Regression Analysis (MATH 1312)

Final Project

"What's driving the health insurance up, in the US?"

Name: Vamika Pardeshi

Student IDs: s3701024

Date: 9th June 2019

Table of Contents

Introduction

As per the saying- “A sound mind lives in a healthy body”. With the increase in health-related problems, People have become more aware about maintaining an insured life to avoid any financial risks. But the fact that getting a health insurance comes at a high cost, is one of the concerns. I have come across an article about the same, published by CNBC in 2018, addressing the concern over high health care costs for the individuals in the United States and factors involved. The United States is one of those countries that overspends on an average ~$9,500 Mn on its healthcare expenditure. This report will scrutinize all those parameters that are responsible for the variation in the costs of health-insurance.

During my proposal I will try and answer the following two questions:

- Within the United States, is it true that the insurance premium cost is high based on the location?
- If not, then what are the key variables apart from the regions explaining higher insurance costs?

Dataset

To support my idea, I have searched related datasets on various public data sites and found “Medical cost personal dataset” to be the most suitable according to my requirements for further analysis. The data is sourced from Kaggle. The dataset comprises 7 variables (predictors) and 1338 observations. All the further actions will be taken out on this dataset. The variables in this dataset are sex, BMI, children, smoker, region, and target variable charges.

Target Feature

The target feature for this dataset is ‘**Charges**’ variable. ‘Charges’ has chosen as a response variable because it is the medical cost for an individual, that is dependent on other features/predictors. It varies based on the factors affecting the cost of insurance.

Descriptive Features (Predictors)

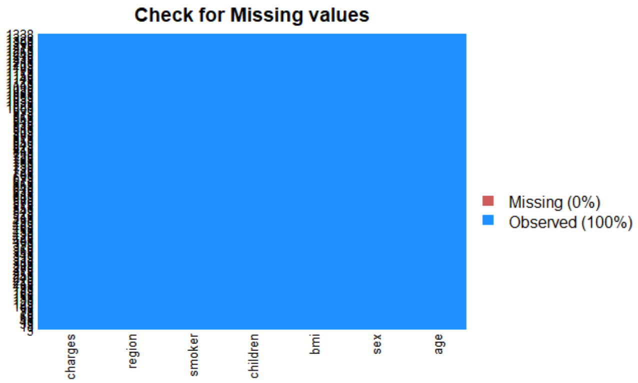
The description of the attributes that has been chosen for descriptive features are readily comprehensible and are as follows-

Age- age of primary beneficiary
Sex- gender of the insurer, female, male
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 ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9.
children: Number of children covered by health insurance / Number of dependents
smoker: Smoking
region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

Data-Preprocessing

Data has been imported to R, using read.csv() function and named it as “US_health”. The size of the data is confirmed using ‘dim()’ function. The structure of the data is viewed by using ‘str()’ function, in order to display list contents compactly. To scan the data quickly on the basis of statistics, summary() function is used.

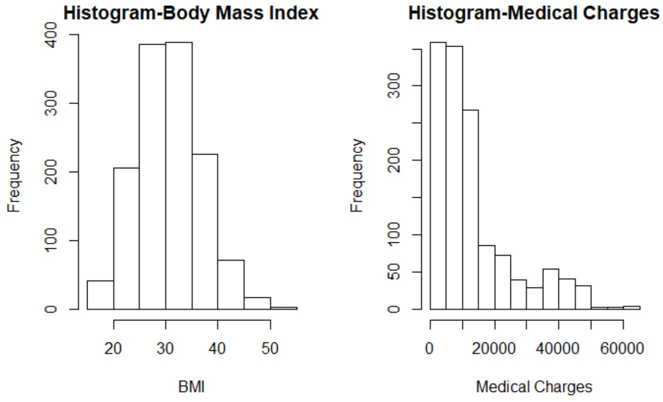
It is very important that the dataset should not contain any missing and NA values. A miss map is created to check the missing values in the data. The map shows that there is no missing value present in the data. Also, colSums(is.na()) is used to check for the NA values, and the result shows that there is no NA value as well.



Data Exploration

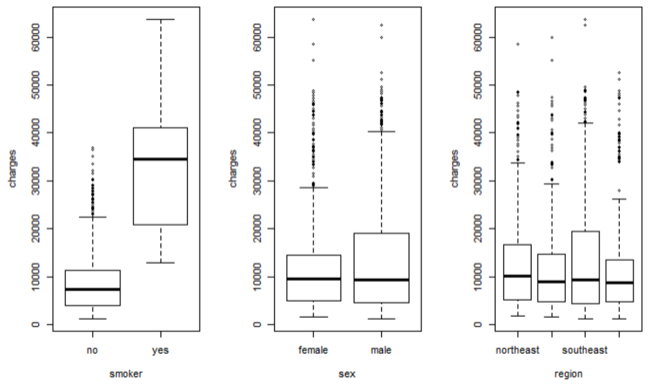
To support the hypothesis and to apprehend the dataset statistically, it is always better to visualize the data. So, histograms are used for "Univariate Visualization", whereas, "Box plot", "Double bar graph" and "scatter plot" are used for "Multi-variate Visualizations".

1.) Univariate Visualization

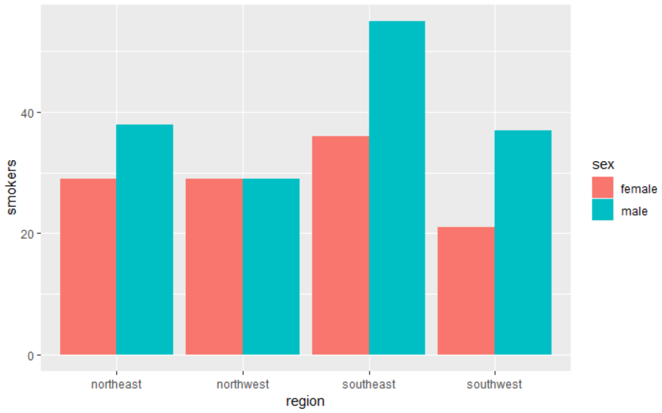


From the above histogram for BMI, it can be seen that it is normally distributed. Although, it is rightly skewed for the medical charges, showing that in this dataset, the observations with high medical costs are lesser.

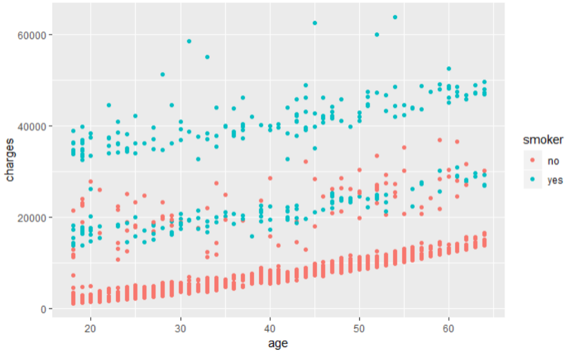
2.) Multivariate Visualization



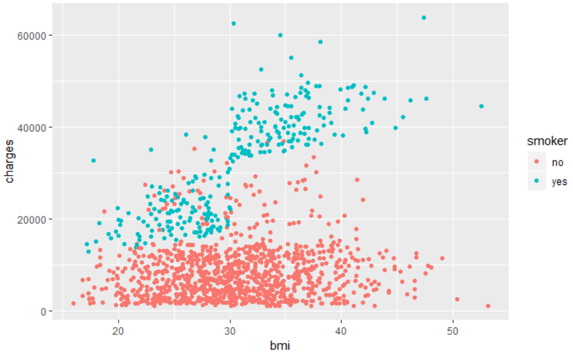
The above boxplots for gender and region are having more outliers for that of high charges than comparatively lower charges. The medians for region and genders are almost the same. The most expected point to note here is medical costs are very high for the people who smoke.



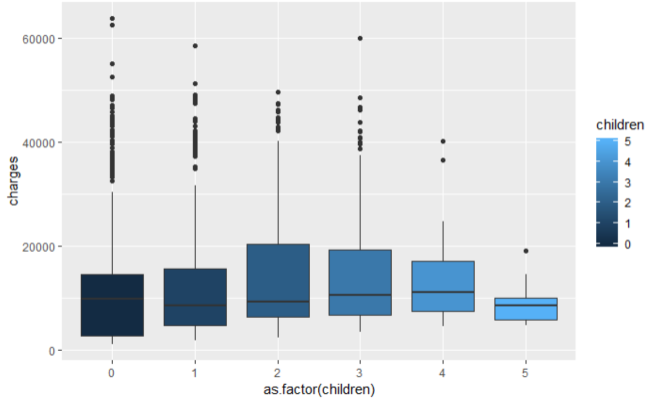
From the above double bar graph, it can be observed that both male and female equal smokers in almost all the 4 regions. Although, in southeast male smokers are more than that of female.



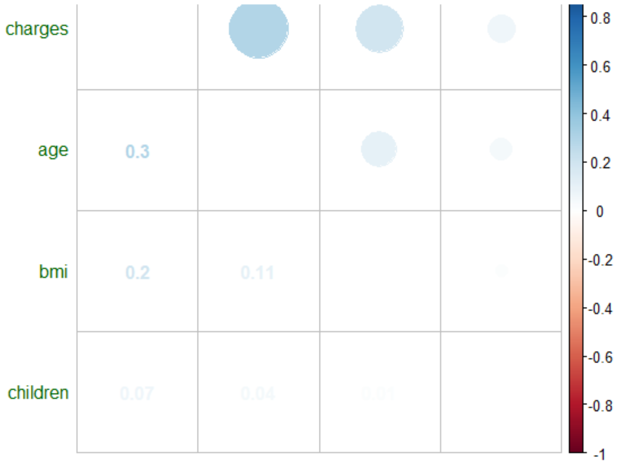
The above scatter plot is showing that higher charges are directly proportional to old-aged smokers. In other words, as the age increases for the smokers so do their health-related issues, that eventually results in higher medical costs.



Based on the above plot, it is found out that those smokers who have higher BMI tend to pay higher for the insurance premiums. In other words, the higher bmi smokers are more vulnerable to diseases, and hence will pay higher medical costs.



The above boxplot is giving an idea that people having 2 or 3 children are paying higher insurance charges than those having lesser or a greater number of children. Again, all the medians are almost at the same level.



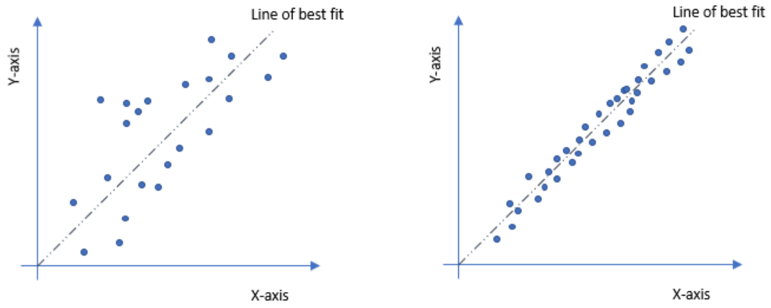
Using correlation plot, it can be figured out that medical cost is highly correlated with ‘age’. In other words, as the age increase the medical cost also increases. With the help of visualisation it can be concluded that ‘charges’ is highly dependent on ‘customer related data’.

Methodology

In order to understand the relationship between independent variables and dependent variable ‘target’. Regression analysis on data set is performed using Linear regression, Lasso Regression (Least absolute shrinkage selector operator), and Ridge regression.

Linear Regression

Linear regression is among the simplest and most widely used technique for predictive modeling. Linear regression model looks for statistical relationship and not for deterministic relationship[1], where the main objective is to obtain a line of best fit for the data. The line of best is a diagonal line for which the prediction errors are very small as possible.



Moderate linear model

Strong and positive linear model

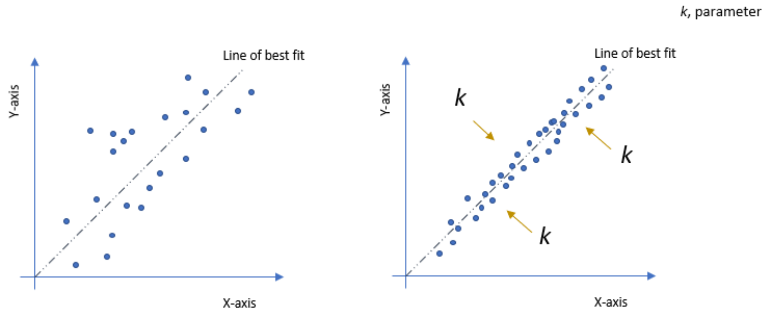
Figure 1, Linear regression model type

The closer the data point are closer to the line of best fit, the better is the model with less variation which is explained by adjusted R^2 value. A scatterplot shown above determines the strength of the relationship between two variables, and numerical measurement of association between two variables is generally explained by coefficient of correlations, between -1 and 1.

Linear regression equation is: $Y = a + bX$. Where, a is the intercept, Y is the dependent variable, slope is b and X is the explanatory variable.

Ridge Regression

Multiple regression data which has multicollinearity, ridge regression is a technique that is commonly used. If there is multicollinearity present then there is an unbiasedness in their least square estimates, but they might be placed far away from their actual values which is explained by higher variation. As it is known that multicollinearity creates inaccurate estimates for the coefficients by inflating the standard errors. Therefore, by adding a parameter or a degree of bias to the estimates, the standard errors explained by higher variance is minimized.



Regression model with multicollinearity

Regression model using Ridge analysis

Figure 2, Ridge regression model type

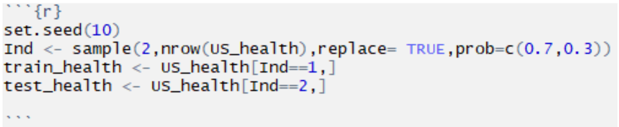
Ridge Regression equation is: *Y = a + be*. Where, *a* is the intercept, Y is the dependent variable, slope is *b* and *e* represents the errors are residuals.

Lasso Regression (Least absolute shrinkage selector operator)

Lasso regression is very similar to the ridge analysis, where Lasso selects only the best features while reducing the coefficients of less important features to 0, generally this practice is known as feature selection. Lasso uses L1 regression technique and is generally preferred when there are large number of variable present in the dataset. However, one of the challenge during the Lasso regression is that even if there are variables with smaller correlation, the lasso model will set the value for that feature to 0. As a result, it will lead to loss of information which further would lead to poor model performance.

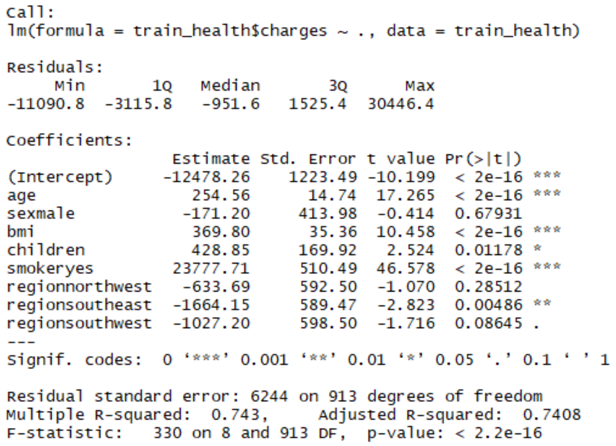
Model Selection

For model building, the first step is to split the data into 70-30 train and test.



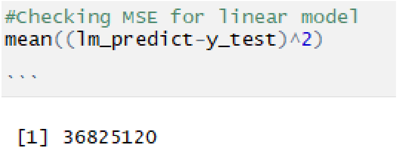
~~~
```{r}
set.seed(10)
Ind <- sample(2,nrow(US_health),replace= TRUE,prob=c(0.7,0.3))
train_health <- US_health[Ind==1,]
test_health <- US_health[Ind==2,]
```
~~~

A linear regression model is then built based on this split. The summary statistics for the same is below for the train set.

