Math 644 Project

Gkeri Pepelasi

12/11/2021

```
library(ggplot2)
library(ggcorrplot)
library(dplyr)          # data wrangling
library(caret)          # machine learning functions
library(MLmetrics)      # machine learning metrics
library(car)            # VIF calculation
library(lmtest)
```

Predicting medical expense

Intro

In order for a health insurance company to make money, it needs to collect more in yearly premiums than it spends on medical care to its beneficiaries. As a result, insurers invest a great deal of time and money in developing models that accurately forecast medical expenses for the insured population.

Medical expenses are difficult to estimate because the most costly conditions are rare and seemingly random. Still, some conditions are more prevalent for certain segments of the population. For instance, lung cancer is more likely among smokers than non-smokers, and heart disease may be more likely among the obese.

The goal of this analysis is to use patient data to estimate the average medical care expenses for such population segments. These estimates can be used to create actuarial tables that set the price of yearly premiums higher or lower, depending on the expected treatment costs.

```
insurance <- read.csv("C:\\Users\\gpepe\\OneDrive\\Documents\\insurance.csv", header = T,stringsAsFacto
insurance <- data.frame(insurance)
```

insurance is a dataframe with 1,338 observations and 7 variables:

1. age: age of primary beneficiary
2. sex: insurance contractor gender, female, male
3. BMI: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ˆ 2) using the ratio of height to weight, ideally 18.5 to 24.9
4. children: Number of children covered by health insurance / Number of dependents
5. smoker: Smoking or not
6. region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest
7. charges: Individual medical costs billed by health insurance

Data exploration

It is important to give some thought to how these variables may be related to billed medical expenses. For instance, we might expect that older people and smokers are at higher risk of large medical expenses. Unlike many other machine learning methods, in regression analysis, the relationships among the features are typically specified by the user rather than being detected automatically.

```
str(insurance)
```

```
## '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 ...
```

Our model's dependent variable is expenses, which measures the medical costs each person charged to the insurance plan for the year. Prior to building a regression model, it is often helpful to check for normality. Although linear regression does not strictly require a normally distributed dependent variable, the model often fits better when this is true. Let's take a look at the summary statistics:

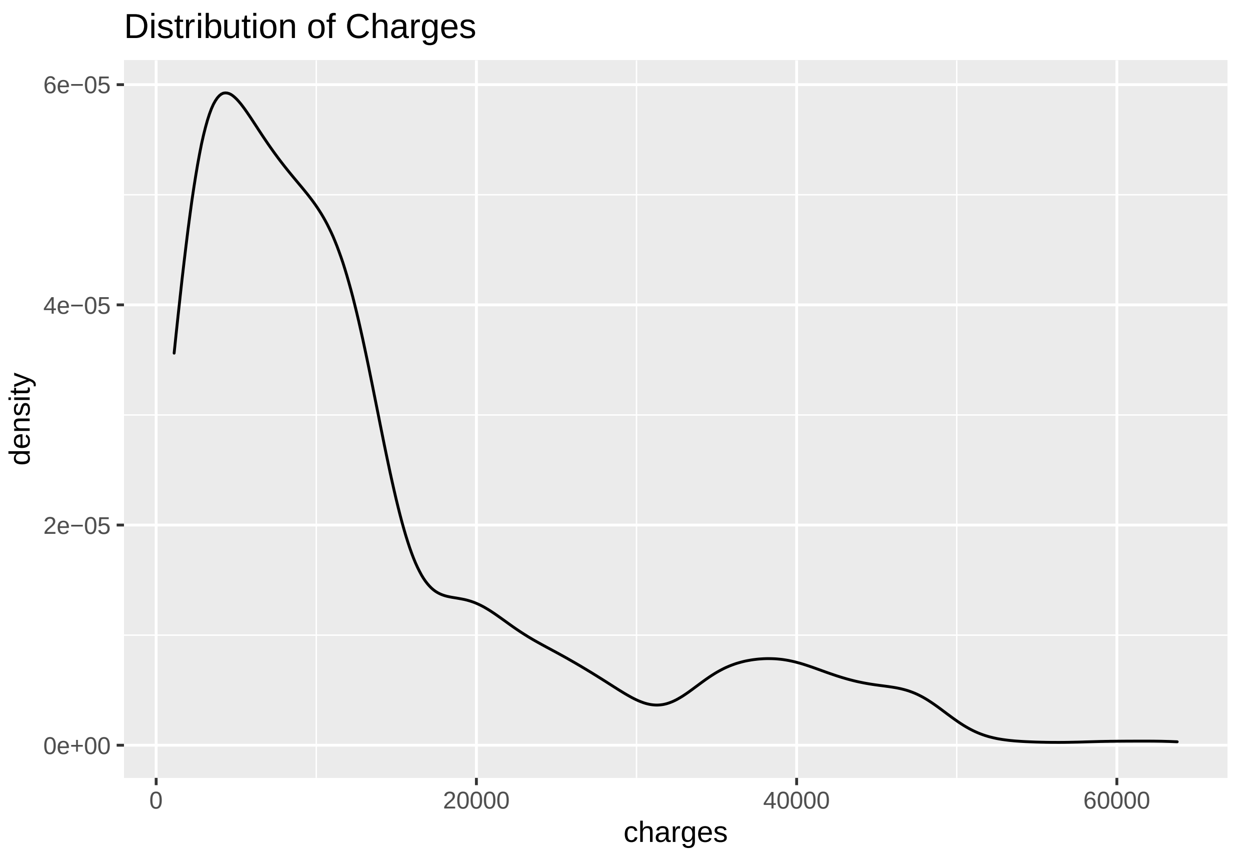
```
summary(insurance)
```

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

Looking at the response variable, the minimum value is 1122 while the maximum value is 63770. Most points cluster between 4740 and 16640. This large variance in the response variable indicates that there are potential outliers. The other quantitative variables are reasonably varied.

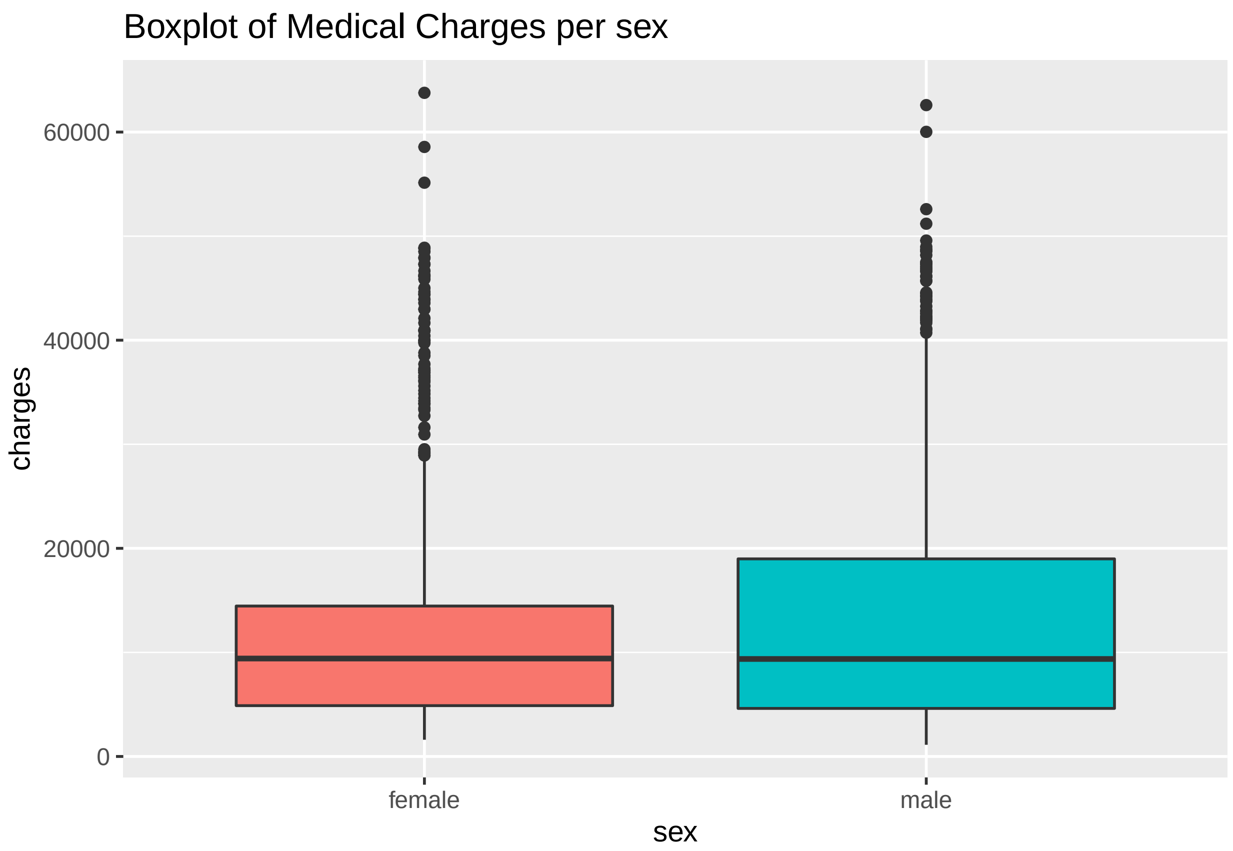
Because the mean value of charges is greater than the median, this implies that the distribution of insurance expenses is right-skewed. We can confirm this visually using a histogram and the output is shown as follows:

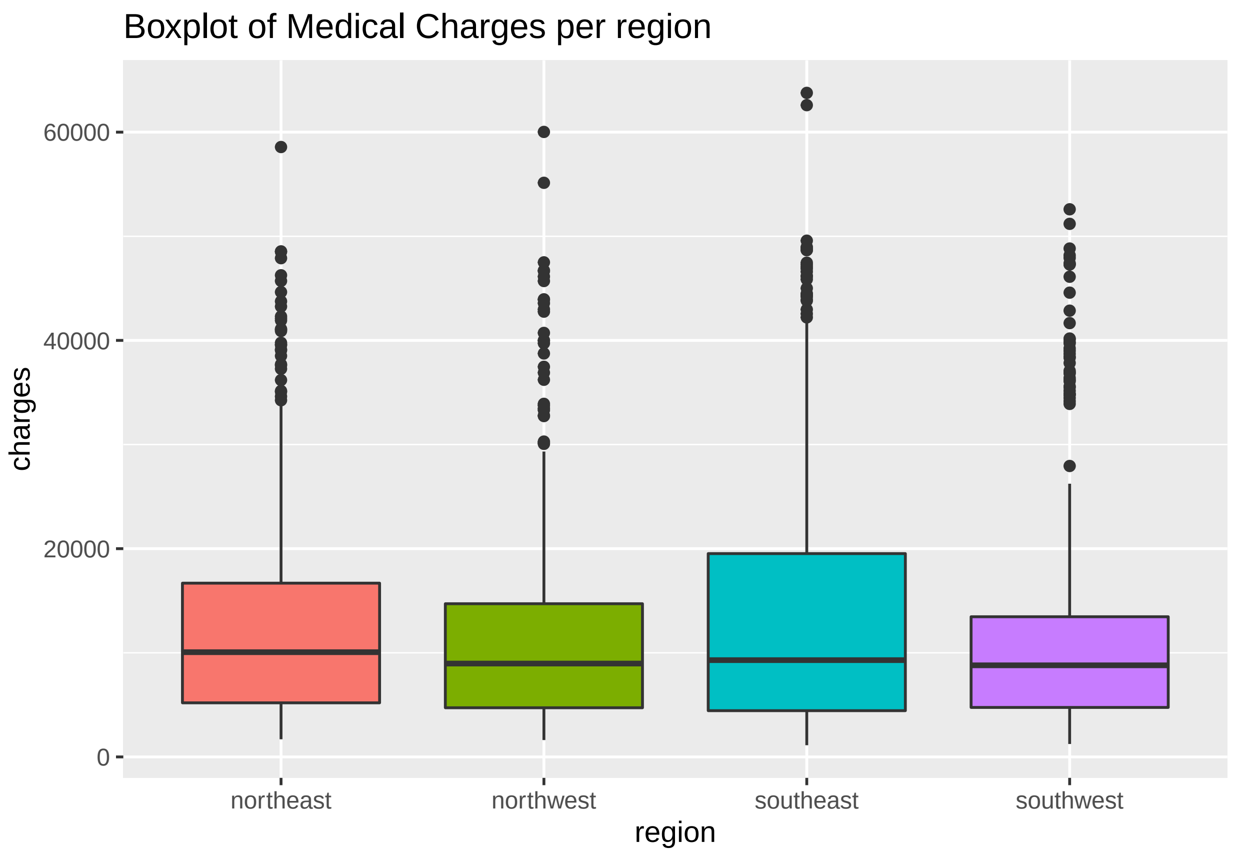
```
ggplot(data = insurance, aes(x = charges)) +

  geom_density(alpha=0.5)+

  ggtitle("Distribution of Charges")
```

