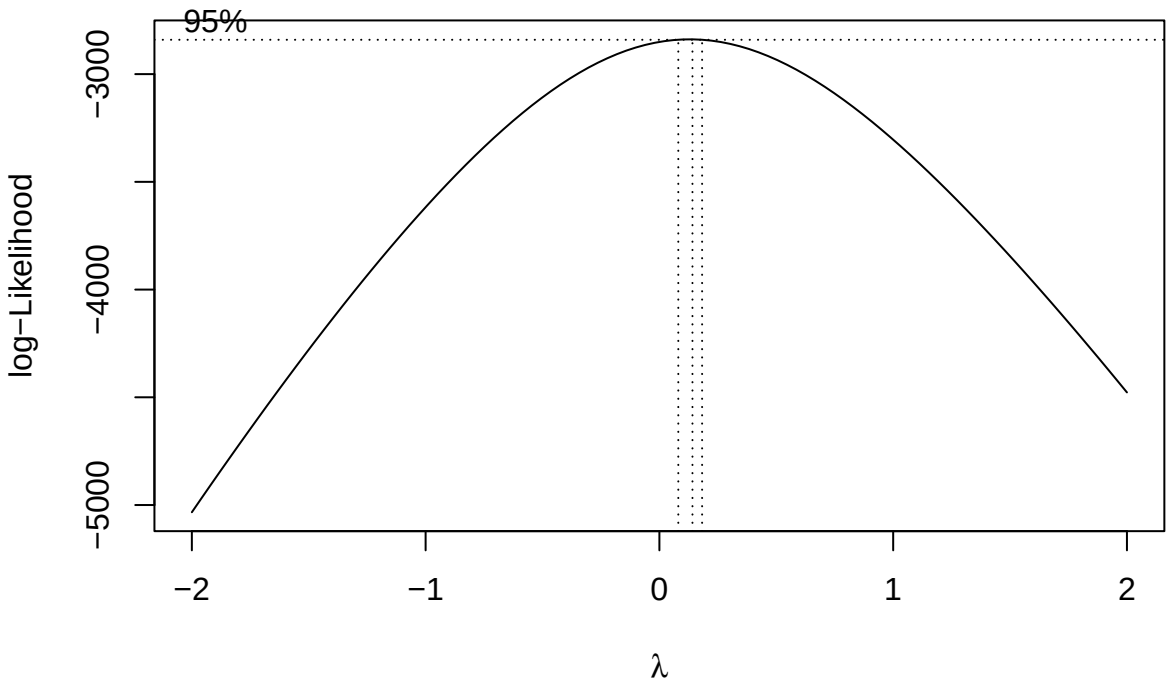
```
##
## Call:
## lm(formula = charges ~ smoker * sex + smoker * poly(bmi, 3) +
##     smoker * children + smoker * region + smoker * poly(age,
##     3), data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8307.0 -1952.6 -1245.6  -431.2 23586.0
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                8588.0      406.0  21.155  < 2e-16 ***
## smokeryes                 24092.4      997.5  24.153  < 2e-16 ***
## sexmale                    -794.4      318.5  -2.494 0.012787 *
## poly(bmi, 3)1              4059.0     5520.0   0.735 0.462305
## poly(bmi, 3)2             -8612.6     5346.4  -1.611 0.107506
```

```
## poly(bmi, 3)3                -983.5     5421.7  -0.181 0.856092
## children1                     1442.5      419.7   3.437 0.000611 ***
## children2                     2620.9      466.3   5.620 2.45e-08 ***
## children3                     1986.3      567.9   3.498 0.000489 ***
## children4                     5147.2     1101.1   4.675 3.33e-06 ***
## children5                     2693.8     1376.7   1.957 0.050651 .
## regionnorthwest               -812.0      451.3  -1.799 0.072259 .
## regionsoutheast              -1101.5      463.9  -2.374 0.017757 *
## regionsouthwest              -1703.8      462.1  -3.687 0.000238 ***
## poly(age, 3)1               125292.0     5227.7  23.967  < 2e-16 ***
## poly(age, 3)2                27744.3     5661.6   4.900 1.11e-06 ***
## poly(age, 3)3                -4047.8     5226.7  -0.774 0.438841
## smokeryes:sexmale              310.7      751.7   0.413 0.679431
## smokeryes:poly(bmi, 3)1     297621.1    12509.8  23.791  < 2e-16 ***
## smokeryes:poly(bmi, 3)2      -8820.9    11915.9  -0.740 0.459310
## smokeryes:poly(bmi, 3)3     -67576.1    11248.7  -6.007 2.60e-09 ***
## smokeryes:children1           -723.5      979.8  -0.738 0.460421
## smokeryes:children2          -1032.4     1050.0  -0.983 0.325736
## smokeryes:children3           -641.5     1160.5  -0.553 0.580540
## smokeryes:children4          -5070.1     2997.8  -1.691 0.091082 .
## smokeryes:children5           -614.9     5173.4  -0.119 0.905415
## smokeryes:regionnorthwest     -286.7     1093.3  -0.262 0.793226
## smokeryes:regionsoutheast    -1300.2     1025.4  -1.268 0.205067
## smokeryes:regionsouthwest      990.9     1113.5   0.890 0.373703
## smokeryes:poly(age, 3)1     -15593.1    12221.1  -1.276 0.202270
## smokeryes:poly(age, 3)2     -10031.7    12199.7  -0.822 0.411104
## smokeryes:poly(age, 3)3      12903.5    11680.8   1.105 0.269555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4642 on 1038 degrees of freedom
## Multiple R-squared:  0.8519, Adjusted R-squared:  0.8475
## F-statistic: 192.6 on 31 and 1038 DF,  p-value: < 2.2e-16
```