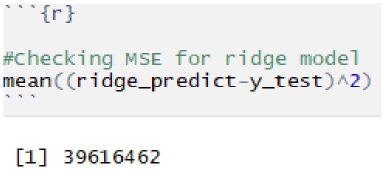


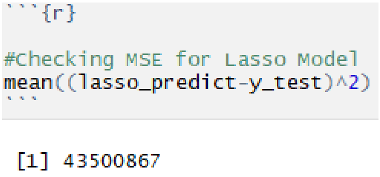
```
Call:
lm(formula = train_health$charges ~ ., data = train_health)

Residuals:
     Min       1Q   Median       3Q      Max 
-11090.8  -3115.8   -951.6   1525.4  30446.4 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -12478.26    1223.49 -10.199  < 2e-16 ***
age                254.56      14.74  17.265  < 2e-16 ***
sexmale           -171.20     413.98  -0.414  0.67931    
bmi                369.80      35.36  10.458  < 2e-16 ***
children           428.85     169.92   2.524  0.01178 *  
smokeryes        23777.71     510.49  46.578  < 2e-16 ***
regionnorthwest   -633.69     592.50  -1.070  0.28512    
regionsoutheast  -1664.15     589.47  -2.823  0.00486 ** 
regionsouthwest  -1027.20     598.50  -1.716  0.08645 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6244 on 913 degrees of freedom
Multiple R-squared:  0.743,	Adjusted R-squared:  0.7408 
F-statistic:   330 on 8 and 913 DF,  p-value: < 2.2e-16
```

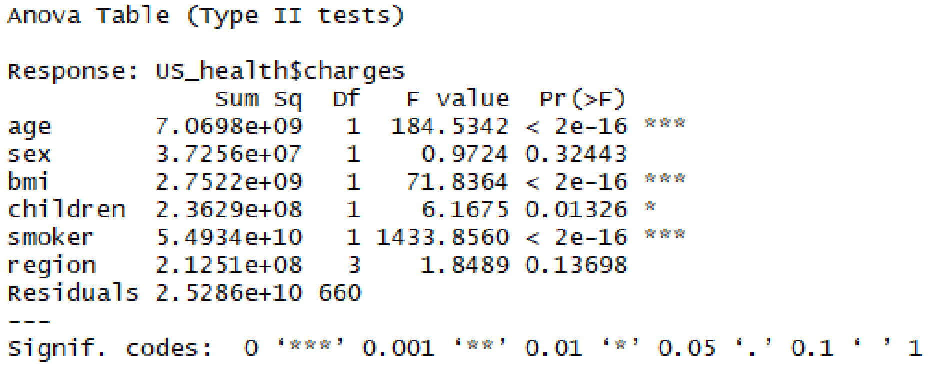
A sample data is the created, in which, the ‘x’ is containing all the variables but the target variable and the ‘y’ is containing only the target variable. The sample train set is then divided into half and the sample test set is made to consist everything but the sample train set. Linear and ridge models were then tried to fit, and the differences came out to be nominal. Hence, ridge model was used to check whether it can improve the OLS estimations. After checking the ridge model for the coefficients, it turned out to be not significant as the coefficients estimates were coming out to be conservative. So, lasso model then tried to fit and check for the coefficients again. Lasso model, as expected, set the unimportant variables to zero and gave preferences to the three most important predictors, i.e., age, bmi and smokers. Lasso can be chosen as the best model, but it needs to have the lowest MSE value. Post checking the MSE values for the three models, linear regression model came out to be with least MSE score, hence, it is selected as the best model among the three.



```
#Checking MSE for linear model
mean((lm_predict-y_test)^2)

```

 [1] 36825120
```



```
```{r}

#Checking MSE for ridge model
mean((ridge_predict-y_test)^2)
```

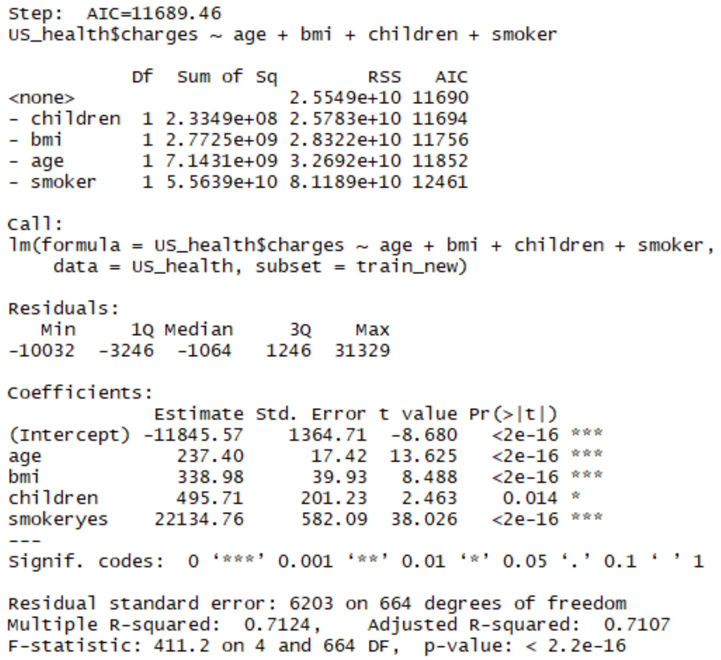
 [1] 39616462
```



```
```{r}

#Checking MSE for Lasso Model
mean((lasso_predict-y_test)^2)
```

 [1] 43500867
```

# Predictors Selection

Summary() function is used to get the descriptive statistics and the coefficients estimates in order to get the equation of the best fit for linear model.  The equation came out to be-

y=  -11544.86 +(236.65*age) +(477.06*sexmale) +(354.18*bmi) + (500.36*children) +(22107.37*smokeryes) -(1101.03*regionnorthwest) -(1409.71*regionsoutheast) -(1363.39*regionsouthwest)

To select the best predictors among all the predictors, 5 methods are used, and those are, ANOVA, All possible subset regression, Forward regression, Backward regression and Stepwise regression. According to ANOVA statistics, age, BMI and smoker variables are the most significant ones. Stepwise and backward regression returned the same results, i.e., age, BMI, smoker and children are the best predictors. And, Forward and all possible subset regression gave the same results, i.e., all the predictors but sex are the best ones. So, to summarize, it can be said that, 3 possible models can be built, and those are-

1. MODEL 1- Based on ANOVA- age, bmi and smoker



```
Anova Table (Type II tests)

Response: US_health$charges
              Sum Sq Df   F value   Pr(>F)
age   7.0698e+09   1 184.5342 < 2e-16 ***
sex   3.7256e+07   1   0.9724 0.32443
bmi   2.7522e+09   1   71.8364 < 2e-16 ***
children 2.3629e+08   1   6.1675 0.01326 *
smoker   5.4934e+10   1 1433.8560 < 2e-16 ***
region   2.1251e+08   3   1.8489 0.13698
Residuals 2.5286e+10 660

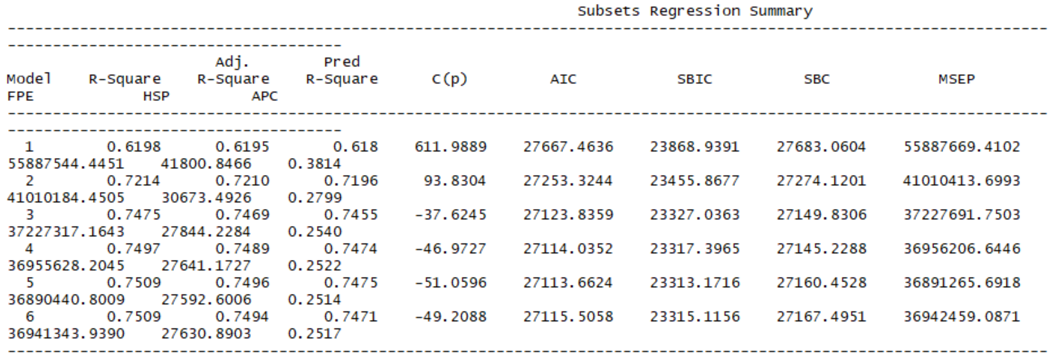
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
```

2. MODEL 2- Based on Stepwise and backward- age, bmi, smoker and children

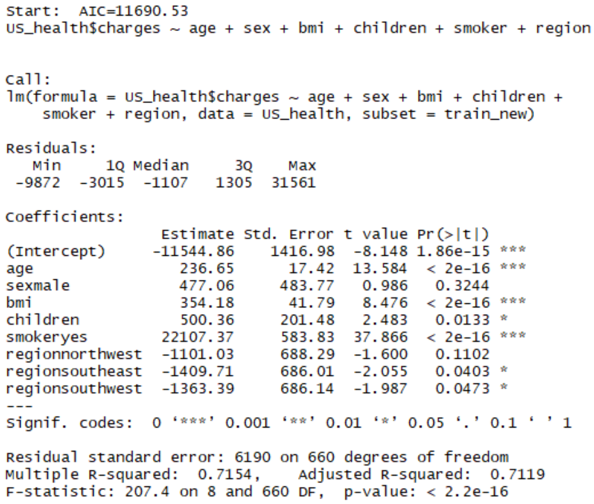


```
Step:  AIC=11689.46
US_health$charges ~ age + bmi + children + smoker

           Df  Sum of Sq        RSS   AIC
<none>                   2.5549e+10 11690
- children  1 2.3349e+08 2.5783e+10 11694
- bmi       1 2.7725e+09 2.8322e+10 11756
- age       1 7.1431e+09 3.2692e+10 11852
- smoker    1 5.5639e+10 8.1189e+10 12461

Call:
lm(formula = US_health$charges ~ age + bmi + children + smoker,
    data = US_health, subset = train_new)

Residuals:
   Min     1Q Median     3Q    Max
-10032  -3246  -1064   1246  31329

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11845.57    1364.71  -8.680   <2e-16 ***
age            237.40      17.42  13.625   <2e-16 ***
bmi            338.98      39.93   8.488   <2e-16 ***
children       495.71     201.23   2.463    0.014 *
smokeryes    22134.76     582.09  38.026   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6203 on 664 degrees of freedom
Multiple R-squared:  0.7124,    Adjusted R-squared:  0.7107
F-statistic: 411.2 on 4 and 664 DF,  p-value: < 2.2e-16
```

3. MODEL 3- Based on Forward and all possible subset regression- age, bmi, smoker, children and region



```
                                                        Subsets Regression Summary
-----------------------------------------------------------------------------------------------------------------
------------------------------------
                 Adj.        Pred
Model    R-Square    R-Square    R-Square      C(p)        AIC          SBIC         SBC           MSEP
FPE            HSP         APC
-----------------------------------------------------------------------------------------------------------------
------------------------------------
  1        0.6198      0.6195       0.618    611.9889    27667.4636    23868.9391    27683.0604    55887669.4102
55887544.4451    41800.8466    0.3814
  2        0.7214      0.7210      0.7196     93.8304    27253.3244    23455.8677    27274.1201    41010413.6993
41010184.4505    30673.4926    0.2799
  3        0.7475      0.7469      0.7455    -37.6245    27123.8359    23327.0363    27149.8306    37227691.7503
37227317.1643    27844.2284    0.2540
  4        0.7497      0.7489      0.7474    -46.9727    27114.0352    23317.3965    27145.2288    36956206.6446
36955628.2045    27641.1727    0.2522
  5        0.7509      0.7496      0.7475    -51.0596    27113.6624    23313.1716    27160.4528    36891265.6918
36890440.8009    27592.6006    0.2514
  6        0.7509      0.7494      0.7471    -49.2088    27115.5058    23315.1156    27167.4951    36942459.0871
36941343.9390    27630.8903    0.2517
-----------------------------------------------------------------------------------------------------------------
```



```
Start:  AIC=11690.53
US_health$charges ~ age + sex + bmi + children + smoker + region


Call:
lm(formula = US_health$charges ~ age + sex + bmi + children + 
    smoker + region, data = US_health, subset = train_new)

Residuals:
   Min     1Q Median     3Q    Max 
 -9872  -3015  -1107   1305  31561 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11544.86    1416.98  -8.148 1.86e-15 ***
age                236.65      17.42  13.584  < 2e-16 ***
sexmale            477.06     483.77   0.986   0.3244    
bmi                354.18      41.79   8.476  < 2e-16 ***
children           500.36     201.48   2.483   0.0133 *  
smokeryes        22107.37     583.83  37.866  < 2e-16 ***
regionnorthwest  -1101.03     688.29  -1.600   0.1102    
regionsoutheast  -1409.71     686.01  -2.055   0.0403 *  
regionsouthwest  -1363.39     686.14  -1.987   0.0473 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6190 on 660 degrees of freedom
Multiple R-squared:  0.7154,    Adjusted R-squared:  0.7119 
F-statistic: 207.4 on 8 and 660 DF,  p-value: < 2.2e-16
```

Model Building

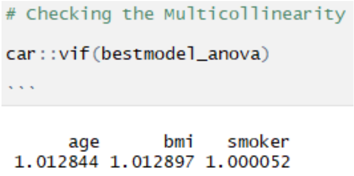
First Model

First model is built on the basis of ANOVA results with three of the best predictors, i.e., age, bmi, smoker. On the basis of summary the equation came out to be-

y= -12037.77 +(256.42*age) +(338.80*bmi) + (23670.40*smoker)

Model diagnostic for first model

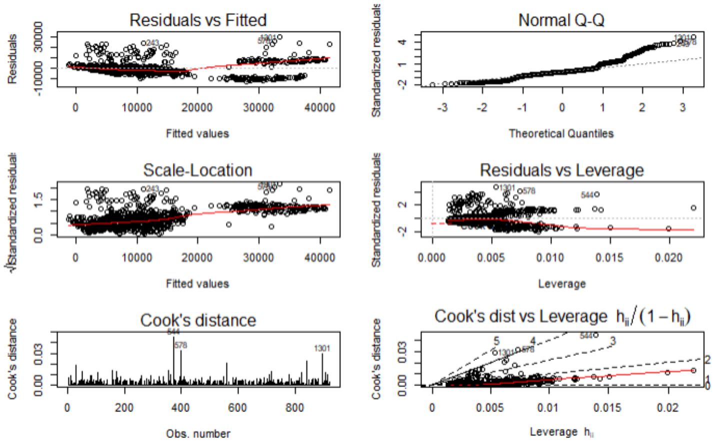
Firstly, the multicollinearity was diagnosed using VIF scores. There are no major differences noticed in the results, and all the values are closed to 1. Hence, it can be said that Multicollinearity is not present among the variables.



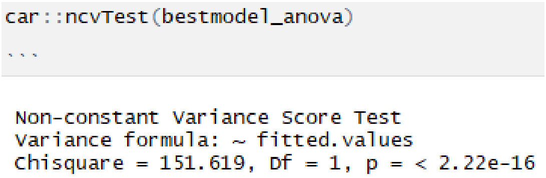
```
# Checking the Multicollinearity
car::vif(bestmodel_anova)
```
```
```

```
     age      bmi   smoker 
1.012844 1.012897 1.000052 
```

Residuals assumptions are applied to this model, and 6 graphs are plotted. It can be noticed that all the variables are independent with zero mean and are mostly normally distributed. Therefore, residuals assumptions are fulfilled.



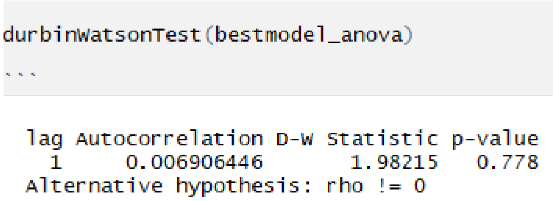
With the 'Non-constant variance score test', the p-value is less than alpha 0.05, therefore the null hypothesis is rejected stating that the residuals variance is constant.



```
car::ncvTest(bestmodel_anova)
```
```
```

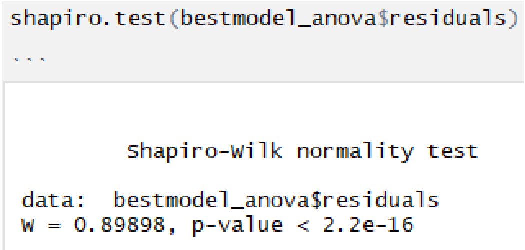
```
Non-constant Variance Score Test 
Variance formula: ~ fitted.values 
Chisquare = 151.619, Df = 1, p = < 2.22e-16
```

Next, is the durbin Watson test for testing the autocorrelated errors. With p-value greater than 0.05, the null hypothesis is not rejected, confirming that NO autocorrelation is present in the model.



```
durbinWatsonTest(bestmodel_anova)
```
```

 lag Autocorrelation D-W Statistic p-value
   1     0.006906446      1.98215   0.778
 Alternative hypothesis: rho != 0
```

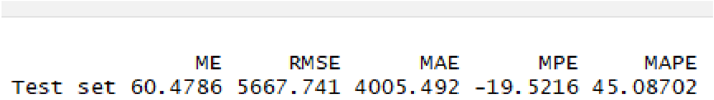
Although by shapiro wilk normality test, the p-value is less than 0.05, hence rejecting the null hypothesis, which states that the data is normally distributed. Hence, for this model the data is not normally distributed.



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


	Shapiro-Wilk normality test

data:  bestmodel_anova$residuals
W = 0.89898, p-value < 2.2e-16
```

Prediction for first model

The prediction for the models are done by using predict() function. In order to solve probabilities of prediction, type is set to “response”. Residual is set with this predict result subtracted from the target variable. A prediction is then made by setting actual and predicted parameters, and accuracy is calculated.



```
              ME     RMSE      MAE      MPE     MAPE
Test set 60.4786 5667.741 4005.492 -19.5216 45.08702
```

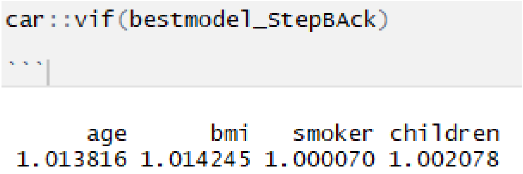
Second Model

Second model is built on the basis of stepwise and backward regression results with four of the best predictors, i.e., age, bmi, smoker and children. On the basis of summary the equation came out to be-

y= -12565.34 +(255.28*age) +(341.87*bmi) + (23675.65*smokeryes) +(421.95*children)

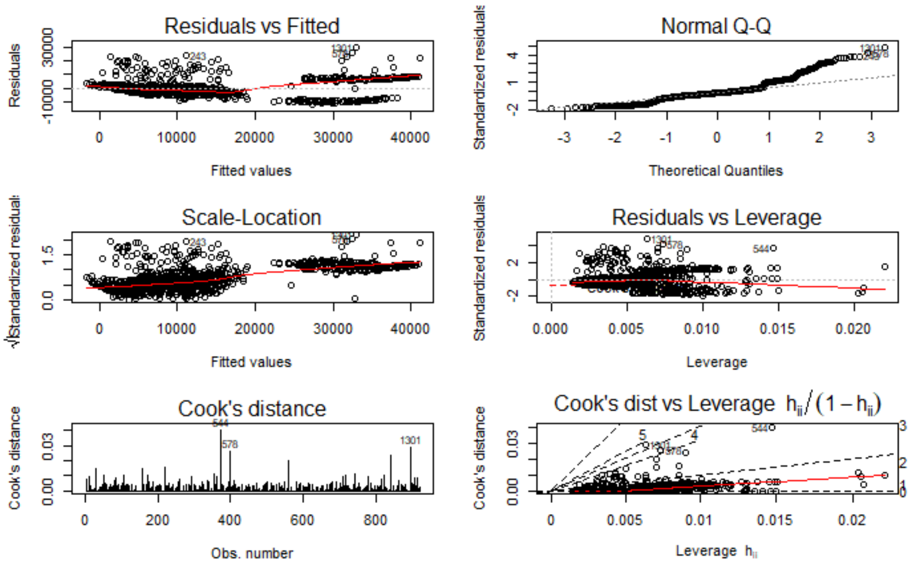
Model diagnostic for second model

Again, There are no major differences noticed in the VIF score results, and all the values are closed to 1. Hence, it can be said that Multicollinearity is not present among the variables.

