Insurance Charge Prediction

AUTHOR
Group19

Introduction

This Dataset is an extensive compilation of health-related information aimed at offering insights into the factors that impact individual medical insurance expenses. This dataset covers diverse attributes related to individuals' demographic details, lifestyles, and medical backgrounds. The main emphasis is on gaining a deeper understanding of the economic aspects of healthcare, with a specific focus on the medical costs incurred by individuals.In this dataset, we have 7 variables/columns in total.

```
setwd("/Users/gaoyuchen/Downloads")
data <- read.csv("insurance.csv")
ncol(data)
```

```
[1] 7
```

1.Age: This is the age of the individual.
2.Sex: This refers to the gender of the individual.This categorical variable takes two values female and male.
3.BMI (Body Mass Index): BMI is a measure of body fat based on height and weight.
4.Children: This indicates the number of children or dependents covered by the insurance.
5.Smoker: This binary feature likely indicates whether the individual is a smoker or not.
6.Region:This categorical variable represents the geographical region where the individual resides.There are four unique inputs, namely "southwest" "southeast" "northwest"and "northeast".
7.Charges: This is likely the target variable, representing the medical costs incurred by the individual.
We have two categorical variables sex(with two unique inputs) and region(with four unique inputs) and one binary variable smoker.

```
unique(data$sex)
```

```
[1] "female" "male"
```

```
unique(data$region)
```

```
[1] "southwest" "southeast" "northwest" "northeast"
```

```
unique(data$smoker)
```

```
[1] "yes" "no"
```

We have 1338 rows in this dataset.

```
nrow(data)
```

```
[1] 1338
```

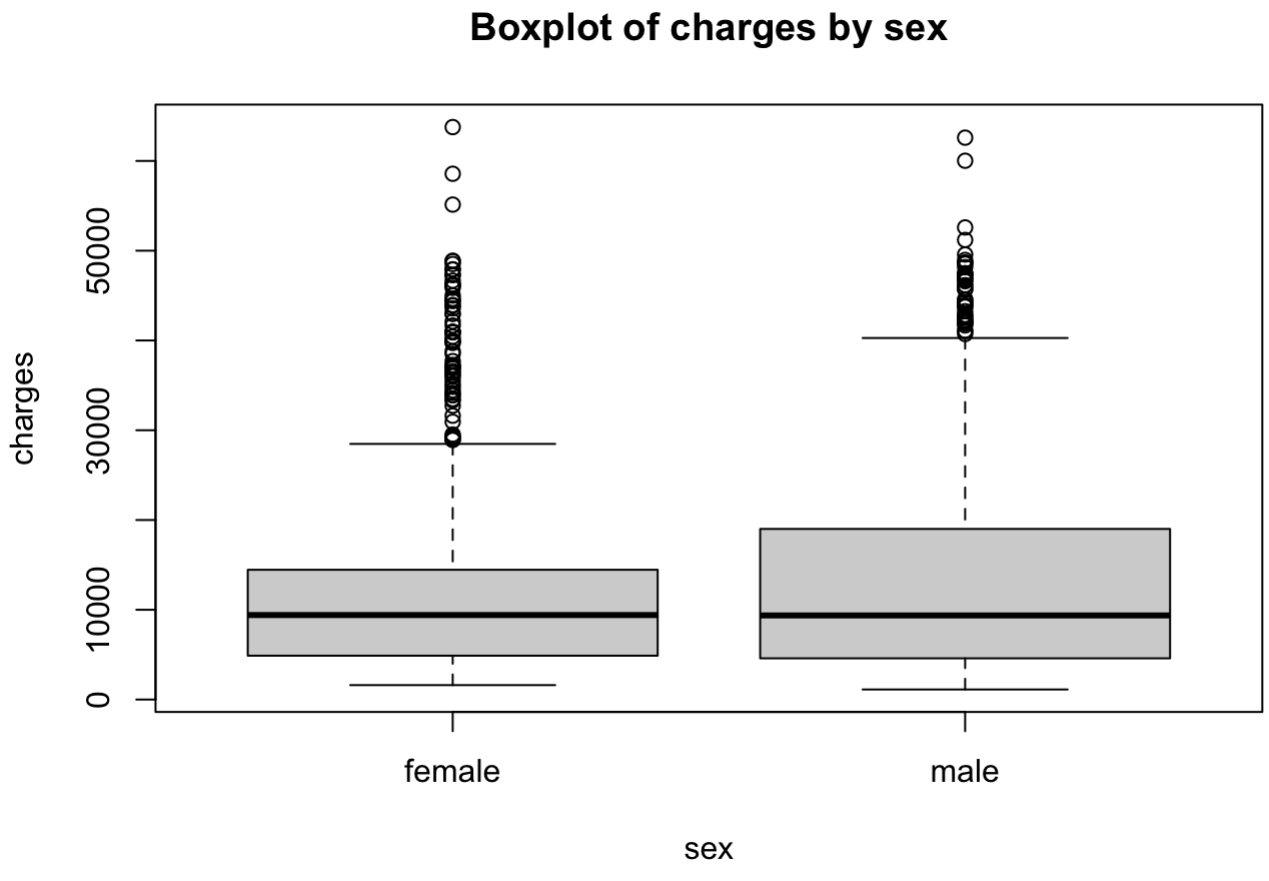
In the health insurance sector, insurance firms frequently encounter difficulties when accurately calculating premiums for individual policyholders. Inaccuracies in assessing health risks can lead to substantial financial setbacks. Therefore, precise determination of health insurance premiums is essential for ensuring the financial stability of insurance companies and delivering equitable services to policyholders. This dataset offers a substantial volume of data encompassing various facets, providing valuable insights that can assist health insurance companies in shaping their operational and business strategies.

Research Questions

The first hypothesis I would like to make is that individuals who smoke are likely to incur higher medical charges in comparison to those who do not smoke.
The second hypothesis I would like to make is that older individuals are more prone to incurring higher medical charges compared to their younger counterparts.
The last hypothesis is that people with higher BMI are likely to incur higher medical charges.