Box-cox:

We used the method of Box-Cox to see if our data set requires transformations to get the best regression. The Box-Cox transformation suggests that using lambda = .2 or transforming Y by Yˆ.2. We will be using this lambda value in our later models to see if there is a significant difference between the models. We ended up getting a Multiple R-squared: 0.8295 and Adjusted R-squared: 0.8244, which was lower than modE.

```
boxcox(charges~.,data=train)
```




```
modJ<- lm(charges^0.2~smoker*sex+smoker*poly(bmi,3)+smoker*children+smoker*region+smoker*poly(age,3),da
summary(modJ)
```

```
##
## Call:
## lm(formula = charges^0.2 ~ smoker * sex + smoker * poly(bmi,
##     3) + smoker * children + smoker * region + smoker * poly(age,
##     3), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64737 -0.20796 -0.12792 -0.00432  2.85136
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                5.88299    0.04085 144.002  < 2e-16 ***
## smokeryes                  2.01925    0.10038  20.116  < 2e-16 ***
## sexmale                   -0.13178    0.03205  -4.111 4.25e-05 ***
## poly(bmi, 3)1              0.44961    0.55550   0.809  0.41849
## poly(bmi, 3)2             -1.02439    0.53803  -1.904  0.05719 .
## poly(bmi, 3)3             -0.11754    0.54560  -0.215  0.82948
## children1                  0.21499    0.04223   5.091 4.23e-07 ***
## children2                  0.38494    0.04693   8.203 6.89e-16 ***
## children3                  0.34906    0.05715   6.108 1.43e-09 ***
## children4                  0.70557    0.11081   6.368 2.88e-10 ***
```

```
## children5                   0.54541    0.13854   3.937 8.81e-05 ***
## regionnorthwest            -0.12130    0.04542  -2.671  0.00768 **
## regionsoutheast            -0.18754    0.04668  -4.017 6.32e-05 ***
## regionsouthwest            -0.24936    0.04650  -5.363 1.01e-07 ***
## poly(age, 3)1              21.82825    0.52608  41.493  < 2e-16 ***
## poly(age, 3)2              -0.19923    0.56974  -0.350  0.72664
## poly(age, 3)3               0.22518    0.52598   0.428  0.66865
## smokeryes:sexmale           0.11588    0.07565   1.532  0.12589
## smokeryes:poly(bmi, 3)1    15.36021    1.25890  12.201  < 2e-16 ***
## smokeryes:poly(bmi, 3)2    -0.77229    1.19914  -0.644  0.51969
## smokeryes:poly(bmi, 3)3    -3.00535    1.13199  -2.655  0.00805 **
## smokeryes:children1        -0.19196    0.09860  -1.947  0.05183 .
## smokeryes:children2        -0.29477    0.10567  -2.790  0.00537 **
## smokeryes:children3        -0.27875    0.11679  -2.387  0.01717 *
## smokeryes:children4        -0.71967    0.30167  -2.386  0.01723 *
## smokeryes:children5        -0.34818    0.52062  -0.669  0.50378
## smokeryes:regionnorthwest   0.07655    0.11002   0.696  0.48672
## smokeryes:regionsoutheast   0.07717    0.10319   0.748  0.45471
## smokeryes:regionsouthwest   0.21459    0.11205   1.915  0.05576 .
## smokeryes:poly(age, 3)1   -15.94112    1.22985 -12.962  < 2e-16 ***
## smokeryes:poly(age, 3)2     0.73346    1.22770   0.597  0.55036
## smokeryes:poly(age, 3)3     0.10140    1.17547   0.086  0.93127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4671 on 1038 degrees of freedom
## Multiple R-squared:  0.8388, Adjusted R-squared:  0.834
## F-statistic: 174.3 on 31 and 1038 DF,  p-value: < 2.2e-16
```

Creating the Full Model and Finding FullMSE

```
full.mod<-lm( charges~smoker*sex+smoker*poly(bmi,3)+smoker*children+smoker*region+smoker*poly(age,3),da
fullMSE<-summary(full.mod)$sig^2
```

Using our training data, we created a full model that contains all the interactions from modE. We will use this model and preform backwards, forwards, and both-direction selection methods