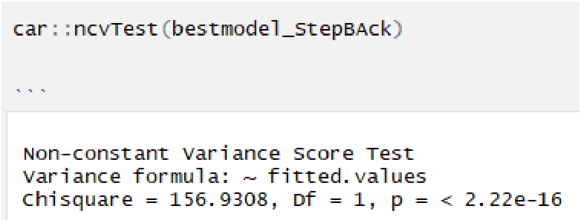


```
car::vif(bestmodel_StepBAck)
```
```
     age      bmi   smoker children 
1.013816 1.014245 1.000070 1.002078 
```

Residuals assumptions are applied to this model, and 6 graphs are plotted. It can be noticed that all the variables are independent with zero mean and are mostly normally distributed. Therefore, residuals assumptions are fulfilled.



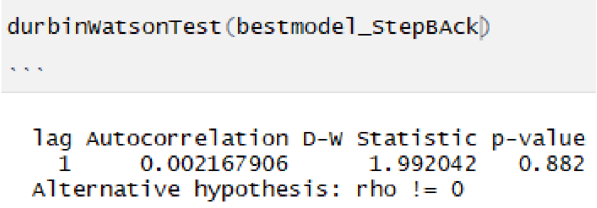
With the ‘Non-constant variance score test’, the p-value is less than alpha 0.05, therefore the null hypothesis is rejected stating that the residuals variance is constant.



```
car::ncvTest(bestmodel_stepBAck)

```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 156.9308, Df = 1, p = < 2.22e-16
```

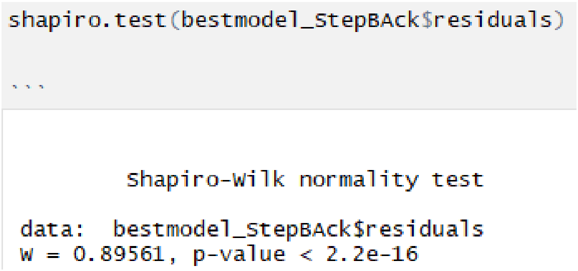
With p-value greater than 0.05, the null hypothesis is not rejected, confirming that NO autocorrelation is present in the model, by durbin Watson test for testing the autocorrelated errors.



```
durbinWatsonTest(bestmodel_stepBAck)
```

```
 lag Autocorrelation D-W Statistic p-value
   1     0.002167906      1.992042   0.882
 Alternative hypothesis: rho != 0
```

Although by shapiro wilk normality test, the p-value is less than 0.05, hence rejecting the null hypothesis, which states that the data is normally distributed. Hence, for this model the data is not normally distributed.

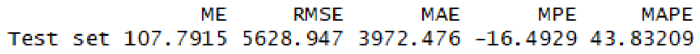


```
shapiro.test(bestmodel_stepBAck$residuals)
```

```

	Shapiro-Wilk normality test

data:  bestmodel_stepBAck$residuals
W = 0.89561, p-value < 2.2e-16
```

Prediction for second model

The prediction for the second model is done by the same process. The statistics that came out from this prediction is-



```
                ME     RMSE      MAE       MPE     MAPE
Test set 107.7915 5628.947 3972.476 -16.4929 43.83209
```

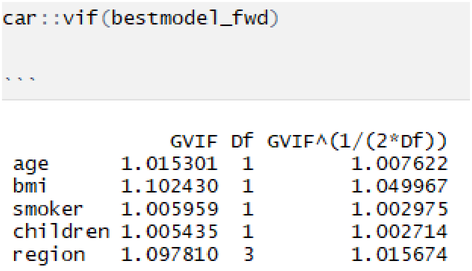
Third model

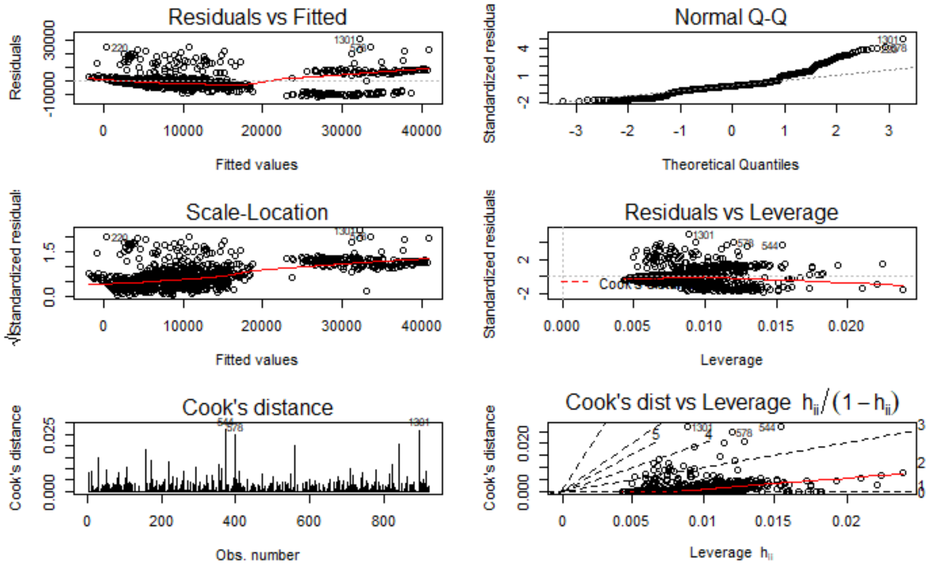
Third model is built on the basis of forward and all possible subset regression results with five of the best predictors, i.e., age, bmi, smoker, region and children. On the basis of summary the equation came out to be-

y= -12546.15 +(254.74*age) +(369.25*bmi) + (23756.57*smokeryes) +(428.42*children) -(638.41*regionnorthwest) -(1668.87*regionsoutheast) -(1030.88*regionsouthwest)

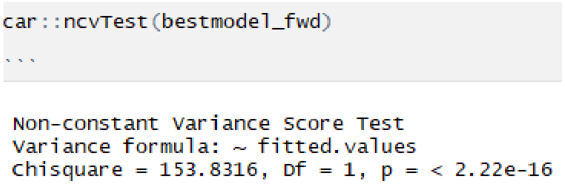
Model diagnostic for third model

Yet again, There are no major differences noticed in the VIF score results, and all the values are closed to 1. Hence, it can be said that Multicollinearity is not present among the variables.



```
car::vif(bestmodel_fwd)

```
```

```
             GVIF Df GVIF^(1/(2*Df))
age      1.015301  1        1.007622
bmi      1.102430  1        1.049967
smoker   1.005959  1        1.002975
children 1.005435  1        1.002714
region   1.097810  3        1.015674
```

Residuals assumptions are applied to this model, and 6 graphs are plotted. It can be noticed that all the variables are independent with zero mean and are mostly normally distributed. Therefore, residuals assumptions are fulfilled.


```

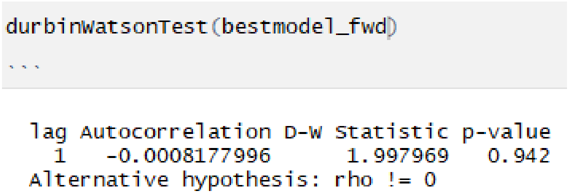
With the ‘Non-constant variance score test’, the p-value is less than alpha 0.05, therefore the null hypothesis is rejected stating that the residuals variance is constant.



~~~
car::ncvTest(bestmodel_fwd)
```
~~~

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 153.8316, Df = 1, p = < 2.22e-16
```

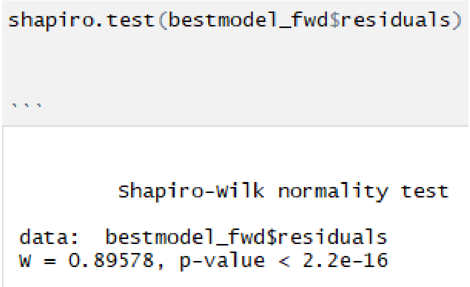
With p-value greater than 0.05, the null hypothesis is not rejected, confirming that NO autocorrelation is present in the model, by durbin Watson test for testing the autocorrelated errors.



~~~
durbinWatsonTest(bestmodel_fwd)
```
~~~

```
 lag Autocorrelation D-W Statistic p-value
   1   -0.0008177996      1.997969   0.942
 Alternative hypothesis: rho != 0
```

Although by shapiro wilk normality test, the p-value is less than 0.05, hence rejecting the null hypothesis, which states that the data is normally distributed. Hence, for this model the data is not normally distributed.

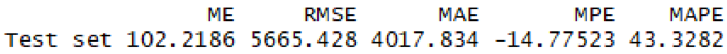


~~~
shapiro.test(bestmodel_fwd$residuals)


```
~~~

```

	Shapiro-Wilk normality test

data:  bestmodel_fwd$residuals
W = 0.89578, p-value < 2.2e-16
```

Prediction for Third model

The prediction for the second model is done by the same process. The statistics that came out from this prediction is-



```
               ME     RMSE      MAE       MPE    MAPE
Test set 102.2186 5665.428 4017.834 -14.77523 43.3282
```

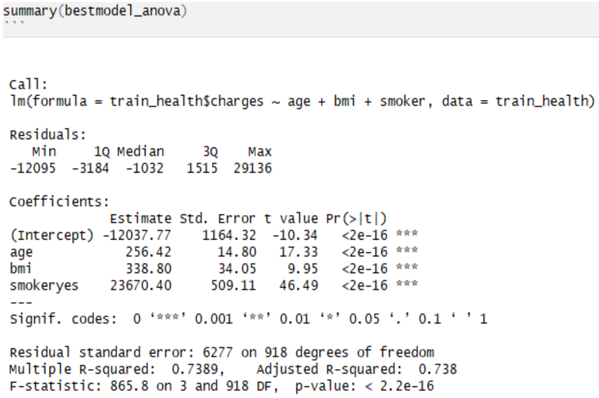
Results

The test assumptions for all the three models are almost the same for ‘Non-constant variance score test’, ‘durbin Watson test’, ‘shapiro wilk normality test’, residual analysis and VIF scores.

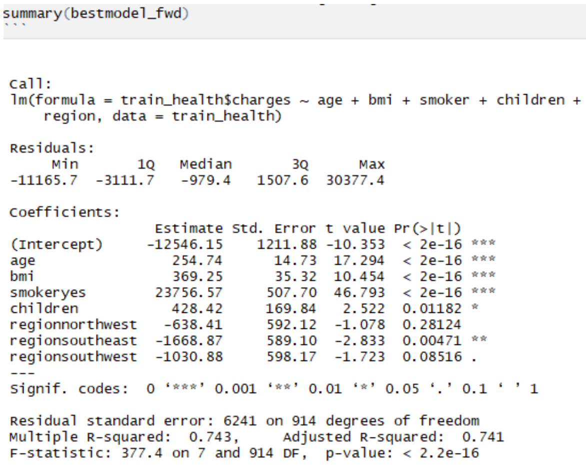
Hence, on this basis the best model can’t be selected. Summary statistics will come into the picture now.

Evaluating Interpretation

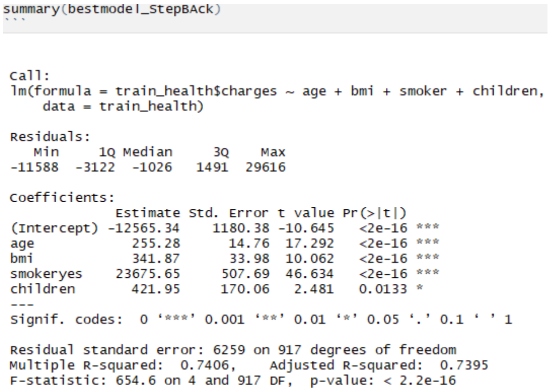
From the **summary statistics** for the three models, the 3 key evaluators are-



~~~
summary(bestmodel_anova)
```


Call:
lm(formula = train_health$charges ~ age + bmi + smoker, data = train_health)

Residuals:
   Min     1Q Median     3Q    Max 
-12095  -3184  -1032   1515  29136 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12037.77    1164.32  -10.34   <2e-16 ***
age            256.42      14.80   17.33   <2e-16 ***
bmi            338.80      34.05    9.95   <2e-16 ***
smokeryes    23670.40     509.11   46.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6277 on 918 degrees of freedom
Multiple R-squared:  0.7389,	Adjusted R-squared:  0.738 
F-statistic: 865.8 on 3 and 918 DF,  p-value: < 2.2e-16
~~~



~~~
summary(bestmodel_fwd)
```


Call:
lm(formula = train_health$charges ~ age + bmi + smoker + children + 
    region, data = train_health)

Residuals:
     Min       1Q   Median       3Q      Max 
-11165.7  -3111.7   -979.4   1507.6  30377.4 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -12546.15    1211.88 -10.353  < 2e-16 ***
age                254.74      14.73  17.294  < 2e-16 ***
bmi                369.25      35.32  10.454  < 2e-16 ***
smokeryes        23756.57     507.70  46.793  < 2e-16 ***
children           428.42     169.84   2.522  0.01182 *  
regionnorthwest   -638.41     592.12  -1.078  0.28124    
regionsoutheast  -1668.87     589.10  -2.833  0.00471 ** 
regionsouthwest  -1030.88     598.17  -1.723  0.08516 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6241 on 914 degrees of freedom
Multiple R-squared:  0.743,	Adjusted R-squared:  0.741 
F-statistic: 377.4 on 7 and 914 DF,  p-value: < 2.2e-16
~~~



~~~
summary(bestmodel_StepBAck)
```


Call:
lm(formula = train_health$charges ~ age + bmi + smoker + children, 
    data = train_health)

Residuals:
   Min     1Q Median     3Q    Max 
-11588  -3122  -1026   1491  29616 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12565.34    1180.38 -10.645   <2e-16 ***
age            255.28      14.76  17.292   <2e-16 ***
bmi            341.87      33.98  10.062   <2e-16 ***
smokeryes    23675.65     507.69  46.634   <2e-16 ***
children       421.95     170.06   2.481   0.0133 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6259 on 917 degrees of freedom
Multiple R-squared:  0.7406,	Adjusted R-squared:  0.7395 
F-statistic: 654.6 on 4 and 917 DF,  p-value: < 2.2e-16
~~~

Residuals- Residuals are actuals negatively predicted. The lowest the "max error", the better it is. Among the three models, the lowest max errors are for ANOVA based model and Stepwise/Backward regression-based model, and those are 29136 and 29616, respectively. In other words, approximately $29136 and $29616 max errors are substantial, as per the model underprediction.

p-value- The smallest the p-value, the better relationship with the response variable "charges" and also the non-zero coefficient. The p-value is less than 0.05 for all the three models, according to the summary. But, the coefficients are the highest significant for ANOVA based best model and next for Stepwise/Backward regression-based model.

Multiple R-squared value- It explains the measurement of variation can be explained by the data. All the three models are having close to 74%, i.e., approximately 74% of the variance can be explained by the model. Hence, ANOVA based model can still be chosen according to this.

So, the results are narrowed down to two of the models that are better than the third, and those are, ANOVA based model and Stepwise/Backward regression-based model.

The **deciding factor** can be the **Root mean squared error (RMSE)** values that came from the prediction and accuracy results of the three models. The lowest the RMSE value the better the model is. The lowest RMSE value is noticed from the Stepwise/Backward regression-based model with 5628.947**.**

Hence, **Stepwise/Backward regression-based model,** i.e**., the second model** can be concluded as the best model among the three models. So, the best regressors/predictors are **age, bmi, smoker and children.**

The equation of best fit for the best model is-

y= -12565.34 +(255.28*age) +(341.87*bmi) + (23675.65*smokeryes) +(421.95*children)

Discussion

Based on the regression results, it can be well concluded that the insurance charges tends to be on the higher side if an individual is a smoker, whether is underweight, overweight or obese, and is directly proportional to the age, which means as age increases the insurance charges tend to increase as well. However, based on all the linear regression analysis techniques, it can be concluded that “smoker” has the most significance as compared to other predictors in determining insurance charges for an individual in the US.