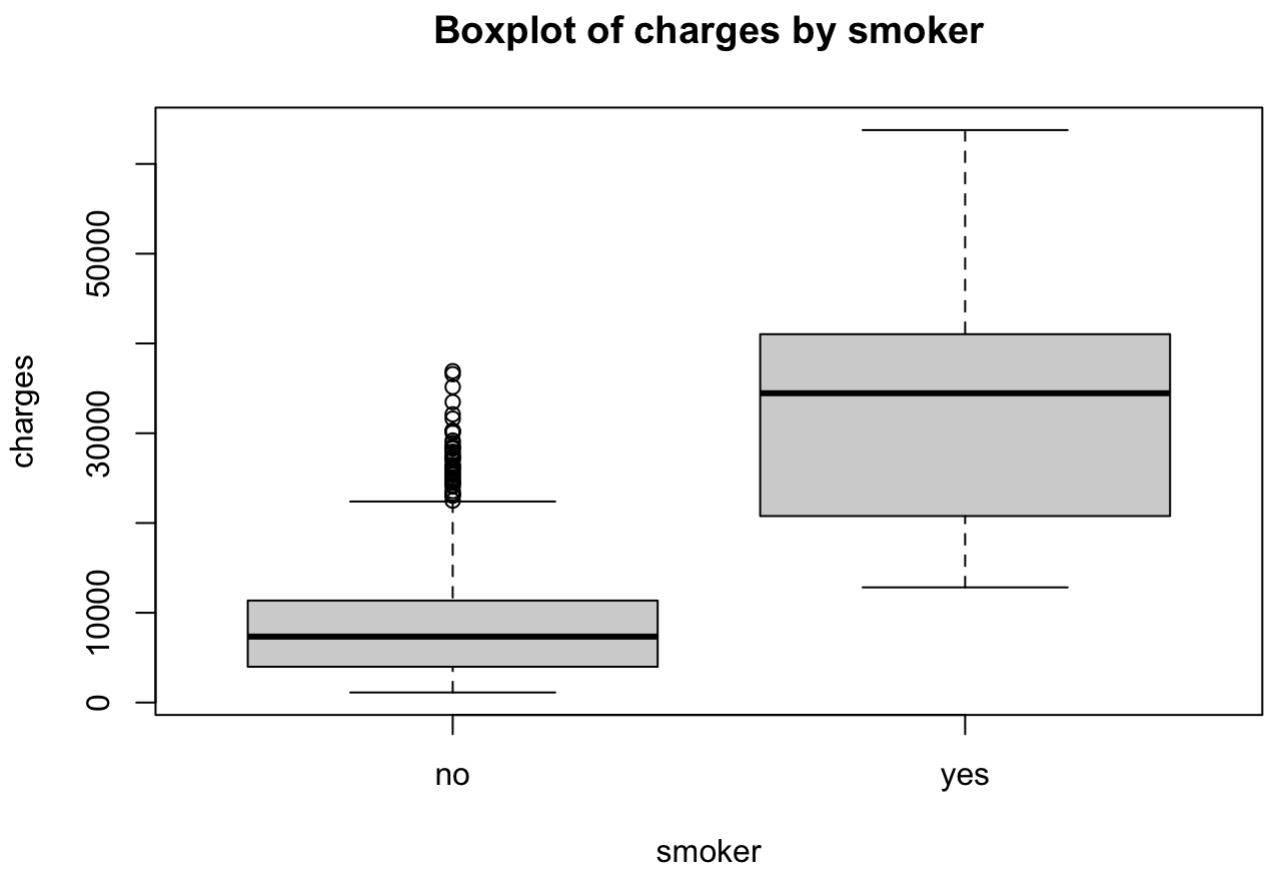
Data Exploration

```
boxplot(charges ~ sex, data = data,
        xlab = "sex", ylab = "charges",
        main = "Boxplot of charges by sex")
```



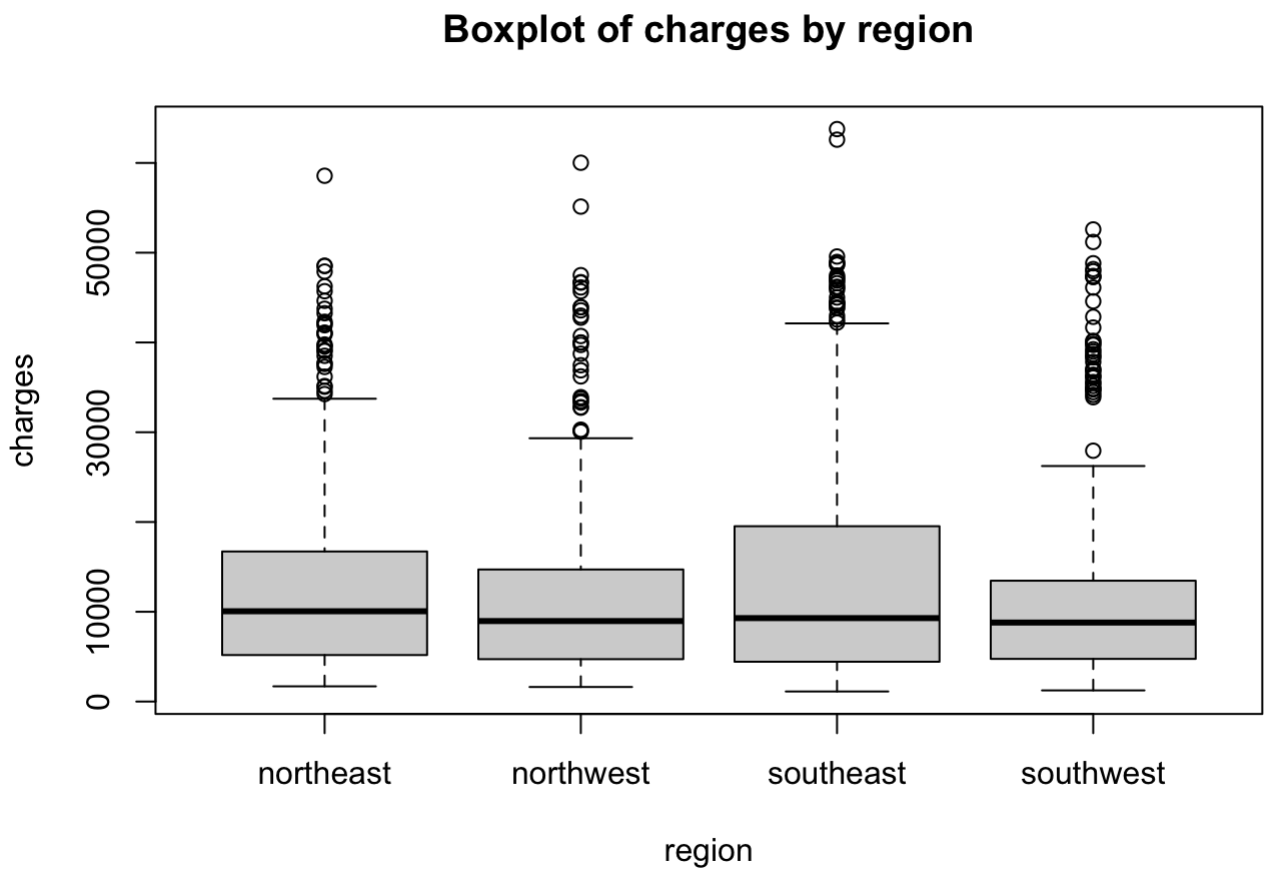
The boxplot above shows that sex seems to has little impact on the medical charges since the median of charges seems to be the same for female and male. However, we can see that the spread of charges value is larger for male.

```
boxplot(charges ~ smoker, data = data,
        xlab = "smoker", ylab = "charges",
        main = "Boxplot of charges by smoker")
```