Backward

```
backstep <- step(full.mod, direction="backward")
```

```
## Start:  AIC=18099.29
## charges ~ smoker * sex + smoker * poly(bmi, 3) + smoker * children +
##     smoker * region + smoker * poly(age, 3)
##
##                       Df  Sum of Sq        RSS   AIC
## - smoker:children      5 7.7152e+07 2.2444e+10 18093
## - smoker:poly(age, 3)  3 7.1553e+07 2.2438e+10 18097
## - smoker:sex           1 3.6817e+06 2.2370e+10 18098
```

```
## - smoker:region        3 1.0750e+08 2.2474e+10 18098
## <none>                              2.2366e+10 18099
## - smoker:poly(bmi, 3)  3 1.2746e+10 3.5112e+10 18576
##
## Step:  AIC=18092.98
## charges ~ smoker + sex + poly(bmi, 3) + children + region + poly(age,
##     3) + smoker:sex + smoker:poly(bmi, 3) + smoker:region + smoker:poly(age,
##     3)
##
##                       Df  Sum of Sq        RSS   AIC
## - smoker:poly(age, 3)  3 5.9081e+07 2.2503e+10 18090
## - smoker:sex           1 1.4179e+06 2.2445e+10 18091
## - smoker:region        3 9.4903e+07 2.2538e+10 18092
## <none>                              2.2444e+10 18093
## - children             5 1.1016e+09 2.3545e+10 18134
## - smoker:poly(bmi, 3)  3 1.2888e+10 3.5332e+10 18573
##
## Step:  AIC=18089.79
## charges ~ smoker + sex + poly(bmi, 3) + children + region + poly(age,
##     3) + smoker:sex + smoker:poly(bmi, 3) + smoker:region
##
##                       Df  Sum of Sq        RSS   AIC
## - smoker:sex           1 1.5633e+06 2.2504e+10 18088
## - smoker:region        3 1.1435e+08 2.2617e+10 18089
## <none>                              2.2503e+10 18090
## - children             5 1.0908e+09 2.3593e+10 18130
## - smoker:poly(bmi, 3)  3 1.2848e+10 3.5350e+10 18567
## - poly(age, 3)         3 1.5189e+10 3.7691e+10 18636
##
## Step:  AIC=18087.86
## charges ~ smoker + sex + poly(bmi, 3) + children + region + poly(age,
##     3) + smoker:poly(bmi, 3) + smoker:region
##
##                       Df  Sum of Sq        RSS   AIC
## - smoker:region        3 1.1522e+08 2.2619e+10 18087
## <none>                              2.2504e+10 18088
## - sex                  1 1.5139e+08 2.2656e+10 18093
## - children             5 1.0938e+09 2.3598e+10 18129
## - smoker:poly(bmi, 3)  3 1.3032e+10 3.5536e+10 18571
## - poly(age, 3)         3 1.5192e+10 3.7696e+10 18634
##
## Step:  AIC=18087.33
## charges ~ smoker + sex + poly(bmi, 3) + children + region + poly(age,
##     3) + smoker:poly(bmi, 3)
##
##                       Df  Sum of Sq        RSS   AIC
## <none>                              2.2619e+10 18087
## - sex                  1 1.4833e+08 2.2768e+10 18092
## - region               3 3.5645e+08 2.2976e+10 18098
## - children             5 1.1211e+09 2.3741e+10 18129
## - smoker:poly(bmi, 3)  3 1.4093e+10 3.6713e+10 18600
## - poly(age, 3)         3 1.5081e+10 3.7700e+10 18628
```

Forward

```
forwardstep <- step(full.mod, direction="forward")
```

```
## Start:  AIC=18099.29
## charges ~ smoker * sex + smoker * poly(bmi, 3) + smoker * children +
##     smoker * region + smoker * poly(age, 3)
```

Both Directions

```
bothstep <- step(full.mod, direction="both")
```

```
## Start:  AIC=18099.29
## charges ~ smoker * sex + smoker * poly(bmi, 3) + smoker * children +
##     smoker * region + smoker * poly(age, 3)
##
##                       Df  Sum of Sq        RSS   AIC
## - smoker:children      5 7.7152e+07 2.2444e+10 18093
## - smoker:poly(age, 3)  3 7.1553e+07 2.2438e+10 18097
## - smoker:sex           1 3.6817e+06 2.2370e+10 18098
## - smoker:region        3 1.0750e+08 2.2474e+10 18098
## <none>                              2.2366e+10 18099
## - smoker:poly(bmi, 3)  3 1.2746e+10 3.5112e+10 18576
##
## Step:  AIC=18092.98
## charges ~ smoker + sex + poly(bmi, 3) + children + region + poly(age,
##     3) + smoker:sex + smoker:poly(bmi, 3) + smoker:region + smoker:poly(age,
##     3)
##
##                       Df  Sum of Sq        RSS   AIC
## - smoker:poly(age, 3)  3 5.9081e+07 2.2503e+10 18090
## - smoker:sex           1 1.4179e+06 2.2445e+10 18091
## - smoker:region        3 9.4903e+07 2.2538e+10 18092
## <none>                              2.2444e+10 18093
## + smoker:children      5 7.7152e+07 2.2366e+10 18099
## - children             5 1.1016e+09 2.3545e+10 18134
## - smoker:poly(bmi, 3)  3 1.2888e+10 3.5332e+10 18573
##
## Step:  AIC=18089.79
## charges ~ smoker + sex + poly(bmi, 3) + children + region + poly(age,
##     3) + smoker:sex + smoker:poly(bmi, 3) + smoker:region
##
##                       Df  Sum of Sq        RSS   AIC
## - smoker:sex           1 1.5633e+06 2.2504e+10 18088
## - smoker:region        3 1.1435e+08 2.2617e+10 18089
## <none>                              2.2503e+10 18090
## + smoker:poly(age, 3)  3 5.9081e+07 2.2444e+10 18093
## + smoker:children      5 6.4680e+07 2.2438e+10 18097
## - children             5 1.0908e+09 2.3593e+10 18130
## - smoker:poly(bmi, 3)  3 1.2848e+10 3.5350e+10 18567
```

```
## - poly(age, 3)         3 1.5189e+10 3.7691e+10 18636
##
## Step:  AIC=18087.86
## charges ~ smoker + sex + poly(bmi, 3) + children + region + poly(age,
##     3) + smoker:poly(bmi, 3) + smoker:region
##
##                       Df  Sum of Sq        RSS   AIC
## - smoker:region        3 1.1522e+08 2.2619e+10 18087
## <none>                              2.2504e+10 18088
## + smoker:sex           1 1.5633e+06 2.2503e+10 18090
## + smoker:poly(age, 3)  3 5.9226e+07 2.2445e+10 18091
## - sex                  1 1.5139e+08 2.2656e+10 18093
## + smoker:children      5 6.2768e+07 2.2441e+10 18095
## - children             5 1.0938e+09 2.3598e+10 18129
## - smoker:poly(bmi, 3)  3 1.3032e+10 3.5536e+10 18571
## - poly(age, 3)         3 1.5192e+10 3.7696e+10 18634
##
## Step:  AIC=18087.33
## charges ~ smoker + sex + poly(bmi, 3) + children + region + poly(age,
##     3) + smoker:poly(bmi, 3)
##
##                       Df  Sum of Sq        RSS   AIC
## <none>                              2.2619e+10 18087
## + smoker:region        3 1.1522e+08 2.2504e+10 18088
## + smoker:sex           1 2.4243e+06 2.2617e+10 18089
## + smoker:poly(age, 3)  3 7.9024e+07 2.2540e+10 18090
## - sex                  1 1.4833e+08 2.2768e+10 18092
## + smoker:children      5 4.8240e+07 2.2571e+10 18095
## - region               3 3.5645e+08 2.2976e+10 18098
## - children             5 1.1211e+09 2.3741e+10 18129
## - smoker:poly(bmi, 3)  3 1.4093e+10 3.6713e+10 18600
## - poly(age, 3)         3 1.5081e+10 3.7700e+10 18628
```