References

[1] Linear Regression — Detailed View
Towards Data Science. (2018). Linear Regression — Detailed View.
Available at: https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86

Appendix

Code ▾

Regression Analysis-Final Project

Vamika Pardeshi-s3701024

June 5, 2019

DATA IMPORT

Hide

```
setwd("C:\\Regression Analysis")
US_health <- read.csv("US_healthinsurance.csv")
head(US_health, n=10)
```

	age	sex	bmi	children	smoker	region	charges
	<int>	<fctr>	<dbl>	<int>	<fctr>	<fctr>	<dbl>
1	19	female	27.900	0	yes	southwest	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.471
5	32	male	28.880	0	no	northwest	3866.855
6	31	female	25.740	0	no	southeast	3756.622
7	46	female	33.440	1	no	southeast	8240.590
8	37	female	27.740	3	no	northwest	7281.506
9	37	male	29.830	2	no	northeast	6406.411
10	60	female	25.840	0	no	northwest	28923.137

1-10 of 10 rows

Hide

```
# checking the data size
dim(US_health)
```

```
[1] 1338    7
```

Hide

```
#checking the data structure
str(US_health)
```

```
'data.frame':   1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
 $ charges : num  16885 1726 4449 21984 3867 ...
```

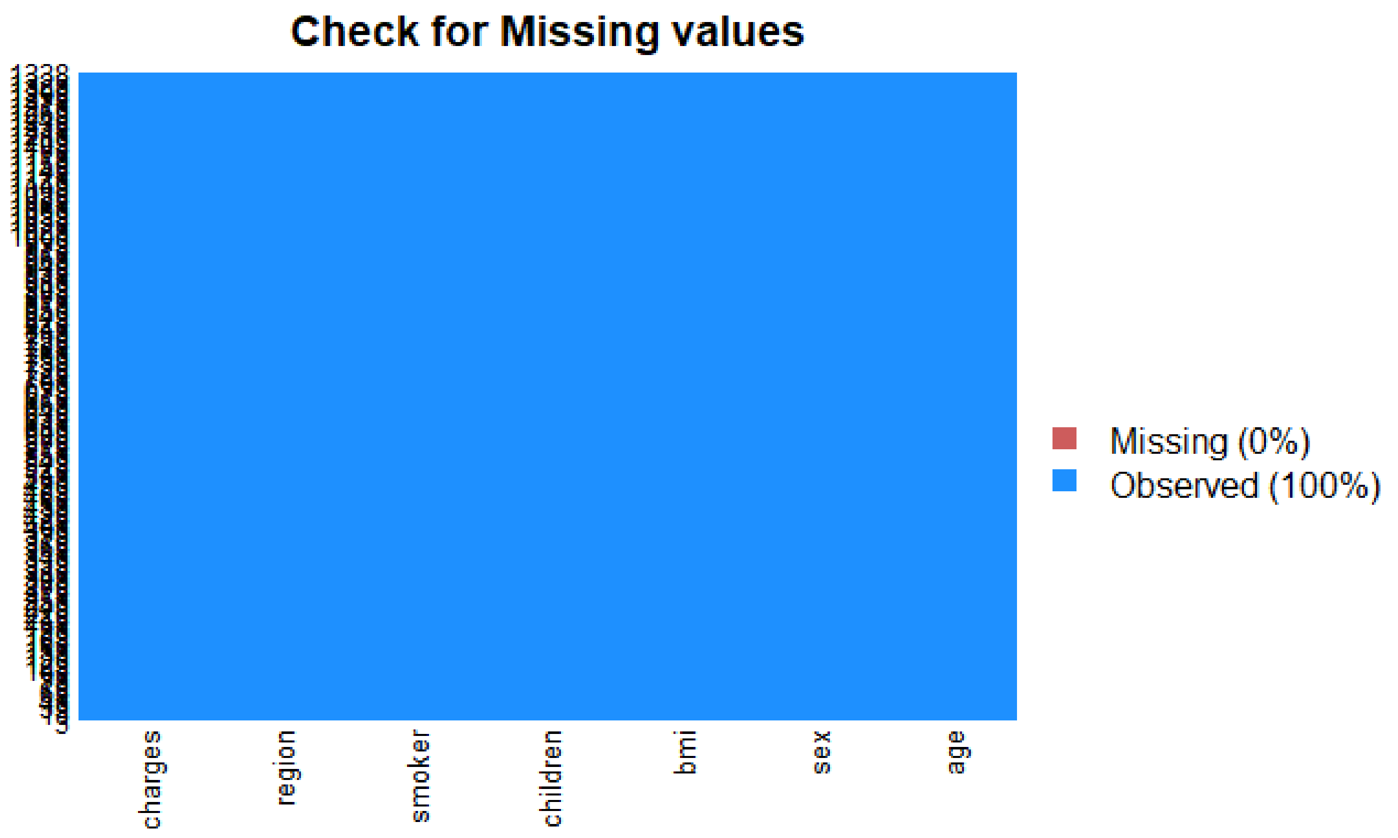
Hide

```
summary(US_health)
```

```
      age            sex           bmi           children       smoker          region        charges     
 Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064   northeast:324   Min.   : 1122  
 1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274   northwest:325   1st Qu.: 4740  
 Median :39.00                Median :30.40   Median :1.000              southeast:364   Median : 9382  
 Mean   :39.21                Mean   :30.66   Mean   :1.095              southwest:325   Mean   :13270  
 3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000                              3rd Qu.:16640  
 Max.   :64.00                Max.   :53.13   Max.   :5.000                              Max.   :63770  
```

Hide

```
missmap(US_health, main = "Check for Missing values")
```



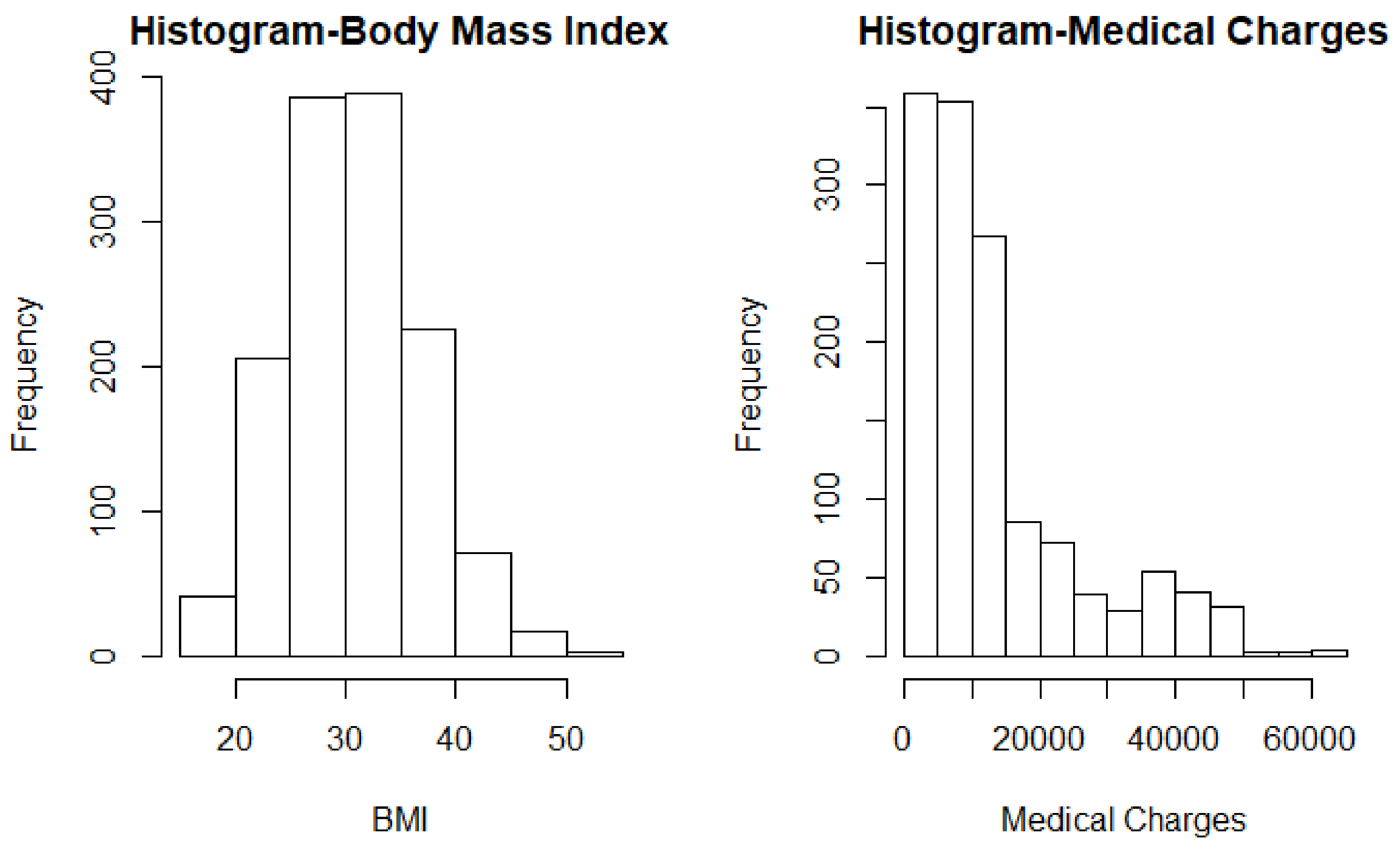
Hide

```
# checking for NA values
colSums(is.na(US_health))
```

```
     age      sex      bmi children   smoker   region  charges 
       0        0        0        0        0        0        0 
```

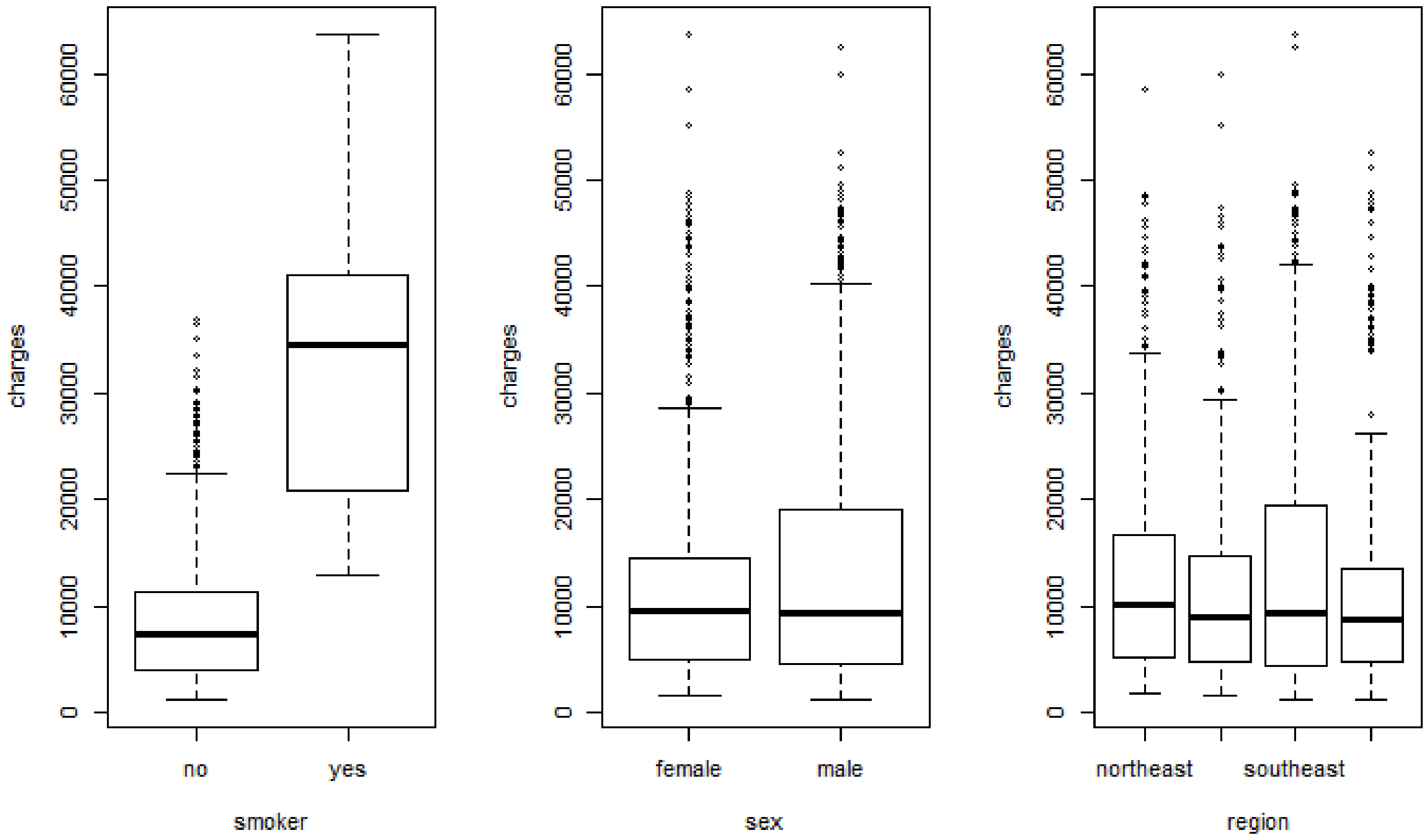
Hide

```
par(mfrow=c(1,2))
hist(US_health$bmi, xlab = "BMI",
     main = "Histogram-Body Mass Index")
hist(US_health$charges, xlab = "Medical Charges",
     main = "Histogram-Medical Charges")
```



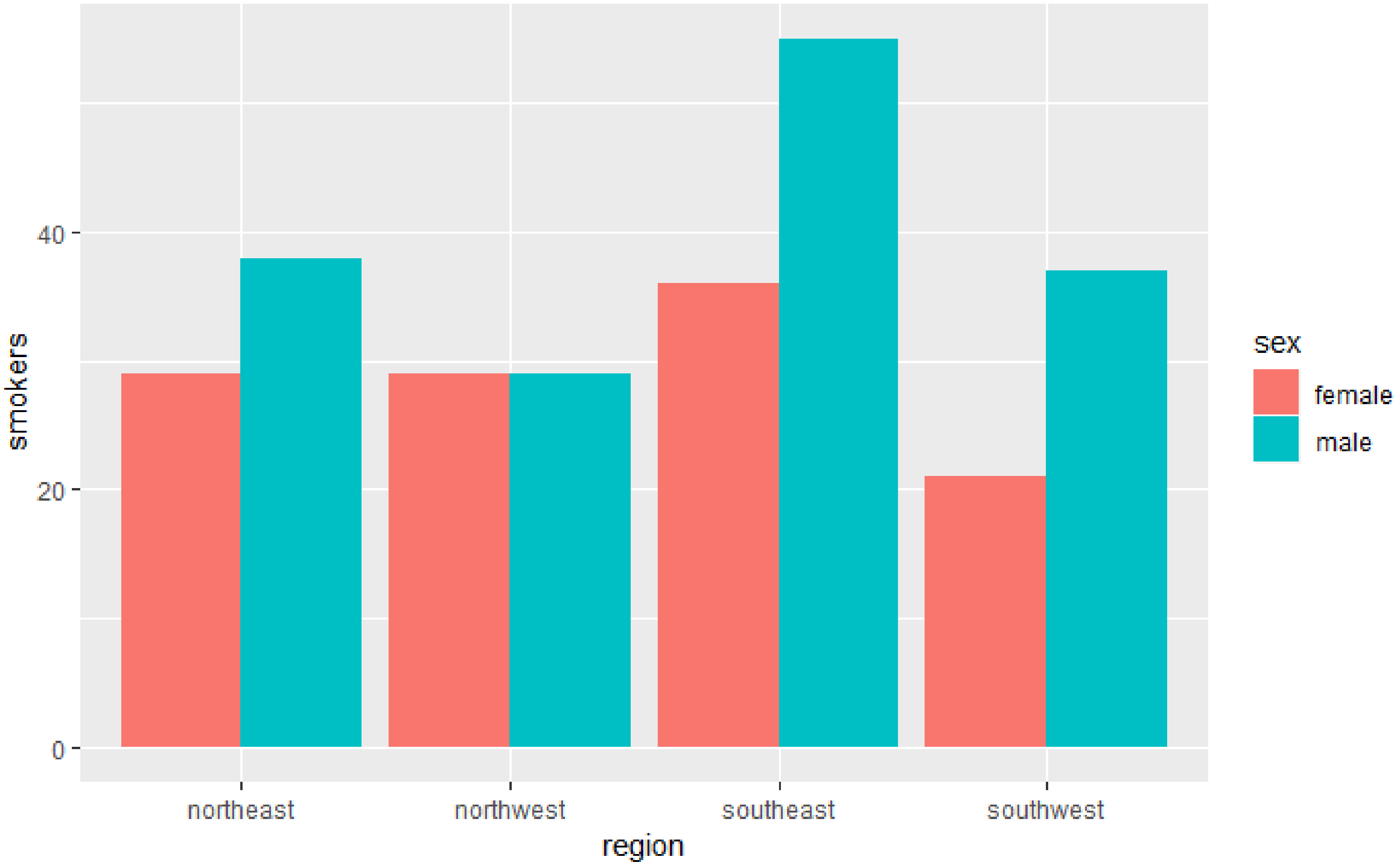
Hide

```
#plotting of charges against other factors like sex and smoker and region
par(mfrow=c(1,3))
with(US_health, plot(charges ~ smoker + sex + region))
```