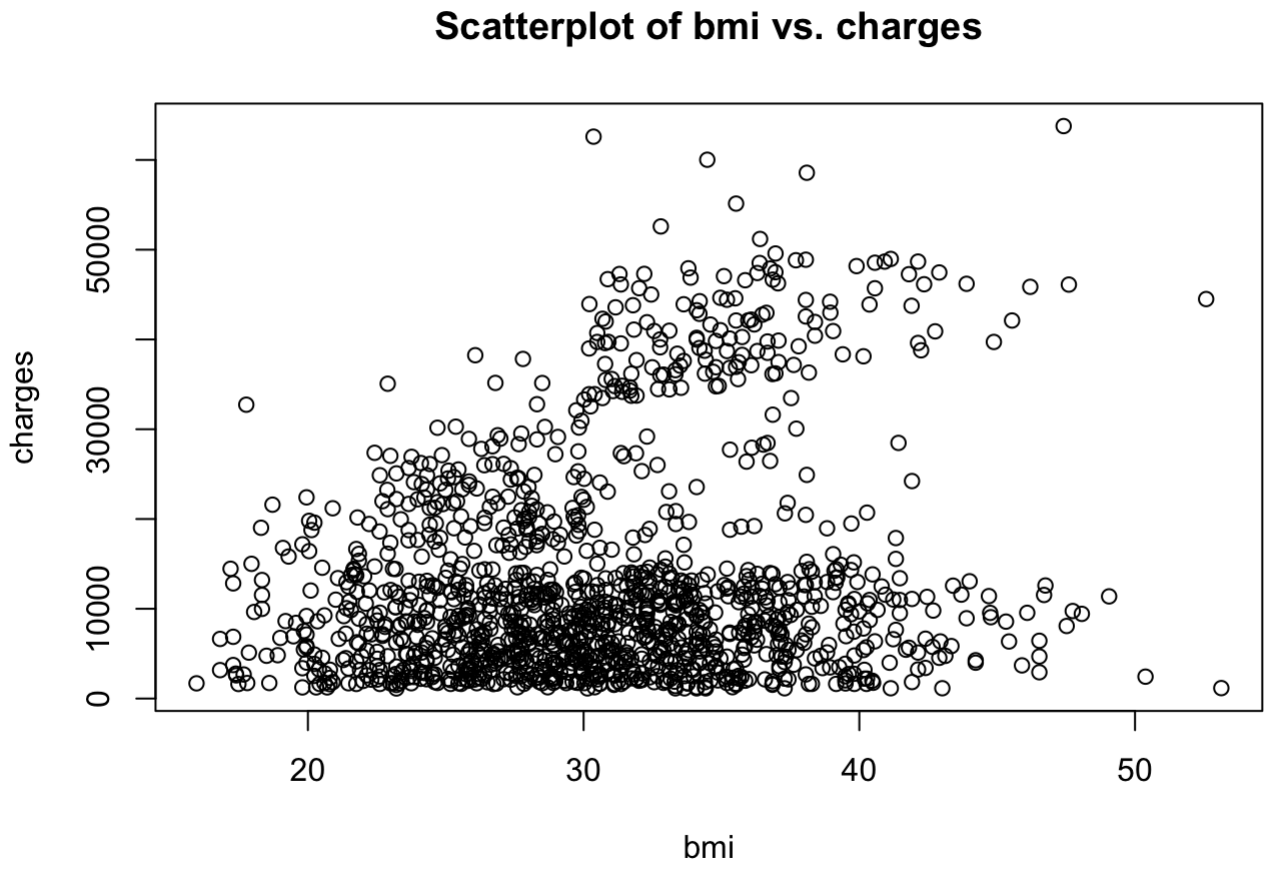
The boxplot shows that people who smoke has a significantly larger charges compared with those who do not since its median is much larger.

```
boxplot(charges ~ region, data = data,
        xlab = "region", ylab = "charges",
        main = "Boxplot of charges by region")
```