Model Comparisons

```
##Function to Calculate PMSE
PMSE<-function(model,testdata){
  fitted <- predict(model,newdata=testdata)
  ytest<- testdata[,7]
  PMSEret<-sum((ytest-fitted)^2)/268
  return(PMSEret)
}

source('C:/Users/Jahnavi Bonagiri/Downloads/modelselfunctionss.R')
rbind(
  full.model = c(Criteria(full.mod,fullMSE,label=T),
         PMSE(full.mod,test)),

  backstep.model = c(Criteria(backstep,fullMSE,label=T),
         PMSE(backstep,test)),

  forwardstep.model = c(Criteria(forwardstep,fullMSE,label=T),
```

```
        PMSE(forwardstep,test)),

  bothstep.model = c(Criteria(bothstep,fullMSE,label=T),
        PMSE(bothstep,test))
  )
```

```
##                   p+1  R2adj    Cp      AIC       PRESS
## full.model         32 0.8475 32.00 18099.29         Inf 24612578
## backstep.model     20 0.8475 19.74 18087.33 24795431258 24186308
## forwardstep.model  32 0.8475 32.00 18099.29         Inf 24612578
## bothstep.model     20 0.8475 19.74 18087.33 24795431258 24186308
```

We noticed that our full model and forward model were similar. Our backward selection and both direction selection are similar as well. So we will compare the full model and the reduced model to find the best results.

Our backstep/both step model has the lowest PMSE. To further test our two models, we will preform a two-way anova test to decide on our final model.