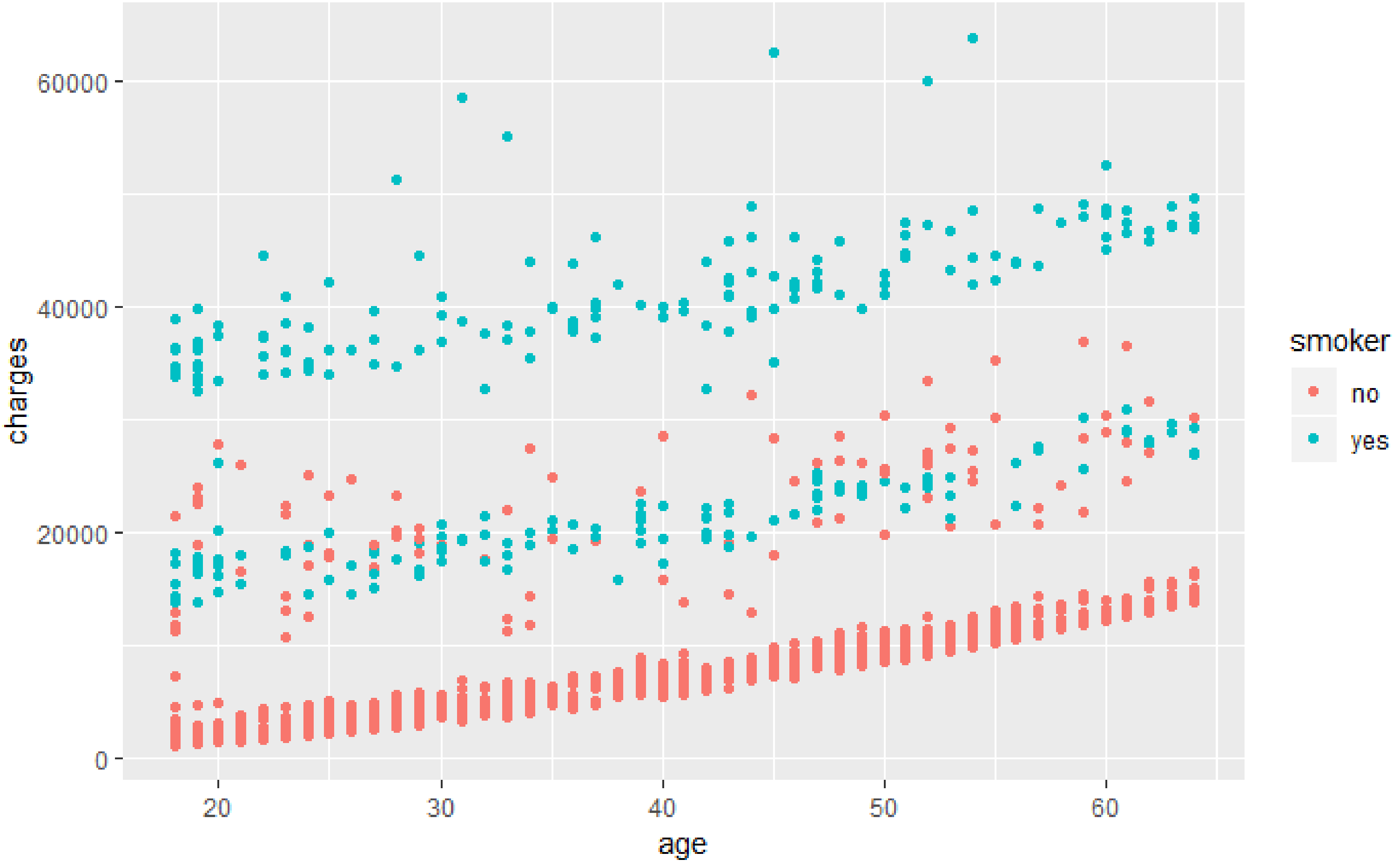
Hide

```
US_health %>% group_by(sex,region) %>% summarise(smokers=sum(smoker=="yes")) %>% ggplot(aes(x=region,y=smokers,fill=sex))+ge
om_bar(stat = "identity",position = "dodge")
```



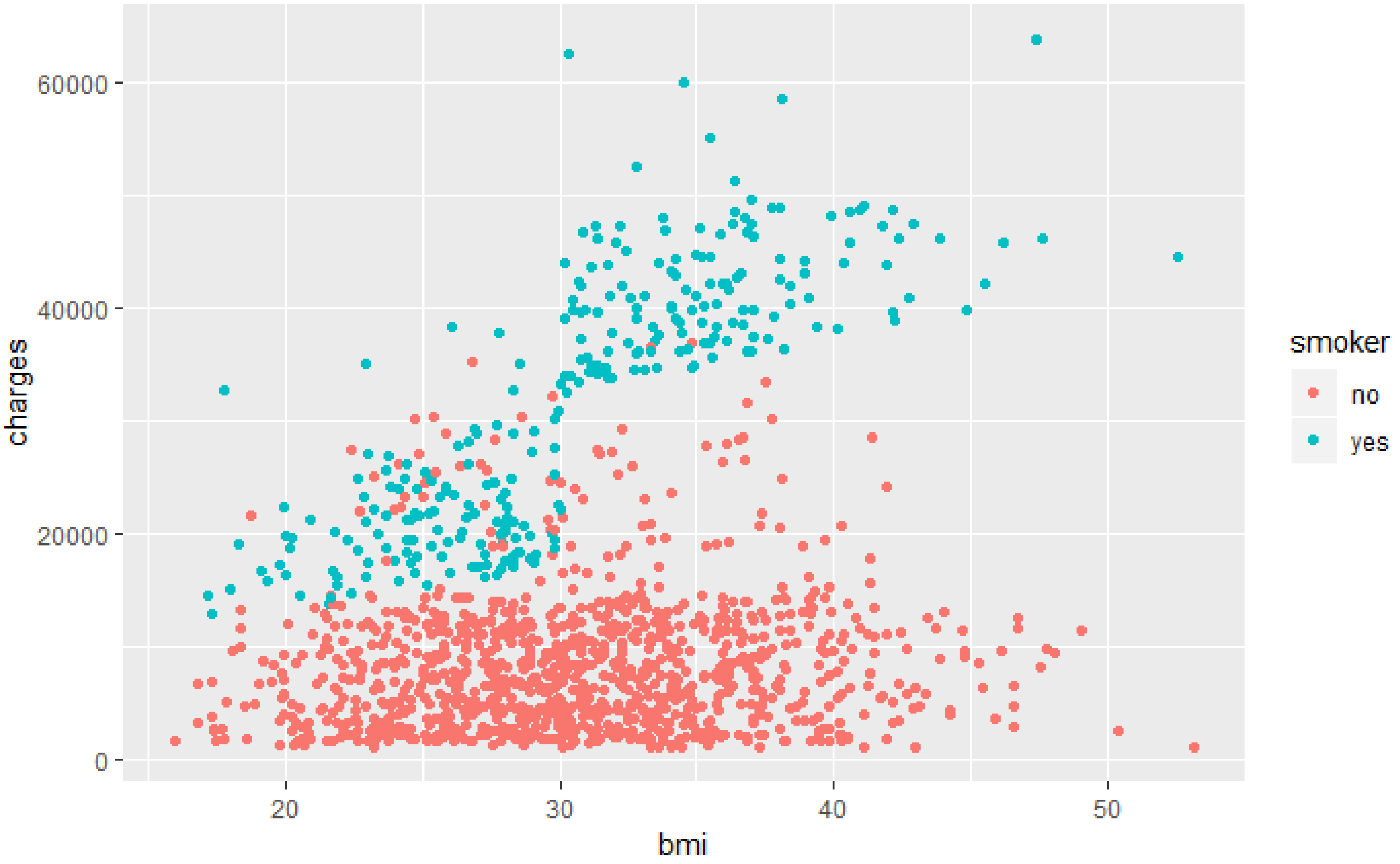
Hide

```
ggplot(US_health, aes(x = age, y = charges)) +
    geom_point(aes(color = smoker))
```



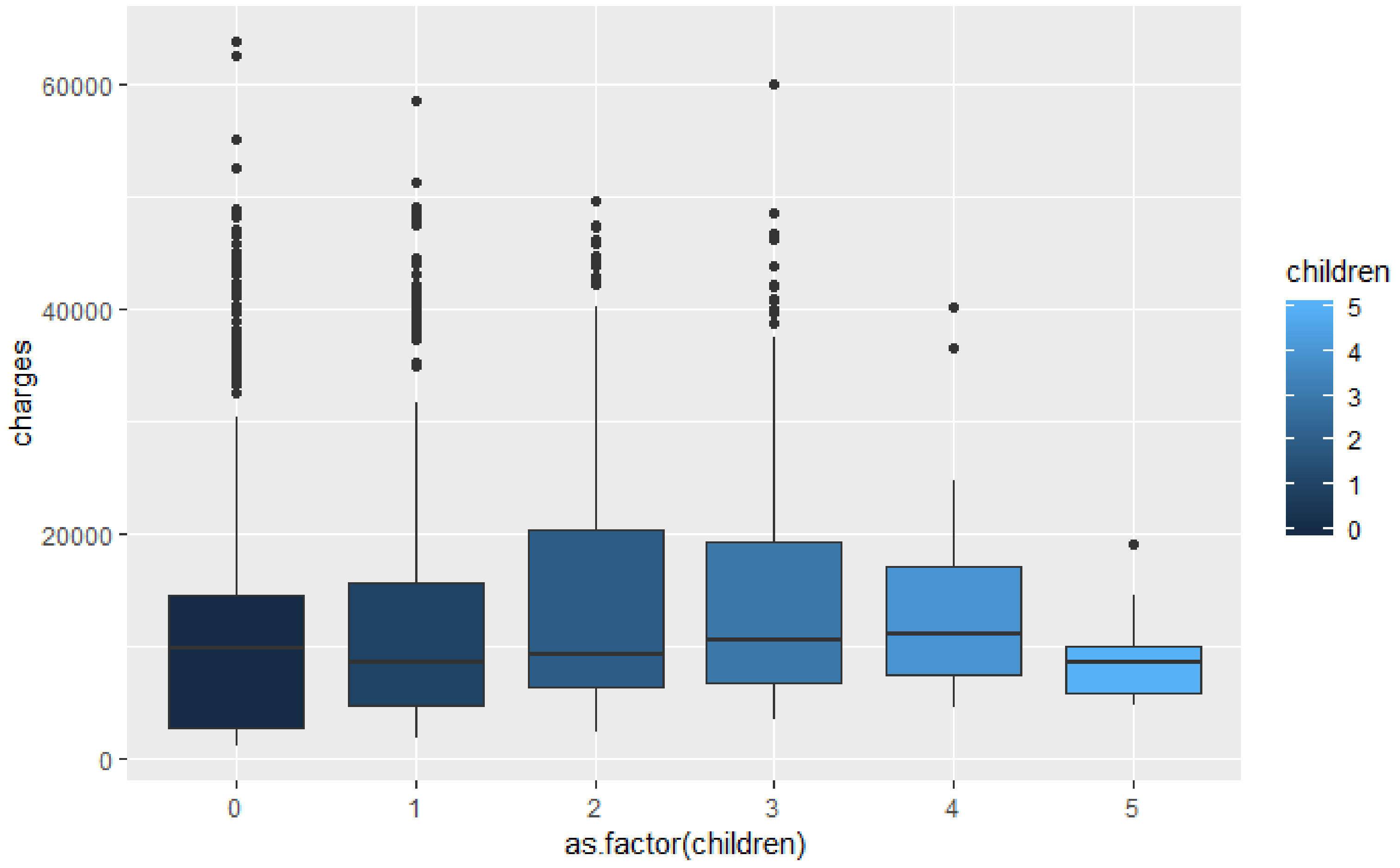
Hide

```
ggplot(US_health, aes(x = bmi, y = charges)) +
    geom_point(aes(color = smoker))
```



Hide

```
ggplot(data = US_health,aes(x=as.factor(children),y=charges))+geom_boxplot(aes(fill=children))
```



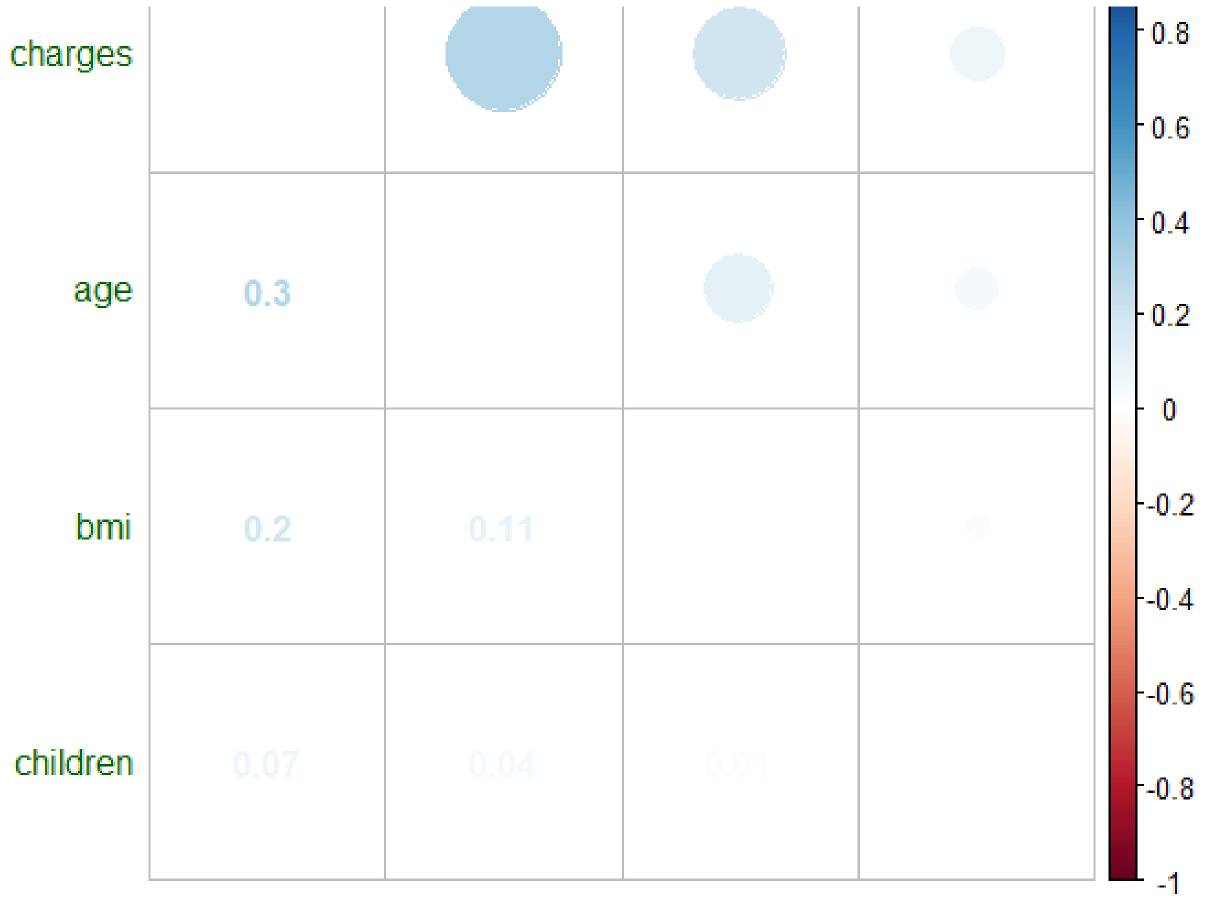
Hide

```
numeric_Var <- which(sapply(US_health, is.numeric)) #index vector numeric variables
numeric_VarNames <- names(numeric_Var) #saving names vector for use later on
cat('There are', length(numeric_Var), 'numeric variables')
```

```
There are 4 numeric variables
```

Hide

```
US_health_numVar <- US_health[, numeric_Var]
cor_health <- cor(US_health_numVar, use="pairwise.complete.obs") #correlations of all numeric variables
#sorting on decreasing correlations with SalePrice
cor_sorted_health <- as.matrix(sort(cor_health[,'charges'], decreasing = TRUE))
#selecting only high corelations
Cor_High <- names(which(apply(cor_sorted_health, 1, function(x) abs(x)>0)))
cor_health <- cor_health[Cor_High, Cor_High]
corrplot.mixed(cor_health, tl.col="dark green", tl.pos = "lt")
```



- Conversions

Hide

```
US_health$age<-as.numeric(US_health$age)
US_health$children<-as.numeric(US_health$children)
US_health$sex<-as.factor(US_health$sex)
US_health$smoker<-as.factor(US_health$smoker)
US_health$region<-as.factor(US_health$region)
```

Creating train and test data

Hide

```
set.seed(10)
Ind <- sample(2,nrow(US_health),replace= TRUE,prob=c(0.7,0.3))
train_health <- US_health[Ind==1,]
test_health <- US_health[Ind==2,]
```

Linear, Ridge and Lasso Models

Hide

```
# Linear regression
lm_health <- lm(train_health$charges~., data =train_health)
```