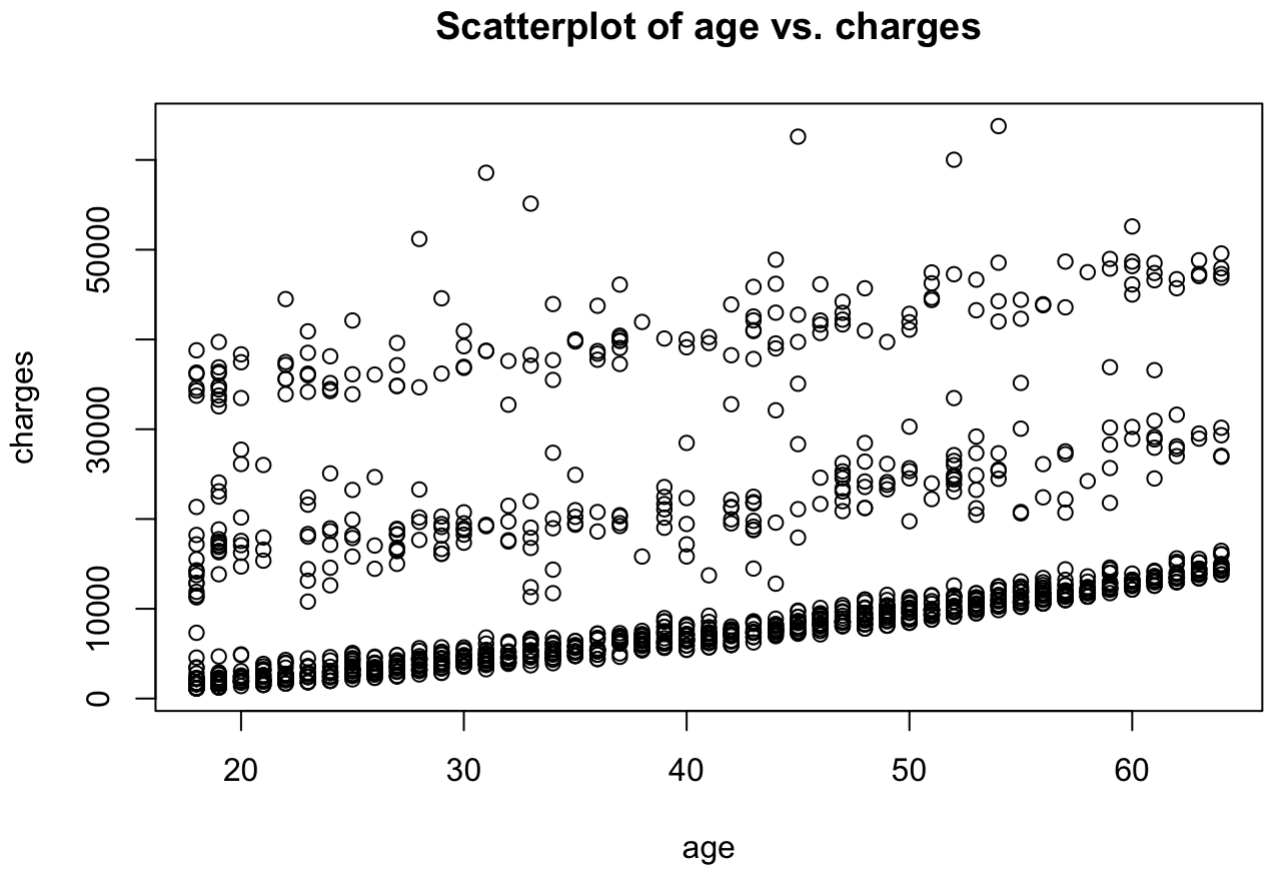
The plot shows that the difference in charges for people living in different regions is not very significant. People living in northeast have a little higher charges compared with other regions.

```
plot(data$bmi, data$charges,
     xlab = "bmi",
     ylab = "charges",
     main = "Scatterplot of bmi vs. charges")
```



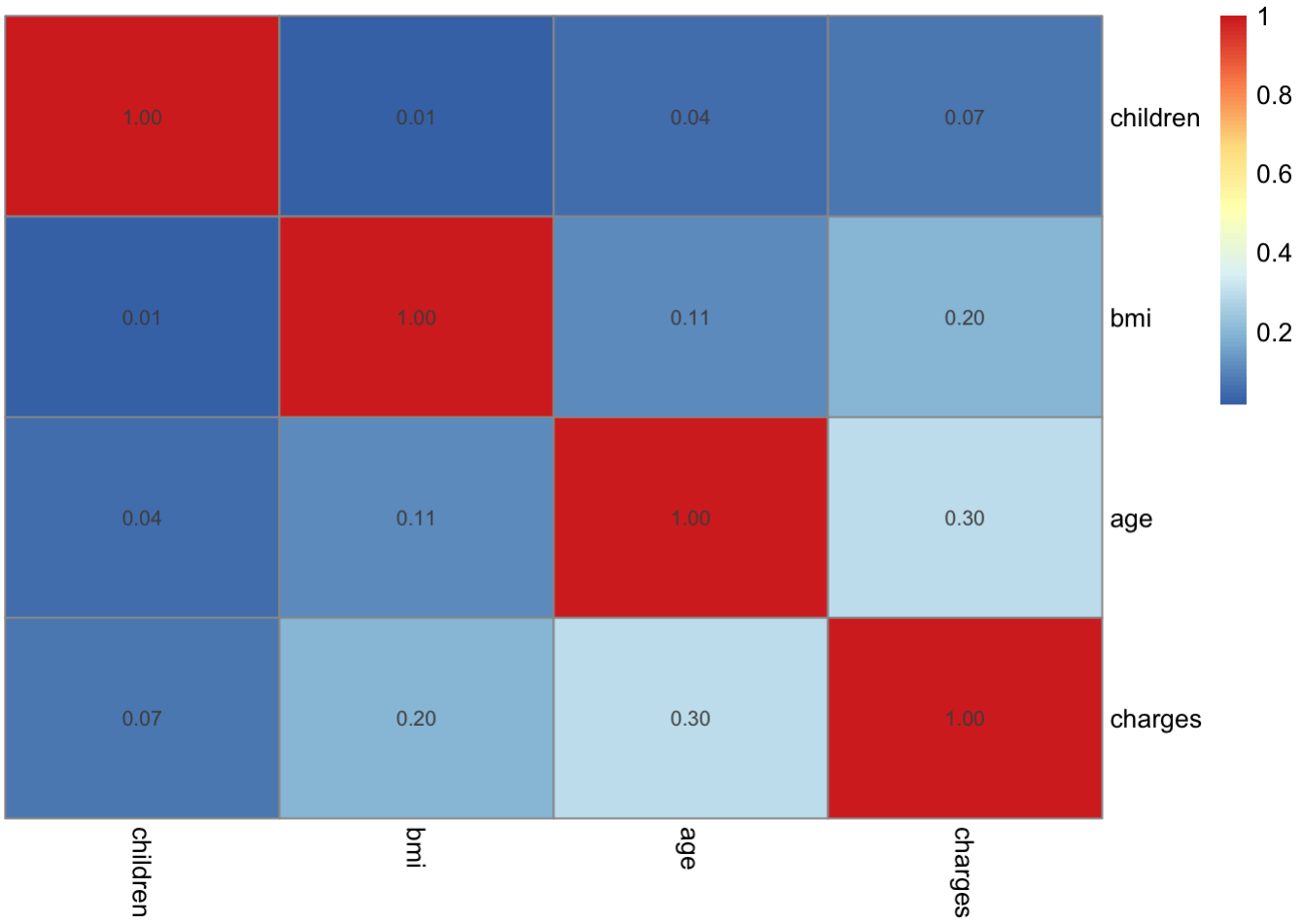
The scatterplot above shows a trend that people with higher bmi tend to have higher medical charges.Besides that, the degree of variability of the charges is higher at higher bmi which is reflected by the more scattered points.

```
plot(data$age, data$charges,
     xlab = "age",
     ylab = "charges",
     main = "Scatterplot of age vs. charges")
```



The trend here is interesting. While we can observe a positive relationship between age and medical charges, the points here seem to be divided into three different groups.

```
cor <- cor(data[, sapply(data, is.numeric)])
## Make the heatmap
pheatmap(cor,
         treeheight_col = 0,
         treeheight_row = 0,
         display_numbers = TRUE)
```



The heatmap above demonstrates the correlation between the numeric variables and charges. Across all the numeric variables, age has the highest correlation with the charges(0.3). From the three boxplots we got previously, we can conclude that smoker has the most significant impact on charges among all the categorical variables, thus has higher correlation with the response variable charges.