Comparing Best Two Models

```
summary(full.mod)
```

```
##
## Call:
## lm(formula = charges ~ smoker * sex + smoker * poly(bmi, 3) +
##     smoker * children + smoker * region + smoker * poly(age,
##     3), data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8307.0 -1952.6 -1245.6  -431.2 23586.0
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                8588.0      406.0  21.155  < 2e-16 ***
## smokeryes                 24092.4      997.5  24.153  < 2e-16 ***
## sexmale                    -794.4      318.5  -2.494 0.012787 *
## poly(bmi, 3)1              4059.0     5520.0   0.735 0.462305
## poly(bmi, 3)2             -8612.6     5346.4  -1.611 0.107506
## poly(bmi, 3)3              -983.5     5421.7  -0.181 0.856092
## children1                  1442.5      419.7   3.437 0.000611 ***
## children2                  2620.9      466.3   5.620 2.45e-08 ***
## children3                  1986.3      567.9   3.498 0.000489 ***
## children4                  5147.2     1101.1   4.675 3.33e-06 ***
## children5                  2693.8     1376.7   1.957 0.050651 .
## regionnorthwest            -812.0      451.3  -1.799 0.072259 .
## regionsoutheast           -1101.5      463.9  -2.374 0.017757 *
## regionsouthwest           -1703.8      462.1  -3.687 0.000238 ***
## poly(age, 3)1            125292.0     5227.7  23.967  < 2e-16 ***
## poly(age, 3)2             27744.3     5661.6   4.900 1.11e-06 ***
```

```
## poly(age, 3)3              -4047.8     5226.7  -0.774 0.438841
## smokeryes:sexmale            310.7      751.7   0.413 0.679431
## smokeryes:poly(bmi, 3)1   297621.1    12509.8  23.791  < 2e-16 ***
## smokeryes:poly(bmi, 3)2    -8820.9    11915.9  -0.740 0.459310
## smokeryes:poly(bmi, 3)3   -67576.1    11248.7  -6.007 2.60e-09 ***
## smokeryes:children1          -723.5      979.8  -0.738 0.460421
## smokeryes:children2         -1032.4     1050.0  -0.983 0.325736
## smokeryes:children3          -641.5     1160.5  -0.553 0.580540
## smokeryes:children4         -5070.1     2997.8  -1.691 0.091082 .
## smokeryes:children5          -614.9     5173.4  -0.119 0.905415
## smokeryes:regionnorthwest    -286.7     1093.3  -0.262 0.793226
## smokeryes:regionsoutheast   -1300.2     1025.4  -1.268 0.205067
## smokeryes:regionsouthwest     990.9     1113.5   0.890 0.373703
## smokeryes:poly(age, 3)1   -15593.1    12221.1  -1.276 0.202270
## smokeryes:poly(age, 3)2   -10031.7    12199.7  -0.822 0.411104
## smokeryes:poly(age, 3)3    12903.5    11680.8   1.105 0.269555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4642 on 1038 degrees of freedom
## Multiple R-squared:  0.8519, Adjusted R-squared:  0.8475
## F-statistic: 192.6 on 31 and 1038 DF,  p-value: < 2.2e-16
```

```
summary(backstep)
```

```
##
## Call:
## lm(formula = charges ~ smoker + sex + poly(bmi, 3) + children +
##     region + poly(age, 3) + smoker:poly(bmi, 3), data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7688.6 -1910.6 -1260.6  -635.3 23772.0
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 8740.0      372.9  23.437  < 2e-16 ***
## smokeryes                  23552.1      359.4  65.531  < 2e-16 ***
## sexmale                     -753.9      287.3  -2.624 0.008816 **
## poly(bmi, 3)1               5754.4     5454.9   1.055 0.291707
## poly(bmi, 3)2              -8644.6     5337.0  -1.620 0.105588
## poly(bmi, 3)3               -989.2     5418.6  -0.183 0.855177
## children1                   1284.9      377.5   3.404 0.000690 ***
## children2                   2425.0      416.2   5.826 7.53e-09 ***
## children3                   1821.4      493.4   3.692 0.000234 ***
## children4                   4535.0     1021.2   4.441 9.90e-06 ***
## children5                   2623.1     1323.6   1.982 0.047758 *
## regionnorthwest             -914.7      407.6  -2.244 0.025050 *
## regionsoutheast            -1402.6      412.3  -3.402 0.000694 ***
## regionsouthwest            -1521.1      415.8  -3.658 0.000267 ***
## poly(age, 3)1             121612.4     4698.5  25.883  < 2e-16 ***
## poly(age, 3)2              25508.7     4999.6   5.102 3.98e-07 ***
## poly(age, 3)3              -1752.2     4655.5  -0.376 0.706711
## smokeryes:poly(bmi, 3)1 290715.1    11568.9  25.129  < 2e-16 ***
```

```
## smokeryes:poly(bmi, 3)2 -11484.6    11218.2  -1.024 0.306191
## smokeryes:poly(bmi, 3)3 -67064.5    10813.4  -6.202 8.01e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4641 on 1050 degrees of freedom
## Multiple R-squared:  0.8502, Adjusted R-squared:  0.8475
## F-statistic: 313.7 on 19 and 1050 DF,  p-value: < 2.2e-16
```

```
anova(backstep,full.mod)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ smoker + sex + poly(bmi, 3) + children + region + poly(age,
##     3) + smoker:poly(bmi, 3)
## Model 2: charges ~ smoker * sex + smoker * poly(bmi, 3) + smoker * children +
##     smoker * region + smoker * poly(age, 3)
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1   1050 2.2619e+10
## 2   1038 2.2366e+10 12 253011382 0.9785 0.4674
```

After preforming the anova test, we decided to choose the backstep model over our full model that had a lower PMSE.

Final Model Summary

```
summary(backstep)
```

```
##
## Call:
## lm(formula = charges ~ smoker + sex + poly(bmi, 3) + children +
##     region + poly(age, 3) + smoker:poly(bmi, 3), data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7688.6 -1910.6 -1260.6  -635.3 23772.0
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               8740.0      372.9  23.437  < 2e-16 ***
## smokeryes                23552.1      359.4  65.531  < 2e-16 ***
## sexmale                   -753.9      287.3  -2.624 0.008816 **
## poly(bmi, 3)1             5754.4     5454.9   1.055 0.291707
## poly(bmi, 3)2            -8644.6     5337.0  -1.620 0.105588
## poly(bmi, 3)3             -989.2     5418.6  -0.183 0.855177
## children1                 1284.9      377.5   3.404 0.000690 ***
## children2                 2425.0      416.2   5.826 7.53e-09 ***
## children3                 1821.4      493.4   3.692 0.000234 ***
## children4                 4535.0     1021.2   4.441 9.90e-06 ***
## children5                 2623.1     1323.6   1.982 0.047758 *
## regionnorthwest           -914.7      407.6  -2.244 0.025050 *
```