Hide

```
summary(lm_health)
```

```


Call:
lm(formula = train_health$charges ~ ., data = train_health)

Residuals:
     Min       1Q   Median       3Q      Max 
-11090.8  -3115.8   -951.6   1525.4  30446.4 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -12478.26    1223.49 -10.199  < 2e-16 ***
age                254.56      14.74  17.265  < 2e-16 ***
sexmale           -171.20     413.98  -0.414  0.67931    
bmi                369.80      35.36  10.458  < 2e-16 ***
children           428.85     169.92   2.524  0.01178 *  
smokeryes        23777.71     510.49  46.578  < 2e-16 ***
regionnorthwest   -633.69     592.50  -1.070  0.28512    
regionsoutheast  -1664.15     589.47  -2.823  0.00486 ** 
regionsouthwest  -1027.20     598.50  -1.716  0.08645 .  
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 6244 on 913 degrees of freedom
Multiple R-squared:  0.743, Adjusted R-squared:  0.7408 
F-statistic:   330 on 8 and 913 DF,  p-value: < 2.2e-16
```

Hide

```
#making predictions
predict_lm <-predict(lm_health, test_health, type = "response")
residual <- test_health$charges - predict_lm
predict_linreg <- data.frame("Predicted"= predict_lm, "Actual" =test_health$charges, "Residuals" = residual)
accuracy(predict_lm, test_health$charges)
```

```
               ME     RMSE      MAE       MPE     MAPE
Test set 104.9532 5665.485 4012.258 -14.62795 43.18357
```

Hide

```
x <- model.matrix(US_health$charges~., US_health)[,-1]
y <- US_health$charges
lambda_val <-10^seq(10, -2, length = 100)
```

Hide

```
set.seed(489)
train_new = sample(1:nrow(x), nrow(x)/2)
test_new = (-train_new)
y_test = y[test_new]
```

- Fitting the models

Hide

```
linear_RegModel <- lm(US_health$charges~., data= US_health)
coef(linear_RegModel)
```

```
    (Intercept)             age         sexmale             bmi        children       smokeryes regionnorthwest 
    -11938.5386        256.8564       -131.3144        339.1935        475.5005      23848.5345       -352.9639 
regionsoutheast regionsouthwest 
     -1035.0220       -960.0510 
```

Hide

```
# Ridge model
ridge_model <- glmnet(x, y, alpha = 0, lambda = lambda_val)
coef(ridge_model)
```

```
9 x 100 sparse Matrix of class "dgCMatrix"
```

```
   [[ suppressing 100 column names <U+393C><U+3E31>s0<U+393C><U+3E32>, <U+393C><U+3E31>s1<U+393C><U+3E32>, <U+393C><U+3E31>s
2<U+393C><U+3E32> ... ]]
```

```
(Intercept)      1.327039e+04  1.327038e+04  1.327036e+04  1.327034e+04  1.327032e+04  1.327028e+04
age              3.119854e-04  4.124261e-04  5.452028e-04  7.207255e-04  9.527558e-04  1.259486e-03
sexmale          1.679238e-03  2.219849e-03  2.934507e-03  3.879241e-03  5.128119e-03  6.779057e-03
bmi              4.768017e-04  6.303034e-04  8.332235e-04  1.101471e-03  1.456079e-03  1.924847e-03
children         8.269116e-04  1.093128e-03  1.445050e-03  1.910269e-03  2.525260e-03  3.338241e-03
smokeryes        2.858823e-02  3.779195e-02  4.995871e-02  6.604242e-02  8.730412e-02  1.154108e-01
regionnorthwest -1.363637e-03 -1.802645e-03 -2.382988e-03 -3.150164e-03 -4.164321e-03 -5.504971e-03
regionsoutheast  2.436197e-03  3.220505e-03  4.257312e-03  5.627906e-03  7.439742e-03  9.834869e-03
regionsouthwest -1.476582e-03 -1.951953e-03 -2.580365e-03 -3.411086e-03 -4.509248e-03 -5.960947e-03

(Intercept)      1.327024e+04  1.327018e+04  1.327010e+04  1.327000e+04  1.326986e+04  1.326968e+04
age              1.664963e-03  2.200979e-03  2.909556e-03  3.846248e-03  5.084489e-03  6.721352e-03
sexmale          8.961489e-03  1.184652e-02  1.566032e-02  2.070188e-02  2.736642e-02  3.617638e-02
bmi              2.544530e-03  3.363710e-03  4.446610e-03  5.878130e-03  7.770493e-03  1.027205e-02
children         4.412949e-03  5.833643e-03  7.711707e-03  1.019438e-02  1.347628e-02  1.781471e-02

smokeryes        1.525660e-01  2.016829e-01  2.666121e-01  3.524443e-01  4.659083e-01  6.158994e-01
regionnorthwest -7.277217e-03 -9.620001e-03 -1.271698e-02 -1.681095e-02 -2.222280e-02 -2.937675e-02
regionsoutheast  1.300106e-02  1.718654e-02  2.271942e-02  3.003344e-02  3.970192e-02  5.248269e-02
regionsouthwest -7.879997e-03 -1.041685e-02 -1.377039e-02 -1.820352e-02 -2.406377e-02 -3.181050e-02

(Intercept)      1.326944e+04  1.326913e+04  1.326871e+04  1.326816e+04  1.326743e+04  1.326646e+04
age              8.885157e-03  1.174552e-02  1.552666e-02  2.052493e-02  2.713205e-02  3.586576e-02
sexmale          4.782230e-02  6.321697e-02  8.356685e-02  1.104665e-01  1.460233e-01  1.930222e-01
bmi              1.357890e-02  1.795025e-02  2.372872e-02  3.136719e-02  4.146421e-02  5.481084e-02
children         2.354975e-02  3.113094e-02  4.115251e-02  5.439986e-02  7.191108e-02  9.505814e-02
smokeryes        8.141759e-01  1.076281e+00  1.422759e+00  1.880767e+00  2.486199e+00  3.286496e+00
regionnorthwest -3.883349e-02 -5.133410e-02 -6.785807e-02 -8.969984e-02 -1.185700e-01 -1.567287e-01
regionsoutheast  6.937743e-02  9.171010e-02  1.212305e-01  1.602511e-01  2.118276e-01  2.799976e-01
regionsouthwest -4.205094e-02 -5.558769e-02 -7.348161e-02 -9.713478e-02 -1.284002e-01 -1.697266e-01

(Intercept)      1.326519e+04  1.326351e+04  1.326128e+04 13258.3379786 13254.4500899 13249.3124272
age              4.741028e-02  6.266987e-02  8.283935e-02     0.1094973     0.1447291     0.1912886
sexmale          2.551428e-01  3.372470e-01  4.457563e-01     0.5891511     0.7786265     1.0289552
bmi              7.245252e-02  9.577065e-02  1.265904e-01     0.1673227     0.2211517     0.2922813
children         1.256541e-01  1.660949e-01  2.195460e-01     0.2901889     0.3835463     0.5069097
smokeryes        4.344359e+00  5.742645e+00  7.590841e+00    10.0336014    13.2620066    17.5284060
regionnorthwest -2.071620e-01 -2.738137e-01 -3.618918e-01    -0.4782708    -0.6320206    -0.8351008
regionsoutheast  3.700950e-01  4.891645e-01  6.465081e-01     0.8544036     1.1290484     1.4917971
regionsouthwest -2.243495e-01 -2.965434e-01 -3.919547e-01    -0.5180392    -0.6846394    -0.9047424

(Intercept)     13242.5240529 13233.5560069 13221.7129155 13206.0734859 13185.4304440 13158.196213 13122.289322
age                 0.2528115     0.3340958     0.4414681     0.5832707     0.7704837     1.017547     1.343416
sexmale             1.3596187     1.7962890     2.3715131     3.1312837     4.1328904     5.452140     7.187735
bmi                 0.3862596     0.5104046     0.6743201     0.8907604     1.1763970     1.553147     2.049720
children            0.6699023     0.8852177     1.1695710     1.5450189     2.0405305     2.694159     3.555764
smokeryes          23.1659554    30.6143089    40.4531668    53.4469986    70.6019409    93.241167   123.101644
regionnorthwest    -1.1032674    -1.4572559    -1.9240799    -2.5397732    -3.3508992    -4.418316    -5.820966
regionsoutheast     1.9707785     2.6030020     3.4367976     4.5362694     5.9845223     7.890007    10.393270
regionsouthwest    -1.1954737    -1.5793987    -2.0862193    -2.7549756    -3.6368880    -4.798992    -6.328738

(Intercept)     13074.988008 13012.745783 12930.963093 12823.711554 12683.412902 12500.485945 12262.994539
age                 1.772919     2.338475     3.082253     4.058811     5.338241     7.009754     9.185478
sexmale             9.467545    12.456108    16.363231    21.453138    28.052937    36.557977    47.429828
bmi                 2.703607     3.563578     4.692744     6.172210     8.105248    10.621757    13.882457
children            4.690487     6.183110     8.143420    10.712612    14.070635    18.444032    24.113304
smokeryes         162.458395   214.282252   282.437288   371.922933   489.162336   642.330915   841.704835
regionnorthwest    -7.660598   -10.067258   -13.205239   -17.278848   -22.536601   -29.271435   -37.812943
regionsoutheast    13.675247    17.966780    23.558774    30.811639    40.161279    52.116799    67.241940
regionsouthwest    -8.339713   -10.978607   -14.433408   -18.942552   -24.804288   -32.384688   -42.121533

(Intercept)     11956.36133 11563.25834 11063.84443 10436.58011 9659.88054 8714.82196 7588.87936 6280.36190
age                12.00398    15.63261    20.26717    26.12676   33.44096   42.42645   53.25166   65.99073
sexmale            61.18000    78.32937    99.33034   124.43852  153.52734  185.85694  219.85020  252.96212
bmi                18.08186    23.44839    30.23933    38.72771   49.17820   61.81036   76.75172   93.98643
children           31.41794    40.75598    52.57336    67.33678   85.48304  107.34055  133.02488  162.32629
smokeryes        1099.98485  1432.51321  1857.24840  2394.29966 3064.76806 3888.63531 4881.55493 6050.69209
regionnorthwest   -48.50575   -61.66639   -77.51055   -96.04658 -116.94202 -139.39357 -162.05491 -183.11601
regionsoutheast    86.10821   109.20348   136.77757   168.61107  203.71191  239.98326  273.96817  300.83254
regionsouthwest   -54.51960   -70.13100   -89.51275  -113.15484 -141.37747 -174.21011 -211.28585 -251.80762

(Intercept)     4802.54248 3186.12989  1478.6367  -260.0970 -1965.6512 -3578.10842 -5049.9191 -6350.5215
age               80.57454   96.75251   114.0831   131.9667   149.7213   166.68272   182.3005   196.2001
sexmale          281.78433  302.48564   311.5830   306.8431   287.9656   256.72668   216.5038   171.3997
bmi              113.31386  134.33142   156.4539   178.9706   201.1302   222.23197   241.7037   259.1504
```

```
children         194.62222  228.86208   263.6618   297.5080   329.0177   357.16348   381.3889   401.5927
smokeryes       7390.27714 8878.11422 10474.6259 12125.6745 13769.2223 15344.30308 16799.6853 18099.7508
regionnorthwest -200.59470 -212.84374  -219.1324  -220.0468  -217.4614  -214.02250  -212.3444  -214.2764
regionsoutheast  314.76429  309.86710   281.4145   227.0899   147.7435    47.35415   -67.7787  -190.2895
regionsouthwest -294.64493 -338.58342  -382.6653  -426.4698  -470.1649  -514.26261  -559.1882  -604.8915

(Intercept)     -7466.7598 -8401.41371 -9167.47509 -9784.484455 -10274.86693 -10659.65076 -10959.15937
age               208.1990   218.29113   226.59309   233.299398    238.64038    242.84061    246.11513
sexmale           125.3218    81.41197    41.70534     7.235238    -21.73044    -45.51947    -64.67286
bmi               274.3772   287.33732   298.14103   306.984003    314.09328    319.76338    324.22789
children          418.0355   431.13606   441.43086   449.436892    455.59573    460.34268    463.97570
smokeryes       19226.3817 20177.32061 20962.21848 21598.183619  22105.79019  22506.07587  22818.74744
regionnorthwest  -221.4758  -231.93110  -245.15889  -259.759219   -273.03313   -286.89587   -299.37951
regionsoutheast  -313.6773  -430.85248  -537.91813  -632.333475   -712.06061   -779.55490   -834.83818
regionsouthwest  -651.1356  -696.03828  -738.51354  -777.352941   -810.94950   -840.51147   -865.34297

(Intercept)     -11190.71705 -11368.80675 -11505.2282 -11609.7400 -11689.0845 -11749.4027 -11795.2145
age                248.64992    250.60136    252.0974    253.2430    254.1142    254.7769    255.2804
sexmale            -79.86839    -91.78835   -101.0581   -108.1978   -113.7048   -117.9232   -121.1441
bmi                327.71268    330.41397    332.4964    334.0786    335.3032    336.2391    336.9520
children           466.74827    468.85984    470.4655    471.6675    472.5949    473.2997    473.8342
smokeryes        23061.16918  23248.04042  23391.4488  23501.1325  23584.7922  23648.4783  23696.8861
regionnorthwest   -310.15567   -319.17689   -326.5525   -331.2598   -336.0847   -339.9782   -343.0190
regionsoutheast   -879.34796   -914.70014   -942.4755   -963.1377   -979.9854   -993.0471  -1003.0750
regionsouthwest   -885.74495   -902.21370   -915.3187   -924.9009   -933.0189   -939.3750   -944.2809

(Intercept)     -11829.9759 -11856.3328 -11876.3059 -11891.4349 -11902.8910 -11911.6778 -11918.2301 -11923.1768
age                255.6625    255.9522    256.1719    256.3382    256.4642    256.5612    256.6331    256.6875
sexmale           -123.5978   -125.4637   -126.8809   -127.9562   -128.7715   -129.3659   -129.8386   -130.1973
bmi                337.4941    337.9058    338.2181    338.4549    338.6344    338.7666    338.8696    338.9482
children           474.2391    474.5458    474.7780    474.9538    475.0869    475.1834    475.2598    475.3182
smokeryes        23733.6381  23761.5163  23782.6492  23798.6608  23810.7874  23819.9703  23826.9207  23832.1809
regionnorthwest   -345.3705   -347.1790   -348.5643   -349.6221   -350.4281   -350.7814   -351.2510   -351.6529
regionsoutheast  -1010.7430  -1016.5910  -1021.0422  -1024.4253  -1026.9936  -1028.7303  -1030.2104  -1031.3687
regionsouthwest   -948.0466   -950.9268   -953.1239   -954.7965   -956.0678   -956.8744   -957.6090   -958.1937

(Intercept)     -11926.9172 -11929.7467 -11931.8874 -11933.5069 -11934.7321 -11935.6591 -11936.3603 -11936.8907
age                256.7286    256.7597    256.7832    256.8010    256.8145    256.8247    256.8324    256.8382
sexmale           -130.4690   -130.6747   -130.8304   -130.9482   -131.0373   -131.1048   -131.1558   -131.1944
bmi                339.0078    339.0530    339.0872    339.1131    339.1326    339.1474    339.1586    339.1671
children           475.3625    475.3961    475.4216    475.4408    475.4553    475.4663    475.4747    475.4810
smokeryes        23836.1615  23839.1736  23841.4527  23843.1770  23844.4815  23845.4684  23846.2151  23846.7799
regionnorthwest   -351.9680   -352.2092   -352.3924   -352.5313   -352.6365   -352.7162   -352.7765   -352.8221
regionsoutheast  -1032.2543  -1032.9269  -1033.4365  -1033.8223  -1034.1143  -1034.3353  -1034.5025  -1034.6290
regionsouthwest   -958.6431   -958.9849   -959.2440   -959.4403   -959.5889   -959.7014   -959.7865   -959.8509

(Intercept)     -11937.2920 -11937.5956 -11937.8252 -11937.9990 -11938.1304 -11938.2298 -11938.3050 -11938.3619
age                256.8426    256.8460    256.8485    256.8504    256.8519    256.8530    256.8538    256.8544
sexmale           -131.2236   -131.2457   -131.2624   -131.2751   -131.2846   -131.2919   -131.2974   -131.3015
bmi                339.1735    339.1784    339.1821    339.1848    339.1869    339.1885    339.1897    339.1906
children           475.4857    475.4893    475.4921    475.4941    475.4957    475.4969    475.4978    475.4984
smokeryes        23847.2072  23847.5304  23847.7750  23847.9600  23848.0999  23848.2057  23848.2858  23848.3464
regionnorthwest   -352.8566   -352.8827   -352.9025   -352.9174   -352.9288   -352.9373   -352.9438   -352.9487
regionsoutheast  -1034.7247  -1034.7971  -1034.8519  -1034.8933  -1034.9247  -1034.9484  -1034.9663  -1034.9799
regionsouthwest   -959.8996   -959.9365   -959.9643   -959.9854   -960.0014   -960.0135   -960.0226   -960.0295

(Intercept)     -11938.4049 -11938.4375 -11938.4621 -11938.4807 -11938.4948 -11938.5055 -11938.5135 -11938.5196
age                256.8549    256.8552    256.8555    256.8557    256.8559    256.8560    256.8561    256.8561
sexmale           -131.3046   -131.3070   -131.3088   -131.3101   -131.3112   -131.3119   -131.3125   -131.3130
bmi                339.1913    339.1918    339.1922    339.1925    339.1928    339.1929    339.1931    339.1932
children           475.4990    475.4993    475.4996    475.4999    475.5000    475.5002    475.5002    475.5003
smokeryes        23848.3922  23848.4269  23848.4531  23848.4729  23848.4879  23848.4993  23848.5079  23848.5144
regionnorthwest   -352.9524   -352.9552   -352.9573   -352.9589   -352.9601   -352.9610   -352.9617   -352.9623
regionsoutheast  -1034.9902  -1034.9979  -1035.0038  -1035.0082  -1035.0116  -1035.0141  -1035.0161  -1035.0175
regionsouthwest   -960.0348   -960.0387   -960.0417   -960.0440   -960.0457   -960.0470   -960.0479   -960.0487
```

- Here, the differences came out to be nominal. Hence, using ridge in order to check whether it can improve the OLS estimation.