Multiple Linear Regression Model

```
model=lm(charges~.,data=data)
summary(model)
```

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

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

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11938.5      987.8 -12.086  < 2e-16 ***
age            256.9       11.9  21.587  < 2e-16 ***
sexmale       -131.3      332.9  -0.394 0.693348    
bmi            339.2       28.6  11.860  < 2e-16 ***
children       475.5      137.8   3.451 0.000577 ***
```

```
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

Interpretation of slope for numeric variables:
Holding other factors constant,the increase in age by 1 unit will lead to an increase in charges by $256.9.
Holding other factors constant,the increase in bmi by 1 unit will lead to an increase in charges by $339.2.
Holding other factors constant,the increase in children by 1 unit will lead to an increase in charges by $475.5.

Interpretation of slope for categorical variables:
Holding other factors constant, on average the person who smoke will lead to $23848.5 increase in charges compared with the one who does not.
On average, holding other factors constant, males are expected to have charges that are $131.3 lower compared to charges incurred by females.
Holding other factors constant, on average the persons who live in northwest region are expected to have charges that are $353 lower compared to charges incurred by people living in northeast.
Holding other factors constant, on average the persons who live in southeast region are expected to have charges that are $1035 lower compared to charges incurred by people living in northeast.
Holding other factors constant, on average the persons who live in southwest region are expected to have charges that are $960 lower compared to charges incurred by people living in northeast.

```
summary(model)
```

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

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

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

The null hypothesis for f test here is that all the coefficients of the numeric variables are zero, and there is no difference in charges based on sex, smoking status, and regions. The alternative hypothesis is that there is at least one variable that has impact on the medical charges of the person. The F-statistic is 500.8 on 8 and 1329 DF and p value is effectively zero, indicating that we should reject the null hypothesis and the model is significant.

```
summary(model)$r.squared
```

```
[1] 0.750913
```

The value of multiple r square is 0.750913.It means that 75.0913% of the variability in the dependent variable charges is accounted for by the predictor variables included in the model.

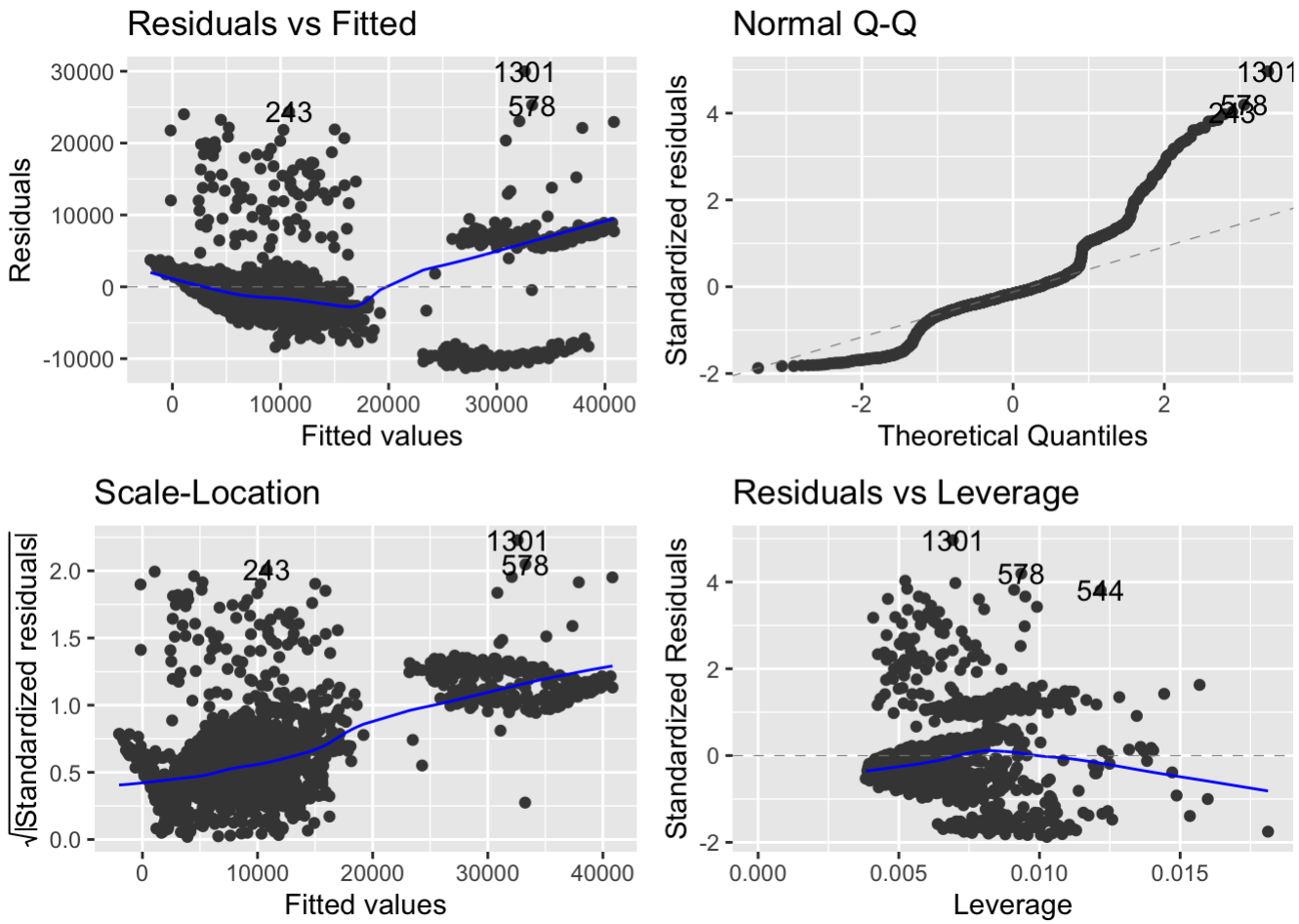
```
summary(model)$sigma
```

```
[1] 6062.102
```

the RSE of 6062 means that, on average, the residuals (the differences between actual and predicted values) are around 6062 units with the degree of freedom of 1329.