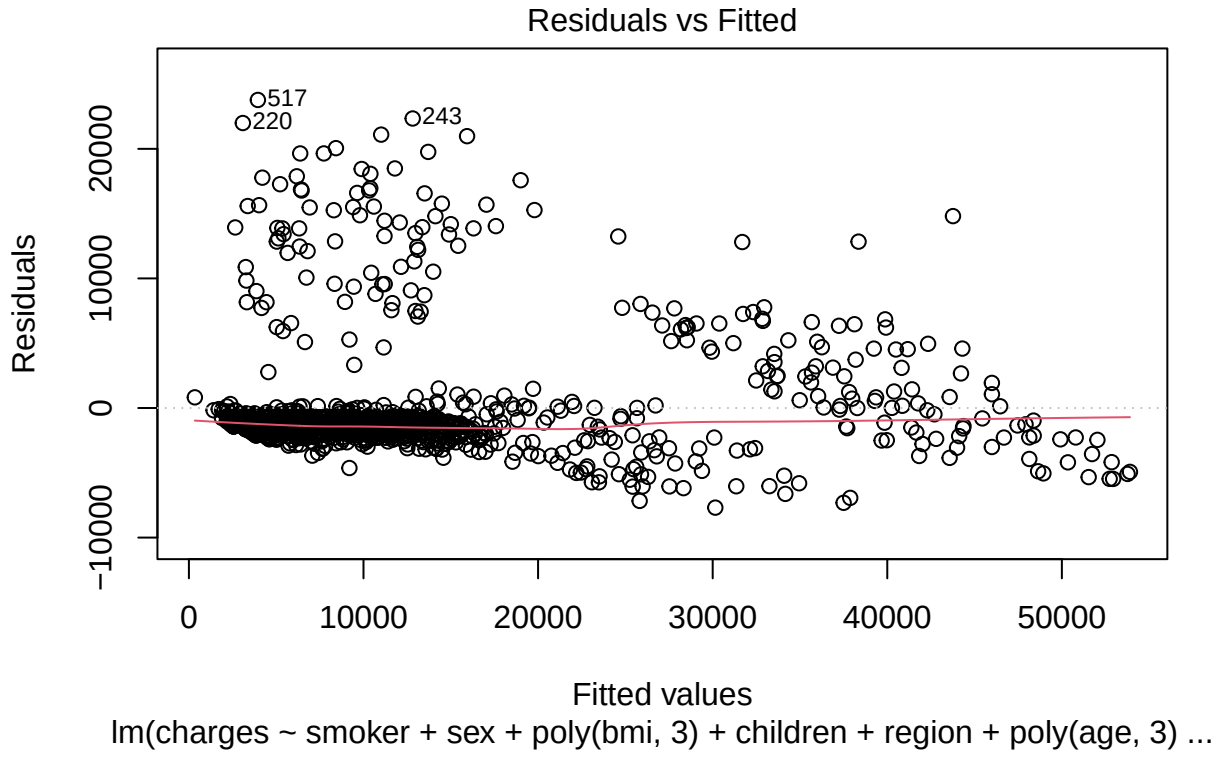
```
## regionsoutheast          -1402.6      412.3  -3.402 0.000694 ***
## regionsouthwest          -1521.1      415.8  -3.658 0.000267 ***
## poly(age, 3)1           121612.4     4698.5  25.883  < 2e-16 ***
## poly(age, 3)2            25508.7     4999.6   5.102 3.98e-07 ***
## poly(age, 3)3            -1752.2     4655.5  -0.376 0.706711
## smokeryes:poly(bmi, 3)1 290715.1    11568.9  25.129  < 2e-16 ***
## smokeryes:poly(bmi, 3)2 -11484.6    11218.2  -1.024 0.306191
## smokeryes:poly(bmi, 3)3 -67064.5    10813.4  -6.202 8.01e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4641 on 1050 degrees of freedom
## Multiple R-squared:  0.8502, Adjusted R-squared:  0.8475
## F-statistic: 313.7 on 19 and 1050 DF,  p-value: < 2.2e-16
```

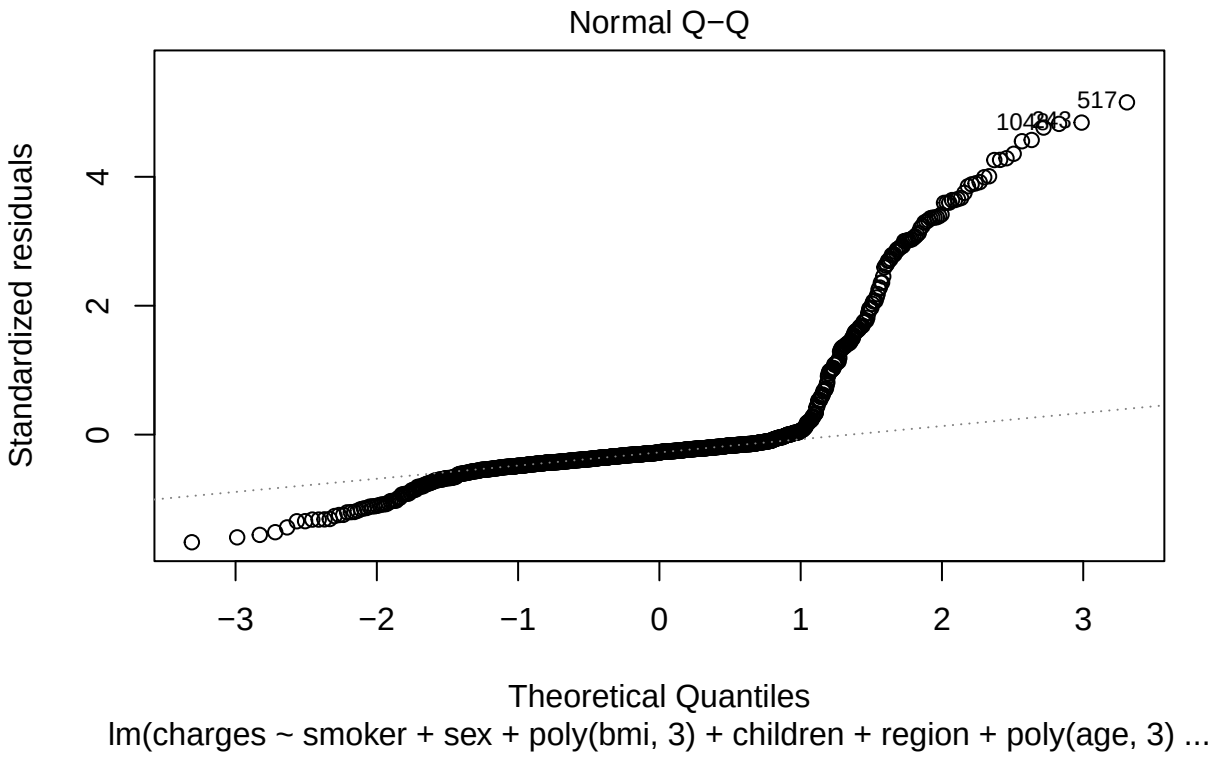
We now have a model where most terms are significant and our adjusted R-squared value is fairly high.