Hide

```
# Creating linear and ridge models based on x and y train/test split
linear_RegModel <- lm(US_health$charges~., data = US_health, subset = train_new)
ridge_model <- glmnet(x[train_new,], y[train_new], alpha = 0, lambda = lambda_val)
#Finding out best lambda value from the list via cross-validation.
CV_out <- cv.glmnet(x[train_new,], y[train_new], alpha =0)
lambda_best <- CV_out$lambda.min
```

Hide

```
# Making predictions based on best lambda
# Linear model prediction
lm_predict <- predict(linear_RegModel, newdata = US_health[test_new,])
# Ridge model prediction
ridge_predict <- predict(ridge_model, s= lambda_best, newx = x[test_new,])
```

Hide

```
# looking for coefficients
out_ridge <-glmnet(x[train_new,], y[train_new], alpha = 0)
predict(ridge_model, type = "coefficients", s= lambda_best)[1:6,]
```

```
(Intercept)         age     sexmale         bmi    children   smokeryes 
 -9585.4564    216.4560    584.9393    318.8245    492.4345  20305.3594 
```

- MOst of the coefficients estimates are coming out to be conservative.
- So, trying to fit lasso as it takes absolute values for coefficients estimates.

Hide

```
#Lasso Model
lasso_model <- glmnet(x[train_new,],y[train_new], alpha = 1, lambda = lambda_val)
lasso_predict <- predict(lasso_model, s= lambda_best, newx = x[test_new,])
```

Hide

```
lasso_coef <- predict(lasso_model, type= "coefficients", s= lambda_best)[1:6]
lasso_coef
```

```
[1] -3651.569   171.284     0.000   190.056     0.000 19538.071
```

- As expected, lasso set the value to 0 for unimportant variables and set the values high for the important ones, i.e., age, bmi, smoker. So, lasso model can be chosen. Let’s, confirm it with MSE values.

Selection of best model among linear, ridge and lasso

Hide

```
#Checking MSE for linear model
mean((lm_predict-y_test)^2)
```

```
[1] 36825120
```

Hide

```
#Checking MSE for ridge model
mean((ridge_predict-y_test)^2)
```

```
[1] 39616462
```

Hide

```
#Checking MSE for Lasso Model
mean((lasso_predict-y_test)^2)
```

```
[1] 43500867
```

- Since, linear regression model has the least MSE, therefore, it is the best model among the three models.

SELECTION OF BEST PREDICTORS BASED ON LINEAR REGRESSION MODEL

Hide

```
summary(linear_RegModel)
```

```


Call:
lm(formula = US_health$charges ~ ., data = US_health, subset = train_new)

Residuals:
   Min     1Q Median     3Q    Max 
 -9872  -3015  -1107   1305  31561 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11544.86    1416.98  -8.148 1.86e-15 ***
age                236.65      17.42  13.584  < 2e-16 ***
sexmale            477.06     483.77   0.986   0.3244    
bmi                354.18      41.79   8.476  < 2e-16 ***
children           500.36     201.48   2.483   0.0133 *  
smokeryes        22107.37     583.83  37.866  < 2e-16 ***
regionnorthwest  -1101.03     688.29  -1.600   0.1102    
regionsoutheast  -1409.71     686.01  -2.055   0.0403 *  
regionsouthwest  -1363.39     686.14  -1.987   0.0473 *  
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 6190 on 660 degrees of freedom
Multiple R-squared:  0.7154,	Adjusted R-squared:  0.7119 
F-statistic: 207.4 on 8 and 660 DF,  p-value: < 2.2e-16
```

- From the above summary(), the equation came out to be:
- y= -11544.86 +(236.65*age) +(477.06*sexmale) +(354.18*bmi) + (500.36*children) +(22107.37*smokeryes) -(1101.03*regionnorthwest) -(1409.71*regionsoutheast) -(1363.39*regionsouthwest)

Anova

Hide

```
Anova(linear_RegModel)
```

```
Anova Table (Type II tests)

Response: US_health$charges
              Sum Sq  Df   F value  Pr(>F)    
age       7.0698e+09   1  184.5342 < 2e-16 ***
sex       3.7256e+07   1    0.9724 0.32443    
bmi       2.7522e+09   1   71.8364 < 2e-16 ***
children  2.3629e+08   1    6.1675 0.01326 *  
smoker    5.4934e+10   1 1433.8560 < 2e-16 ***
region    2.1251e+08   3    1.8489 0.13698    
Residuals 2.5286e+10 660                      
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

All possible subset regression

Hide

```
All_subset <- lm(linear_RegModel, data = US_health)
ols_step_best_subset(All_subset)
```

```
             Best Subsets Regression
-------------------------------------------------
Model Index    Predictors
-------------------------------------------------
     1         smoker
     2         age smoker
     3         age bmi smoker
     4         age bmi children smoker
     5         age bmi children smoker region
     6         age sex bmi children smoker region
-------------------------------------------------


                                                               Subsets Regression Summary
-----------------------------------------------------------------------------------------------------------------------------
-----------------------------
                       Adj.        Pred
Model    R-Square    R-Square    R-Square      C(p)         AIC           SBIC          SBC             MSEP              FPE
HSP         APC
-----------------------------------------------------------------------------------------------------------------------------
-----------------------------
  1        0.6198      0.6195       0.618    611.9889    27667.4636    23868.9391    27683.0604    55887669.4102    5588754
4.4451    41800.8466    0.3814
  2        0.7214      0.7210      0.7196     93.8304    27253.3244    23455.8677    27274.1201    41010413.6993    4101018
4.4505    30673.4926    0.2799
  3        0.7475      0.7469      0.7455    -37.6245    27123.8359    23327.0363    27149.8306    37227691.7503    3722731
7.1643    27844.2284    0.2540
  4        0.7497      0.7489      0.7474    -46.9727    27114.0352    23317.3965    27145.2288    36956206.6446    3695562
8.2045    27641.1727    0.2522
  5        0.7509      0.7496      0.7475    -51.0596    27113.6624    23313.1716    27160.4528    36891265.6918    3689044
0.8009    27592.6006    0.2514
  6        0.7509      0.7494      0.7471    -49.2088    27115.5058    23315.1156    27167.4951    36942459.0871    3694134
3.9390    27630.8903    0.2517
-----------------------------------------------------------------------------------------------------------------------------
-----------------------------
AIC: Akaike Information Criteria
 SBIC: Sawa's Bayesian Information Criteria
 SBC: Schwarz Bayesian Criteria
 MSEP: Estimated error of prediction, assuming multivariate normality
 FPE: Final Prediction Error
 HSP: Hocking's Sp
 APC: Amemiya Prediction Criteria
```

Forward Regression

Hide

```
Model_ForwardReg <- step(linear_RegModel , direction = "forward")
```

```
Start:  AIC=11690.53
US_health$charges ~ age + sex + bmi + children + smoker + region
```

Hide

```
summary(Model_ForwardReg)
```

```
Call:
lm(formula = US_health$charges ~ age + sex + bmi + children + 
    smoker + region, data = US_health, subset = train_new)

Residuals:
   Min     1Q Median     3Q    Max 
 -9872  -3015  -1107   1305  31561 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11544.86    1416.98  -8.148 1.86e-15 ***
age                236.65      17.42  13.584  < 2e-16 ***
sexmale            477.06     483.77   0.986   0.3244    
bmi                354.18      41.79   8.476  < 2e-16 ***
children           500.36     201.48   2.483   0.0133 *  
smokeryes        22107.37     583.83  37.866  < 2e-16 ***
regionnorthwest  -1101.03     688.29  -1.600   0.1102    
regionsoutheast  -1409.71     686.01  -2.055   0.0403 *  
regionsouthwest  -1363.39     686.14  -1.987   0.0473 *  
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 6190 on 660 degrees of freedom
Multiple R-squared:  0.7154,	Adjusted R-squared:  0.7119 
F-statistic: 207.4 on 8 and 660 DF,  p-value: < 2.2e-16
```

Backward Regression

Hide

```
Model_BackwardReg <- step(linear_RegModel , direction = "backward")
```

```
Start:  AIC=11690.53
US_health$charges ~ age + sex + bmi + children + smoker + region

           Df  Sum of Sq        RSS   AIC
- sex       1 3.7256e+07 2.5323e+10 11690
- region    3 2.1251e+08 2.5498e+10 11690
<none>                   2.5286e+10 11690
- children  1 2.3629e+08 2.5522e+10 11695
- bmi       1 2.7522e+09 2.8038e+10 11758
- age       1 7.0698e+09 3.2356e+10 11854
- smoker    1 5.4934e+10 8.0219e+10 12461

Step:  AIC=11689.51
US_health$charges ~ age + bmi + children + smoker + region

           Df  Sum of Sq        RSS   AIC
- region    3 2.2628e+08 2.5549e+10 11690
<none>                   2.5323e+10 11690
- children  1 2.4049e+08 2.5564e+10 11694
- bmi       1 2.8350e+09 2.8158e+10 11758
- age       1 7.0387e+09 3.2362e+10 11852
- smoker    1 5.5663e+10 8.0986e+10 12465

Step:  AIC=11689.46
US_health$charges ~ age + bmi + children + smoker

           Df  Sum of Sq        RSS   AIC
<none>                   2.5549e+10 11690
- children  1 2.3349e+08 2.5783e+10 11694
- bmi       1 2.7725e+09 2.8322e+10 11756
- age       1 7.1431e+09 3.2692e+10 11852
- smoker    1 5.5639e+10 8.1189e+10 12461
```

Hide

```
summary(Model_BackwardReg)
```

```

Call:
lm(formula = US_health$charges ~ age + bmi + children + smoker, 
    data = US_health, subset = train_new)

Residuals:
   Min     1Q Median     3Q    Max 
-10032  -3246  -1064   1246  31329 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11845.57    1364.71  -8.680   <2e-16 ***
age            237.40      17.42  13.625   <2e-16 ***
bmi            338.98      39.93   8.488   <2e-16 ***
children       495.71     201.23   2.463    0.014 *  
smokeryes    22134.76     582.09  38.026   <2e-16 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 6203 on 664 degrees of freedom
Multiple R-squared:  0.7124,    Adjusted R-squared:  0.7107 
F-statistic: 411.2 on 4 and 664 DF,  p-value: < 2.2e-16
```

Stepwise Regression

Hide

```
Model_StepwiseReg <- step(linear_RegModel , direction = "both")
```

```
Start:  AIC=11690.53
US_health$charges ~ age + sex + bmi + children + smoker + region

           Df  Sum of Sq        RSS   AIC
- sex       1 3.7256e+07 2.5323e+10 11690
- region    3 2.1251e+08 2.5498e+10 11690
<none>                   2.5286e+10 11690
- children  1 2.3629e+08 2.5522e+10 11695
- bmi       1 2.7522e+09 2.8038e+10 11758
- age       1 7.0698e+09 3.2356e+10 11854
- smoker    1 5.4934e+10 8.0219e+10 12461

Step:  AIC=11689.51
US_health$charges ~ age + bmi + children + smoker + region

           Df  Sum of Sq        RSS   AIC
- region    3 2.2628e+08 2.5549e+10 11690
<none>                   2.5323e+10 11690
+ sex       1 3.7256e+07 2.5286e+10 11690
- children  1 2.4049e+08 2.5564e+10 11694
- bmi       1 2.8350e+09 2.8158e+10 11758
- age       1 7.0387e+09 3.2362e+10 11852
- smoker    1 5.5663e+10 8.0986e+10 12465

Step:  AIC=11689.46
US_health$charges ~ age + bmi + children + smoker

           Df  Sum of Sq        RSS   AIC
<none>                   2.5549e+10 11690
+ region    3 2.2628e+08 2.5323e+10 11690
+ sex       1 5.1028e+07 2.5498e+10 11690
- children  1 2.3349e+08 2.5783e+10 11694
- bmi       1 2.7725e+09 2.8322e+10 11756
- age       1 7.1431e+09 3.2692e+10 11852
- smoker    1 5.5639e+10 8.1189e+10 12461
```

Hide

```
summary(Model_StepwiseReg)
```

```
Call:
lm(formula = US_health$charges ~ age + bmi + children + smoker, 
    data = US_health, subset = train_new)

Residuals:
   Min     1Q Median     3Q    Max 
-10032  -3246  -1064   1246  31329 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11845.57    1364.71  -8.680   <2e-16 ***
age            237.40      17.42  13.625   <2e-16 ***
bmi            338.98      39.93   8.488   <2e-16 ***
children       495.71     201.23   2.463    0.014 *  
smokeryes    22134.76     582.09  38.026   <2e-16 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 6203 on 664 degrees of freedom
Multiple R-squared:  0.7124,    Adjusted R-squared:  0.7107 
F-statistic: 411.2 on 4 and 664 DF,  p-value: < 2.2e-16
```

BEST MODEL ACCORDING TO ANNOVA

Hide

```
bestmodel_anova <- lm(train_health$charges~ age + bmi + smoker, data =train_health)
summary(bestmodel_anova)
```

```
Call:
lm(formula = train_health$charges ~ age + bmi + smoker, data = train_health)

Residuals:
   Min     1Q Median     3Q    Max 
-12095  -3184  -1032   1515  29136 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12037.77    1164.32  -10.34   <2e-16 ***
age            256.42      14.80   17.33   <2e-16 ***
bmi            338.80      34.05    9.95   <2e-16 ***
smokeryes    23670.40     509.11   46.49   <2e-16 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 6277 on 918 degrees of freedom
Multiple R-squared:  0.7389,    Adjusted R-squared:  0.738 
F-statistic: 865.8 on 3 and 918 DF,  p-value: < 2.2e-16
```

- From the above summary(), the equation came out to be:
- y= -12037.77 +(256.42*age) +(338.80*bmi) + (23670.40*smokeryes)

Model Diagnostic Tests for annova based fit model

Hide

```
# Checking the Multicollinearity
car::vif(bestmodel_anova)
```

```
     age      bmi   smoker 
1.012844 1.012897 1.000052 
```

Hide

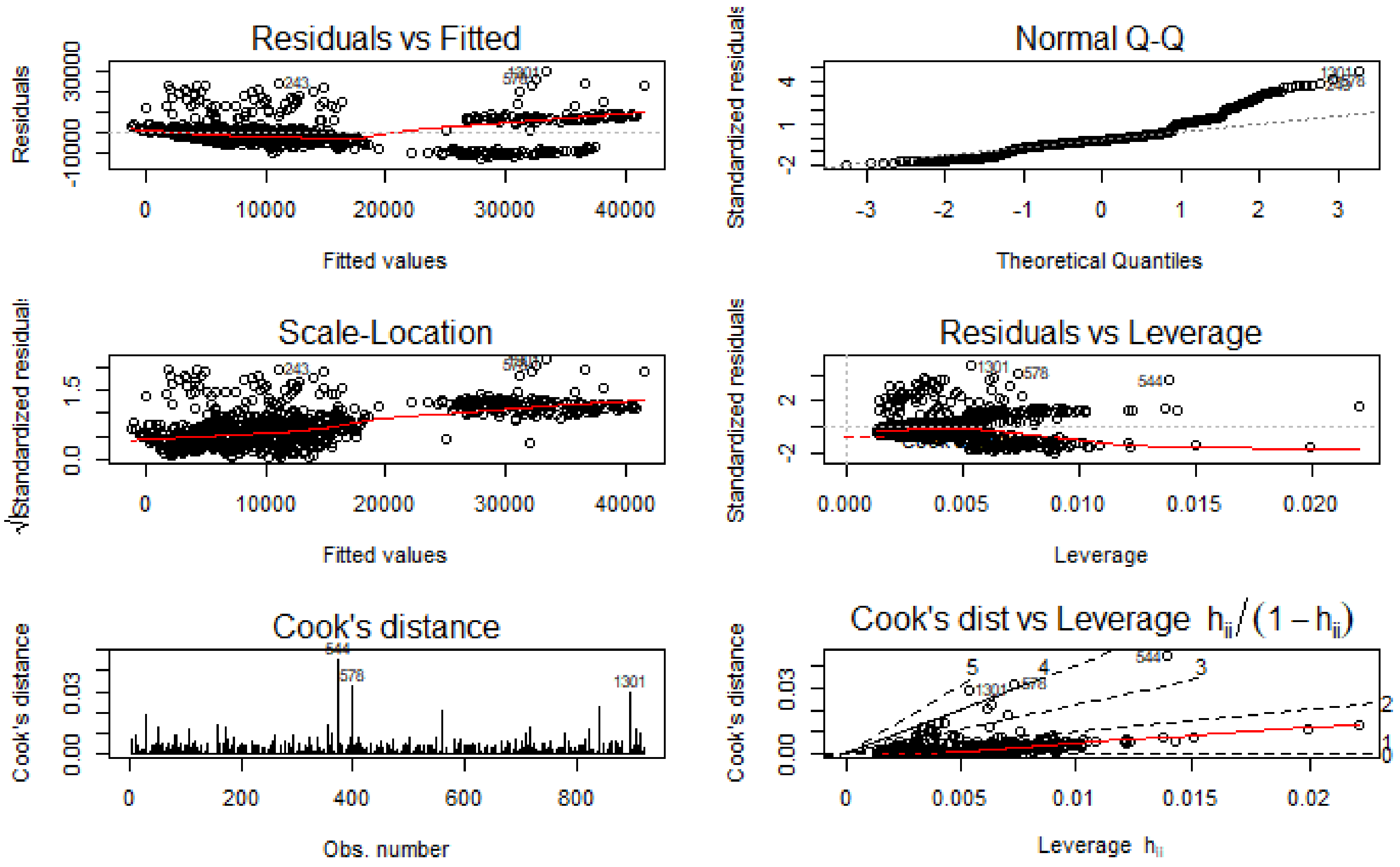
```
par(mfrow = c(3, 2))
plot(bestmodel_anova, which=1)  #Plotted Residuals v/s fitted graph
plot(bestmodel_anova, which=2) # Normality plot for residuals
```

Hide