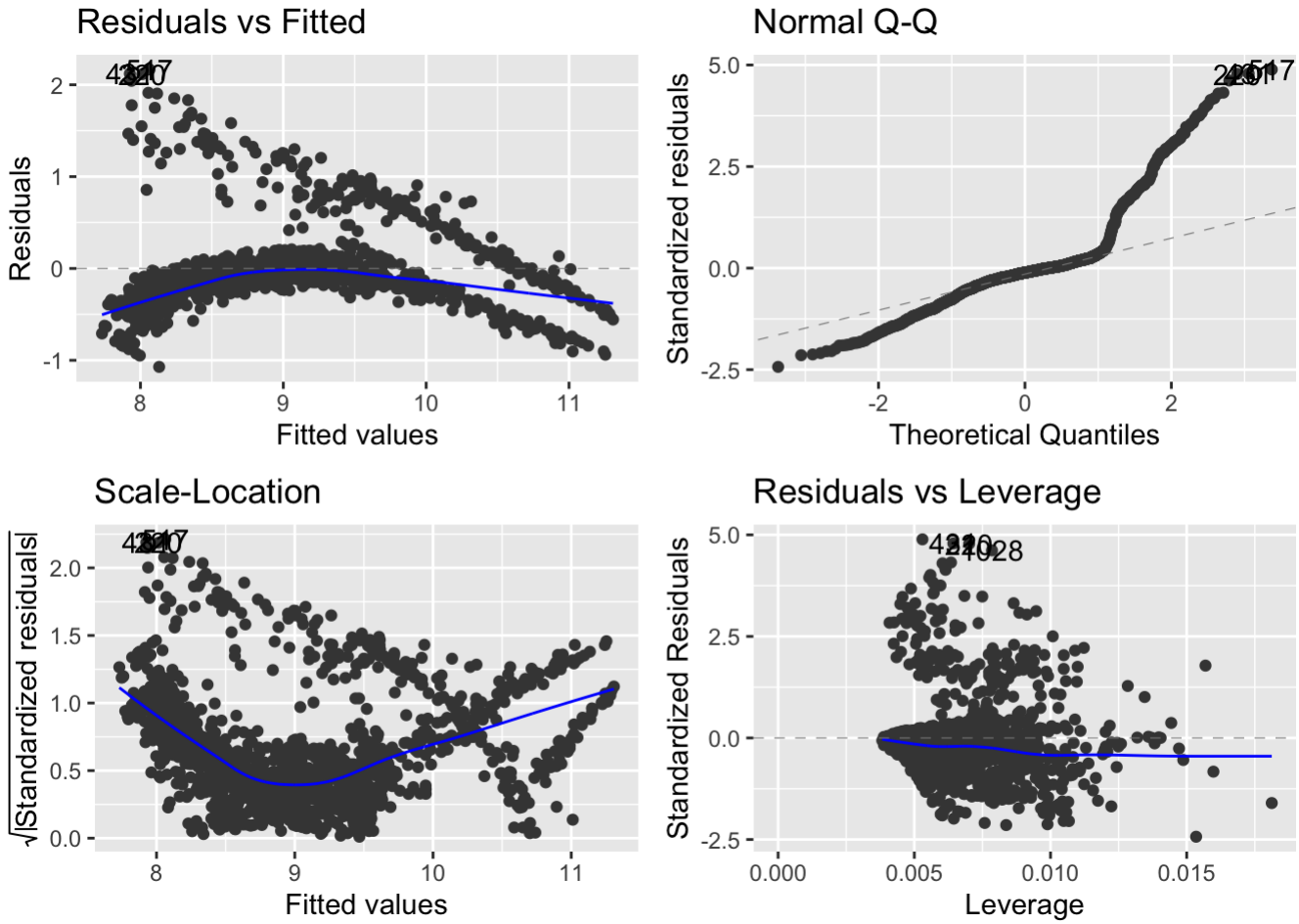
Improving the Model

```
autoplot(model)
```



Since the blue line on the first plot increases drastically from the middle part, we can conclude that, the errors on average is not zero. There is a obvious increasing trend for scale-location plot, indicating that errors do not have a constant variance. The QQ-plot shows systematic deviation from the line, showing that normality assumption is violated.

```
model2=lm(log(charges)~.,data=data)
autoplot(model2)
```



By conducting the log transformation on the dependent variable, we can observe a significant improvement on the residuals vs fitted values plot. We can approximately conclude that the errors on average is near to zero.However, the assumptions on variance and normality are still violated.

```
library(splines)
model3=lm(charges~bmi+ns(age,2)+sex+children+smoker+region,data=data)
summary(model3)
```

```
Call:
lm(formula = charges ~ bmi + ns(age, 2) + sex + children + smoker +
    region, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-11660.0  -2875.4   -952.5   1271.5  30884.5

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -6332.44     954.91  -6.631 4.82e-11 ***
bmi               335.08      28.45  11.777  < 2e-16 ***
ns(age, 2)1     10721.08     923.54  11.609  < 2e-16 ***
ns(age, 2)2     11059.94     603.66  18.322  < 2e-16 ***
sexmale          -139.47     331.03  -0.421   0.6736
children          650.18     143.61   4.528 6.50e-06 ***
smokeryes       23867.36     410.79  58.101  < 2e-16 ***
regionnorthwest  -370.85     473.55  -0.783   0.4337
```

```
regionsoutheast -1028.54     475.93  -2.161   0.0309 *
regionsouthwest  -960.58     475.17  -2.022   0.0434 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6027 on 1328 degrees of freedom
Multiple R-squared:  0.754, Adjusted R-squared:  0.7523
F-statistic: 452.2 on 9 and 1328 DF,  p-value: < 2.2e-16
```

```
anova(model,model3)
```

```
Analysis of Variance Table

Model 1: charges ~ age + sex + bmi + children + smoker + region
Model 2: charges ~ bmi + ns(age, 2) + sex + children + smoker + region
  Res.Df        RSS Df Sum of Sq     F    Pr(>F)
1   1329 4.8840e+10
2   1328 4.8241e+10  1 598661836 16.48 5.203e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p value is less than 0.05, the model with the natural spline is better.