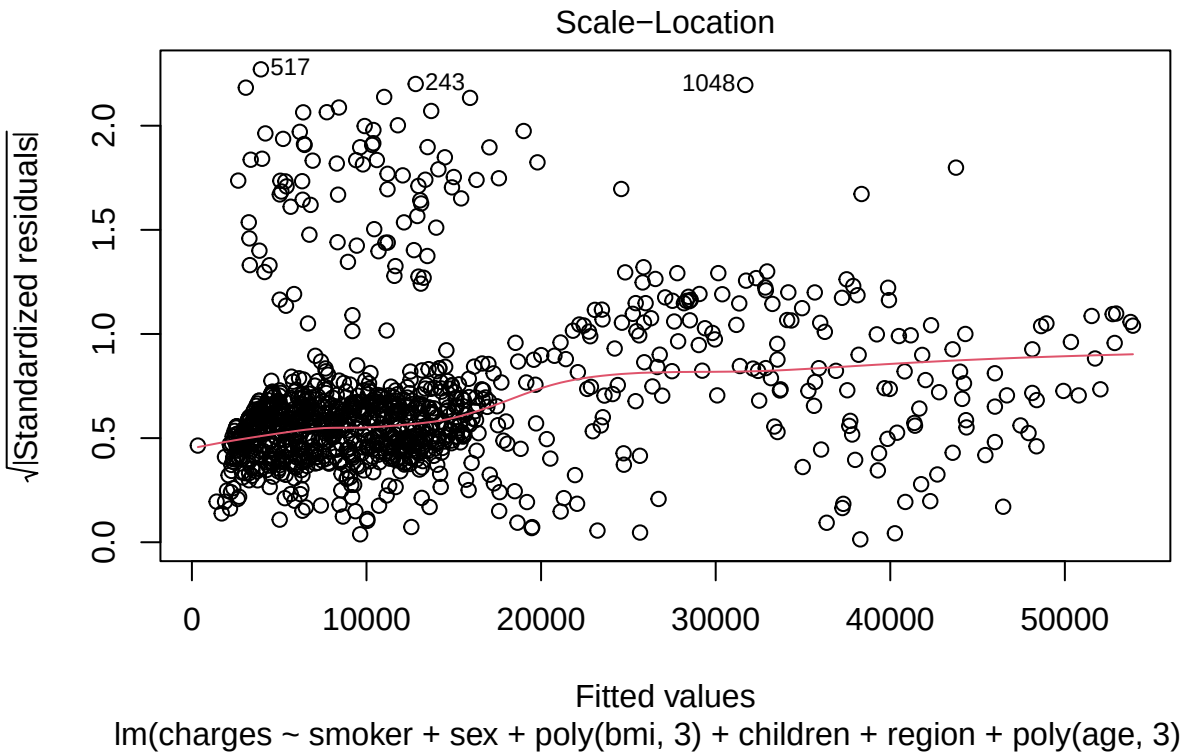
Plot Chosen Model

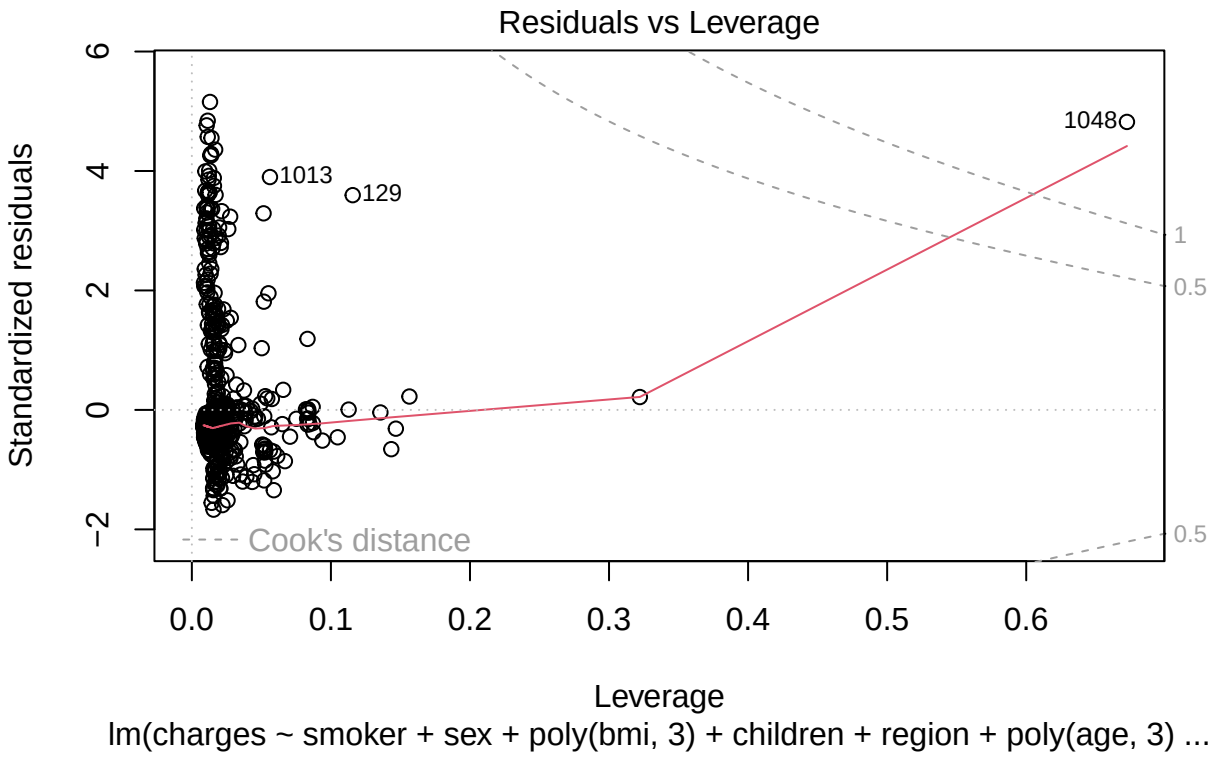
We plotted backstep to see the variance and residuals. The residuals were more scattered but there was a cloud present in the bottom left corner. This indicates that there is a problem with our residual plot. Moreover, the normality plot also shows us that the data is not normal. This made us realize that there might be an underlying issue with the data set that will need to be explored more.

```
plot(backstep)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(charges ~ smoker + sex + poly(bmi, 3) + children + region + poly(age, 3) ...
517
1047
843
−3
−2
−1
0
1
2
3
0
2
4


Scale−Location
517
243
1048
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
10000
20000
30000
40000
50000
Fitted values
lm(charges ~ smoker + sex + poly(bmi, 3) + children + region + poly(age, 3) ...



Shapiro and Bp Test

```
shapiro.test(backstep$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  backstep$residuals
## W = 0.62488, p-value < 2.2e-16
```

```
bptest(backstep)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  backstep
## BP = 27.834, df = 19, p-value = 0.08667
```

The Shapiro-Wilks test gives us a p-value of about 0, indicating the data is not normal, which aligns with our findings from the normality plot.

We performed the bptest to see how backstep was doing with the variance of the residuals. After performing the Bruesch Pagan test, we got a p value greater than 0.05, which tells us that the variance of the residuals is fairly constant.

Conclusion

Upon performing several tests and running various models, we conclude that the backstep model generated by the back selection was the best model. It was also the simpler model favored in the anova test. Based on the AIC and PRESS value as well, it had the lowest score. Moreover, based on the summary results we generated, it had a high adjusted R^2 value and had many terms that were significant based on their p-values.

During the regression analysis of the project, we faced many challenges using this data set due to lack of numerical variables. We only had 7 variables, half of them being categorical which limited us to make better predictive models. It was also difficult to find a model that had scattered residuals and good normality plots, which we concluded came from underlying problems in the data set.