```
plot(bestmodel_anova, which=3)  #Plotted Scale-Location graph
plot(bestmodel_anova, which=5)  #Plotted Residuals v/s Leverage graph
```

Hide

```
plot(bestmodel_anova, which=4)  #Graph for determining Cook's Distance
plot(bestmodel_anova, which=6)  #Plotted Cook's Distance-Leverage graph
```



Hide

```
# Non-constant variance score test
car::ncvTest(bestmodel_anova)
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 151.619, Df = 1, p = < 2.22e-16
```

Hide

```
# Testing for Autocorrelated errors
durbinWatsonTest(bestmodel_anova)
```

```
 lag Autocorrelation D-W Statistic p-value
   1     0.006906446       1.98215   0.778
 Alternative hypothesis: rho != 0
```

Hide

```
#Test for normal distribution
shapiro.test(bestmodel_anova$residuals)
```

```

    Shapiro-Wilk normality test

data:  bestmodel_anova$residuals
W = 0.89898, p-value < 2.2e-16
```

Predictions for annova based fit model

Hide

```
predict_Abest <-predict(bestmodel_anova, test_health, type = "response")
residual_Abest <- test_health$charges - predict_Abest
predict_Aregbest <- data.frame("Predicted"= predict_Abest, "Actual" =test_health$charges, "Residuals" = residual_Abest)
accuracy(predict_Abest, test_health$charges)
```

```
             ME     RMSE      MAE      MPE     MAPE
Test set 60.4786 5667.741 4005.492 -19.5216 45.08702
```

BEST MODEL ACCORDING TO STEPWISE AND BACKWARD

Hide

```
bestmodel_StepBAck <- lm(train_health$charges~ age + bmi + smoker +children, data =train_health)
summary(bestmodel_StepBAck)
```

```


Call:
lm(formula = train_health$charges ~ age + bmi + smoker + children, 
    data = train_health)

Residuals:
   Min     1Q Median     3Q    Max 
-11588  -3122  -1026   1491  29616 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12565.34    1180.38 -10.645   <2e-16 ***
age            255.28      14.76  17.292   <2e-16 ***
bmi            341.87      33.98  10.062   <2e-16 ***
smokeryes    23675.65     507.69  46.634   <2e-16 ***
children       421.95     170.06   2.481   0.0133 *  
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 6259 on 917 degrees of freedom
Multiple R-squared:  0.7406,	Adjusted R-squared:  0.7395 
F-statistic: 654.6 on 4 and 917 DF,  p-value: < 2.2e-16
```

- From the above summary(), the equation came out to be:
- y= -12565.34 +(255.28*age) +(341.87*bmi) + (23675.65*smokeryes) +(421.95*children)

Model Diagnostic Tests for stepwise and backward based fit model

Hide

```
# Checking the Multicollinearity
car::vif(bestmodel_StepBAck)
```

```
     age      bmi   smoker children 
1.013816 1.014245 1.000070 1.002078 
```

Hide

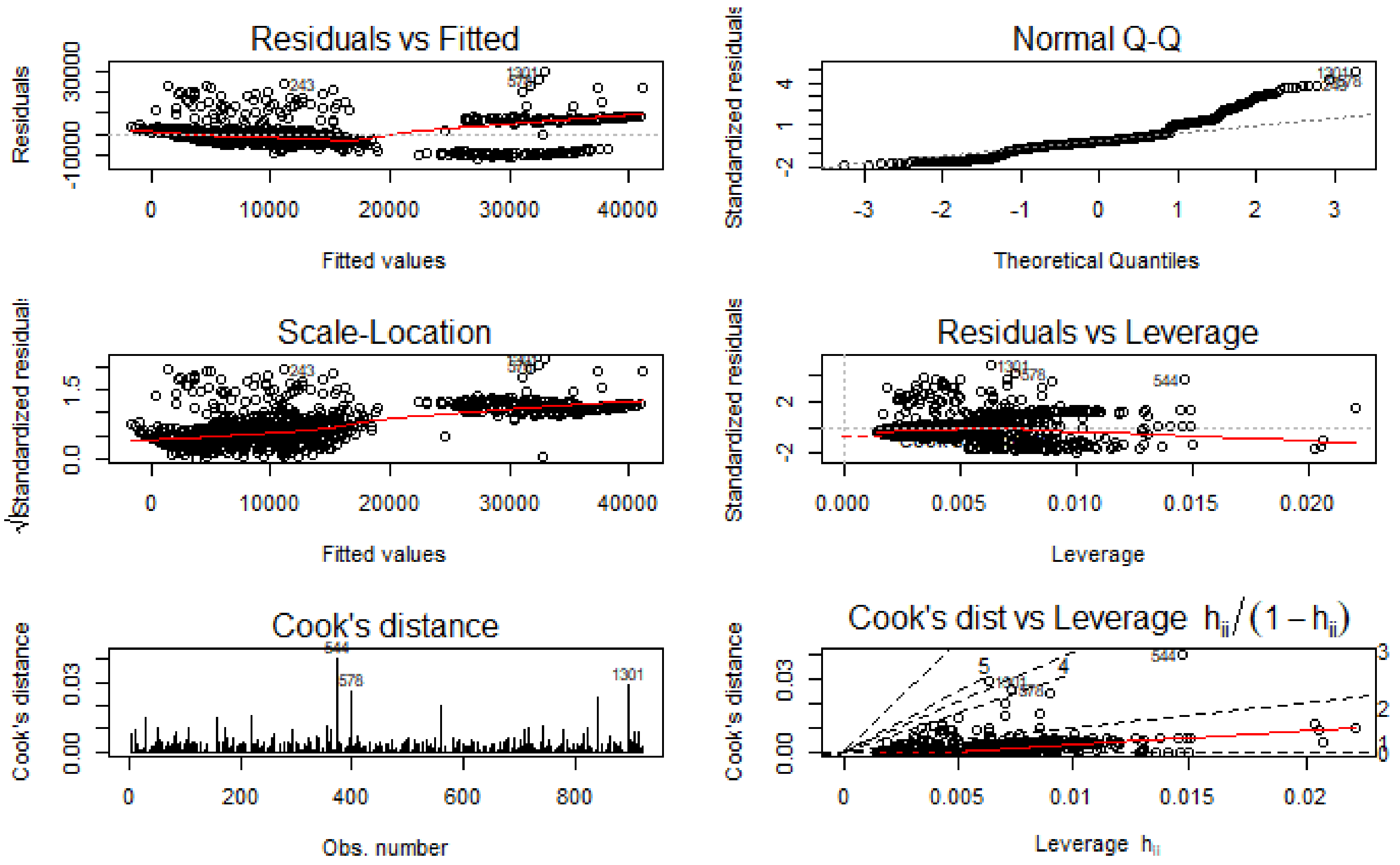
```
par(mfrow = c(3, 2))
plot(bestmodel_StepBAck, which=1)  #Plotted Residuals v/s fitted graph
plot(bestmodel_StepBAck, which=2) # Normality plot for residuals
```

Hide

```
plot(bestmodel_StepBAck, which=3)  #Plotted Scale-Location graph
plot(bestmodel_StepBAck, which=5)  #Plotted Residuals v/s Leverage graph
```

Hide

```
plot(bestmodel_StepBAck, which=4)  #Graph for determining Cook's Distance
plot(bestmodel_StepBAck, which=6)
```



Hide

```
# Non-constant variance score test
car::ncvTest(bestmodel_StepBAck)
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 156.9308, Df = 1, p = < 2.22e-16
```

Hide

```
# Testing for Autocorrelated errors
durbinWatsonTest(bestmodel_StepBAck)
```

```
 lag Autocorrelation D-W Statistic p-value
   1     0.002167906      1.992042   0.884
 Alternative hypothesis: rho != 0
```

Hide

```
#Test for normal distribution
shapiro.test(bestmodel_StepBAck$residuals)
```

```

    Shapiro-Wilk normality test

data:  bestmodel_StepBAck$residuals
W = 0.89561, p-value < 2.2e-16
```

Prediction for stepwise and backward based fit model

Hide

```
predict_SBbest <-predict(bestmodel_StepBAck, test_health, type = "response")
residual_SBbest <- test_health$charges - predict_SBbest
predict_SBregbest <- data.frame("Predicted"= predict_SBbest, "Actual" =test_health$charges, "Residuals" = residual_SBbest)
accuracy(predict_SBbest, test_health$charges)
```

```
                ME     RMSE      MAE      MPE     MAPE
Test set 107.7915 5628.947 3972.476 -16.4929 43.83209
```

BEST MODEL ACCORDING TO FORWARD AND ALL POSSIBLE SUBSET REGRESSION

Hide

```
bestmodel_fwd <- lm(train_health$charges~ age + bmi + smoker +children + region, data =train_health)
summary(bestmodel_fwd)
```

```

Call:
lm(formula = train_health$charges ~ age + bmi + smoker + children + 
    region, data = train_health)

Residuals:
     Min       1Q   Median       3Q      Max 
-11165.7  -3111.7   -979.4   1507.6  30377.4 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -12546.15    1211.88 -10.353  < 2e-16 ***
age                254.74      14.73  17.294  < 2e-16 ***
bmi                369.25      35.32  10.454  < 2e-16 ***
smokeryes        23756.57     507.70  46.793  < 2e-16 ***
children           428.42     169.84   2.522  0.01182 *  
regionnorthwest   -638.41     592.12  -1.078  0.28124    
regionsoutheast  -1668.87     589.10  -2.833  0.00471 ** 
regionsouthwest  -1030.88     598.17  -1.723  0.08516 .  
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 6241 on 914 degrees of freedom
Multiple R-squared:  0.743,	Adjusted R-squared:  0.741 
F-statistic: 377.4 on 7 and 914 DF,  p-value: < 2.2e-16
```

- From the above summary(), the equation came out to be:
- y= -12546.15 +(254.74*age) +(369.25*bmi) + (23756.57*smokeryes) +(428.42*children) -(638.41*regionnorthwest) -(1668.87*regionsoutheast) -(1030.88*regionsouthwest)

Model Diagnostic Tests for forward and all possible subset based fit model

Hide

```
# Checking the Multicollinearity
car::vif(bestmodel_fwd)
```

```
             GVIF Df GVIF^(1/(2*Df))
age      1.015301  1        1.007622
bmi      1.102430  1        1.049967
smoker   1.005959  1        1.002975
children 1.005435  1        1.002714
region   1.097810  3        1.015674
```

Hide

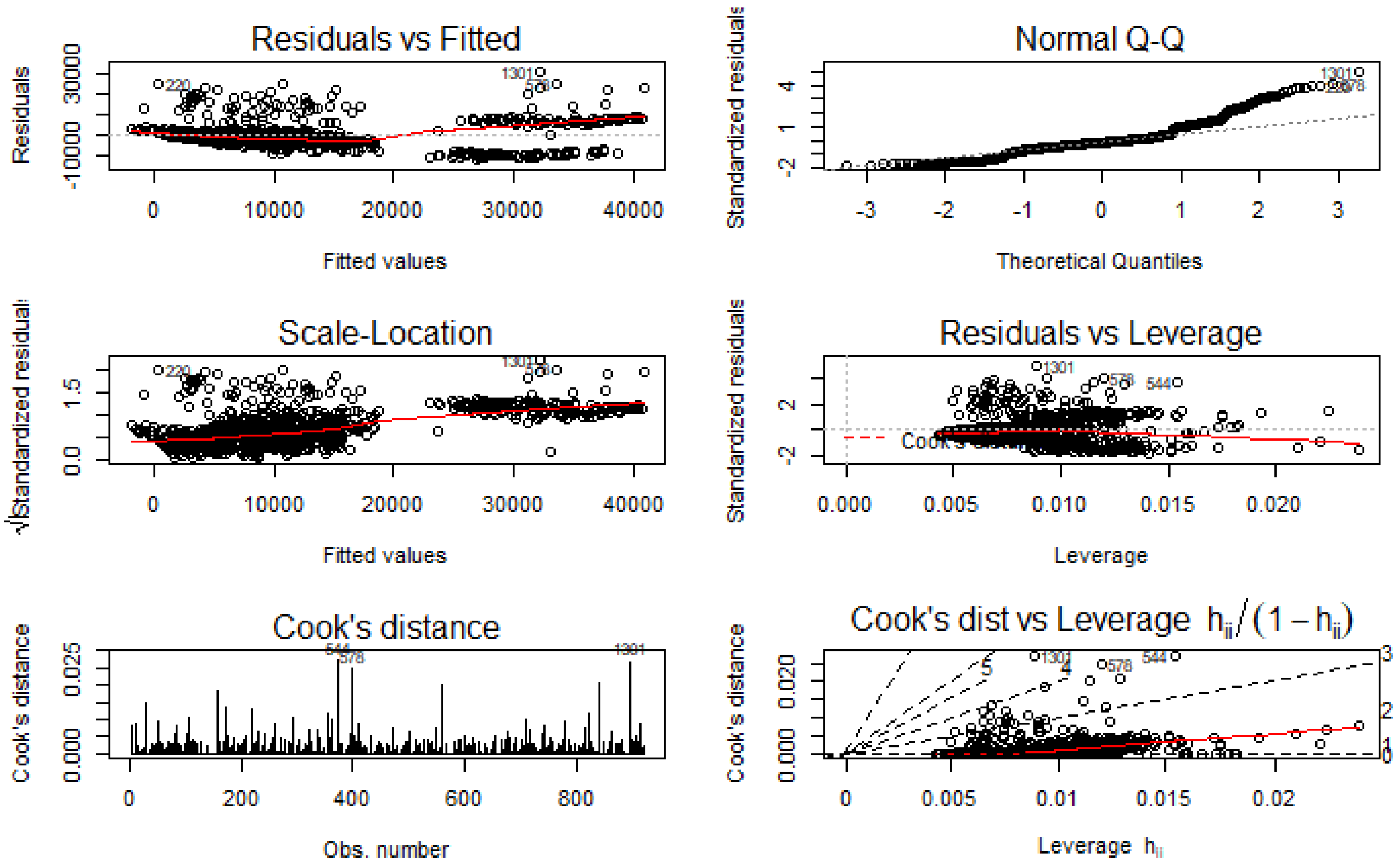
```
par(mfrow = c(3, 2))
plot(bestmodel_fwd, which=1)  #Plotted Residuals v/s fitted graph
plot(bestmodel_fwd, which=2) # Normality plot for residuals
```

Hide

```
plot(bestmodel_fwd, which=3)  #Plotted Scale-Location graph
plot(bestmodel_fwd, which=5)  #Plotted Residuals v/s Leverage graph
```

Hide

```
plot(bestmodel_fwd, which=4)  #Graph for determining Cook's Distance
plot(bestmodel_fwd, which=6)
```



Hide

```
# Non-constant variance score test
car::ncvTest(bestmodel_fwd)
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 153.8316, Df = 1, p = < 2.22e-16
```

Hide

```
# Testing for Autocorrelated errors
durbinWatsonTest(bestmodel_fwd)
```

```
 lag Autocorrelation D-W Statistic p-value
   1   -0.0008177996      1.997969   0.952
 Alternative hypothesis: rho != 0
```

Hide

```
#Test for normal distribution
shapiro.test(bestmodel_fwd$residuals)
```

```

    Shapiro-Wilk normality test

data:  bestmodel_fwd$residuals
W = 0.89578, p-value < 2.2e-16
```

Prediction for forward and all possible subset based fit model

Hide

```
predict_fwdbest <-predict(bestmodel_fwd, test_health, type = "response")
residual_fwdbest <- test_health$charges - predict_fwdbest
predict_linregbest2 <- data.frame("Predicted"= predict_fwdbest, "Actual" =test_health$charges, "Residuals" = residual_fwdbes
t)
accuracy(predict_fwdbest, test_health$charges)
```

```
               ME     RMSE      MAE       MPE    MAPE
Test set 102.2186 5665.428 4017.834 -14.77523 43.3282
```