```
model4=lm(charges~children*age+bmi+sex+smoker+region,data=data)
anova(model,model4)
```

```
Analysis of Variance Table

Model 1: charges ~ age + sex + bmi + children + smoker + region
Model 2: charges ~ children * age + bmi + sex + smoker + region
  Res.Df        RSS Df Sum of Sq     F Pr(>F)
1   1329 4.8840e+10
2   1328 4.8839e+10  1    476282 0.013 0.9094
```

Null hypothesis: There is no interaction between children and age.

Alternative Hypothesis: There is interaction between children and age.

Since the p value is more than 0.05, we cannot reject the null hypothesis. We should not include interaction between children and age.

How to interpret the interaction term: The effect of children on charges depends on the value of age. The effect of age on charges depends on the value of children. The effect of variables children and age on the outcome is not constant and dependent of the values of the other。

Formal Hypothesis Tests

The first hypothesis I would like to make is that individuals who smoke are likely to incur higher medical charges in comparison to those who do not smoke.
Hypothesis1: Null:Variable Smoker has a coefficient of zero or negative number.
Alternative: The coefficient of variable Smoker is a positive number.

```
summary(model)
```

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

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

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348    
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769    
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *  
regionsouthwest   -960.0      477.9  -2.009 0.044765 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

From the summary table, we can see that the p value of smokeryes is < 2e-16 which is less than 0.05 and the coefficient is 23848.5 which is a positive number.Thus,we can reject the null hypothesis and conclude that individuals who smoke are likely to incur higher medical charges in comparison to those who do not smoke

The second hypothesis I would like to make is that older individuals are more prone to incurring higher medical charges compared to their younger counterparts.
Hypothesis2: Null:Variable Age has a coefficient of zero or negative number
Alternative: The coefficient of variable Age is a positive number.

```
summary(model)
```

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

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

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
```

```
sexmale           -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

From the summary table, we can see that the p value of age is < 2e-16 which is less than 0.05 and the coefficient is 256.9 which is a positive number.Thus,we can reject the null hypothesis and conclude that older individuals are more prone to incurring higher medical charges compared to their younger counterparts.

The last hypothesis is that people with higher BMI are likely to incur higher medical charges.
Hypothesis3: Null:Variable BMI has a coefficient of zero or negative number.
Alternative: The coefficient of variable BMI is a positive number.

```
summary(model)
```

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

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

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

From the summary table, we can see that the p value of age is < 2e-16 which is less than 0.05 and the coefficient is 339.2 which is a positive number.Thus,we can reject the null hypothesis and

conclude that people with higher BMI are likely to incur higher medical charges.

Thus, for all three hypotheses that we posited, they are all supported by our result. However, we only have 1338 observations in our dataset which is too little to represent the whole population. Therefore, we should also be cautious about this limitation.

```
#Since we have three tests in total.
0.05/3
```

```
[1] 0.01666667
```

Since all the p value are still less than 0.01666667, none are insignificant after making a Bonferonni correction.The conclusion does not change.

Robustness of Results

```
confint_perclm <- function(object, level = 0.95) {
  L <- (1 - level) / 2
  U <- 1 - L
  t(perc.lm(object, c(L, U)))
}

get_se_lm <- function(object) {
  sqrt(diag(vcov(object)))
}

se_lm_boot <- function(object) {
  summary(object)[["stdev.params"]]
}#HERE IS STANDARD DEVIATION
```

```
boot_results <- lm.boot(model, R = 999)
sterror_boot=se_lm_boot(boot_results)
sterror_lm=get_se_lm(model)
sterror_boot
```

```
    (Intercept)             age         sexmale             bmi        children 
     1045.41403        12.07119       329.38044        31.50560       135.29186 
      smokeryes regionnorthwest regionsoutheast regionsouthwest 
      564.68250       461.38019       513.69503       455.29211 
```

```
sterror_lm
```

```
    (Intercept)             age         sexmale             bmi        children 
      987.81918        11.89885       332.94544        28.59947       137.80409 
      smokeryes regionnorthwest regionsoutheast regionsouthwest 
      413.15335       476.27579       478.69221       477.93302 
```

```
## Compute the T-statistic
t_boot <- coef(model) / sterror_boot
```

```
## Compute the P-values
p_boot <- 2 * pt(abs(t_boot), df =1338-13-1, lower.tail = FALSE)

## Print T statistics and P-values
print(cbind(`t value` = t_boot, `Pr(>|t|)` = p_boot))
```

```
                   t value      Pr(>|t|)
(Intercept)     -11.4199142  7.105650e-29
age              21.2784664  1.175746e-86
sexmale          -0.3986708  6.902001e-01
bmi              10.7661334  5.705292e-26
children          3.5146279  4.552202e-04
smokeryes        42.2335289 1.442407e-247
regionnorthwest  -0.7650175  4.443975e-01
regionsoutheast  -2.0148570  4.412146e-02
regionsouthwest  -2.1086484  3.516237e-02
```

```
summary(model)
```

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

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

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

The values are close but the standard error obtained using boot is higher compared with the one obtained using lm function for variables age,sexmale,bmi,smokeryes and regionsouthest. For the rest variables, the standard errors obtained are lower. we used the bootstrap standard errors to compute the T-statistic and associated P-value for testing whether each of the slopes is equal to zero and compare these P-values to the output of `summary`. There is no variables that were statistically significant at the 0.05 significance level when using the `summary` function no longer significant when the bootstrap is used.

```
library(DAAG)
loo_mse=press(model)/nrow(data)
loo_mse
```

```
[1] 37056293
```

the leave one out cross validated mean squared error is 37056293 which is quite high.