We tried to solve this issue by using the categorical variables as the interaction terms and found that smoker produced the best results and was the most interesting interaction. From there, we performed various transformations, used forward and backward selection, and also cross validation to make sure that we found the best predictive model. After trying to find interactions between the variables, we found an interesting interaction between the variables smoker, age and charges. The interaction plot showed us that we might potentially need another variable to explain the interaction.

From this project, we realized that it is very important to thoroughly analyze the data and perform various methods to get the best regression model. In doing so, it can help families better manage their finances and potentially help maintain or lower insurance costs. Overall, we gained a deeper understanding of how regression analysis can be applied to scenarios such as this. We hope to improve upon our analysis and investigate outside variables that might be affecting our model.

Literature Review:

There were many international papers that analyzed the medical costs using these data sets. In the paper Regression Analysis and Prediction Of Medical Insurance cost by Ayushi Bharti and Lokesh Malik they used 7 attributes and performed regression techniques that are Ridge Regression, Lasso Regression, Random forest, and Elastic Net. They were able to conclude that the best model was using the Random forest. In another paper called Predict Health Insurance Cost by using Machine Learning and DNN Regression Models by Mohamed hanafy and Omar M. A. Mahmoud, were able to get significant results as well. The findings they had showed that Stochastic Gradient Boosting offers the best efficiency, with an RMSE value of 0.380189, an MAE value of 0.17448, and an accuracy of 85.82. They concluded that Stochastic gradient boosting can be used in the estimation of insurance costs with better performance than other regression models.

Works Cited:

https://www.kaggle.com/datasets/mirichoi0218/insurance?select=insurance.csv https://www.researchgate.net/publication/348559741_Predict_Health_Insurance_Cost_by_using_Machine_Learning_and_DNN_Regression_Models https://ijcrt.org/papers/IJCRT2203462.pdf https://healthpayerintelligence.com/news/health-insurance-costs-placing-stress-on-majority-of-americans#:~:text=Seventy%2Dtwo%20percent%20of%20m
https://www.bls.gov/news.release/cesan.nr0.htm