```
predict_loo <- function(model) {
  y <- model.frame(model)[,1]
  loo_r <- residuals(model) / (1 - hatvalues(model))
  return(y - loo_r)
}
rsq_loo <- function(model) {
  y <- model.frame(model)[,1]
  yhat <- predict_loo(model)
  return(cor(y, yhat)^2)
}
rsq_loo(model)
```

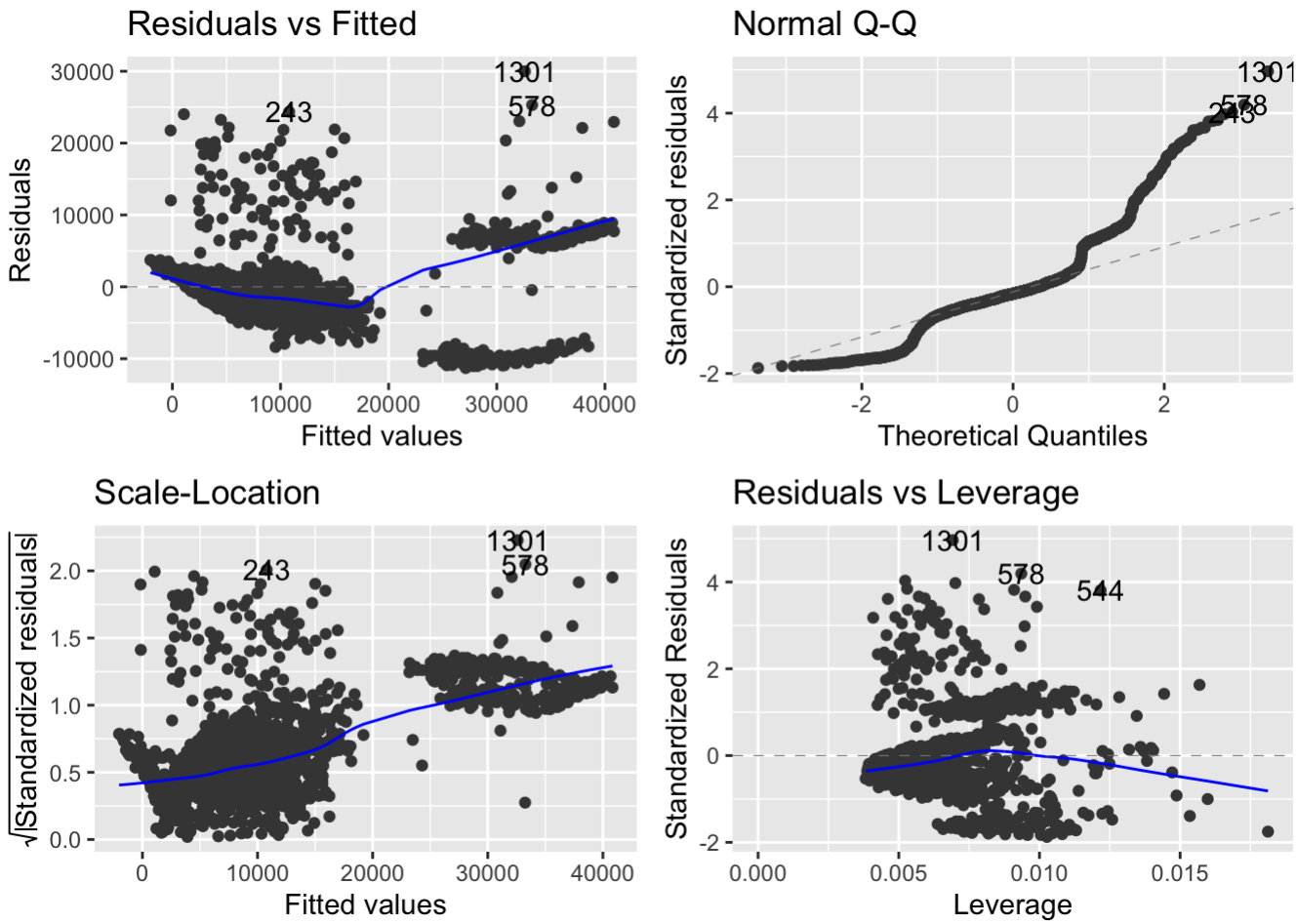
```
[1] 0.7471345
```

```
summary(model)$r.squared
```

```
[1] 0.750913
```

since the difference between loo r^2 and r^2 we got from summary(model) is not significant. We can conclude that the model is not overfitted.

```
autoplot(model)
```



Using the residuals vs leverage plot under autoplot, observation at row 544 and 578 and 1301 are considered as influential since they have a high leverage and high standardized residuals.

```
dfbetas(model)[c(533,578,1301),]
```

```
       (Intercept)         age      sexmale          bmi      children
533   0.0001379068 -0.00100750 -0.000718706  0.0005831097 -0.0004922093
578  -0.0389753041 -0.08751233 -0.144361316  0.1954424354 -0.0042267962
1301  0.0044980153  0.08179037  0.120332537 -0.0788917383 -0.1263376608
        smokeryes regionnorthwest regionsoutheast regionsouthwest
533  0.0004607436    1.917074e-05    -0.001021721   -2.094491e-05
578  0.2392327170   -1.617922e-01    -0.222562536   -1.763930e-01
1301 0.2484126702    1.072963e-02     0.184776301    1.602036e-02
```

From the results of dfbetas,we can see that deleting row 533 will has a very insignificant effect on coefficient of predictors. Deleting row 578 will shift the regression coefficient of smokeryes and regionsoutheast by 0.2392327170 and 0.222562536 of a standard deviation respectively. Deleting row 1301 will shift the regression coefficient of smokeryes by 0.2484126702 of a standard deviation.

```
library(car)
```

```
Loading required package: carData


Attaching package: 'car'
```

```
The following object is masked from 'package:DAAG':

    vif

The following object is masked from 'package:boot':

    logit

The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some
```

```
vif_results <- vif(model)
vif_results
```

```
             GVIF Df GVIF^(1/(2*Df))
age      1.016822  1        1.008376
sex      1.008900  1        1.004440
bmi      1.106630  1        1.051965
children 1.004011  1        1.002003
smoker   1.012074  1        1.006019
region   1.098893  3        1.015841
```

Since the value of vif for all the variables here are less than 5, we can conclude that none of the variables is concerning.

Conclusions

From my dataset, I have found out that the medical charges of an individual is closely related to his or her age, bmi, smoking or not and no of children has. The findings can serve as a reference for the insurance companies when they determine premium of the insurance. It also reminds them the importance of conducting background survey on the individual's habits, demographics and health history before determining the premium.

This dataset only provide us with 1338 observations which is not sufficient for us to make our conclusion here. If possible, we can acquire more data for future more accurate analysis. Besides that, this dataset only provides us with six features. Predictors like sleeping hours could also potentially be important in determining the medical charges. Thus, we can use a dataset with more variables in the future for more comprehensive analysis.