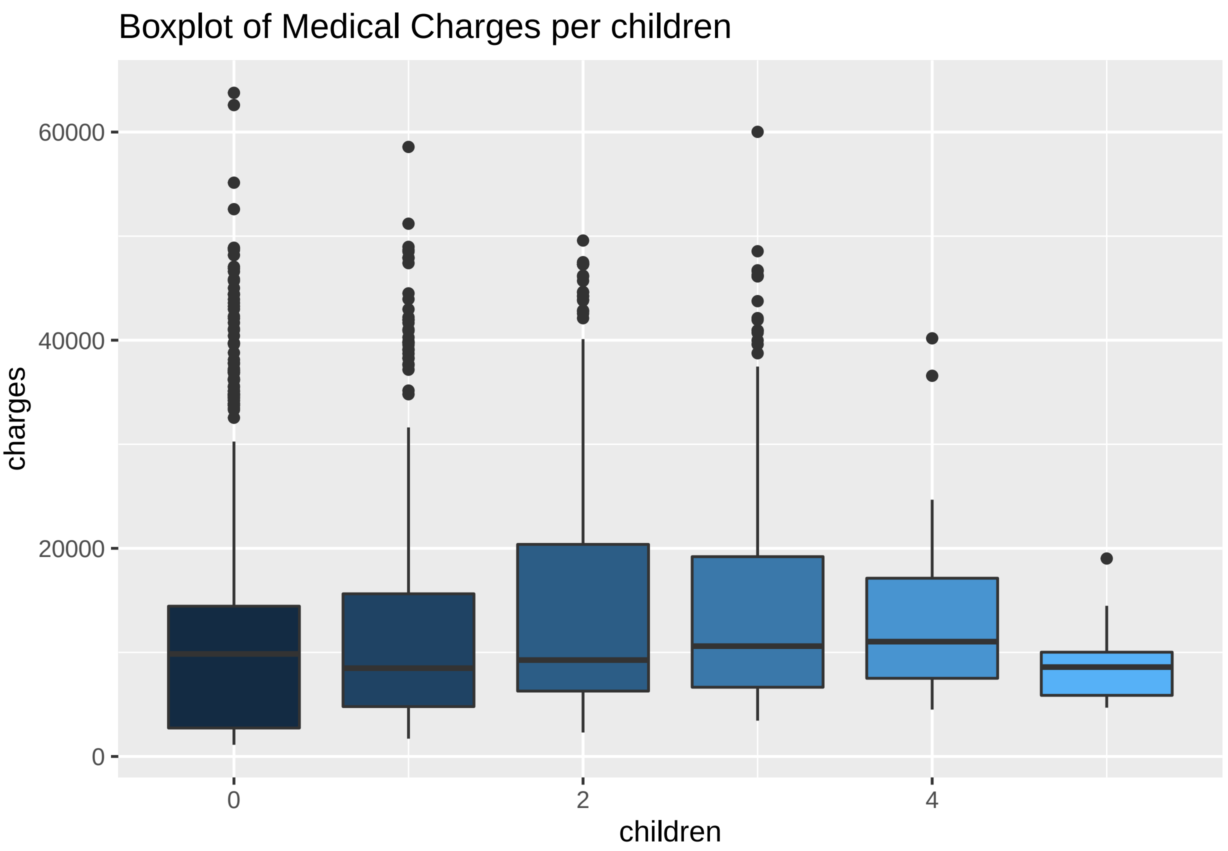


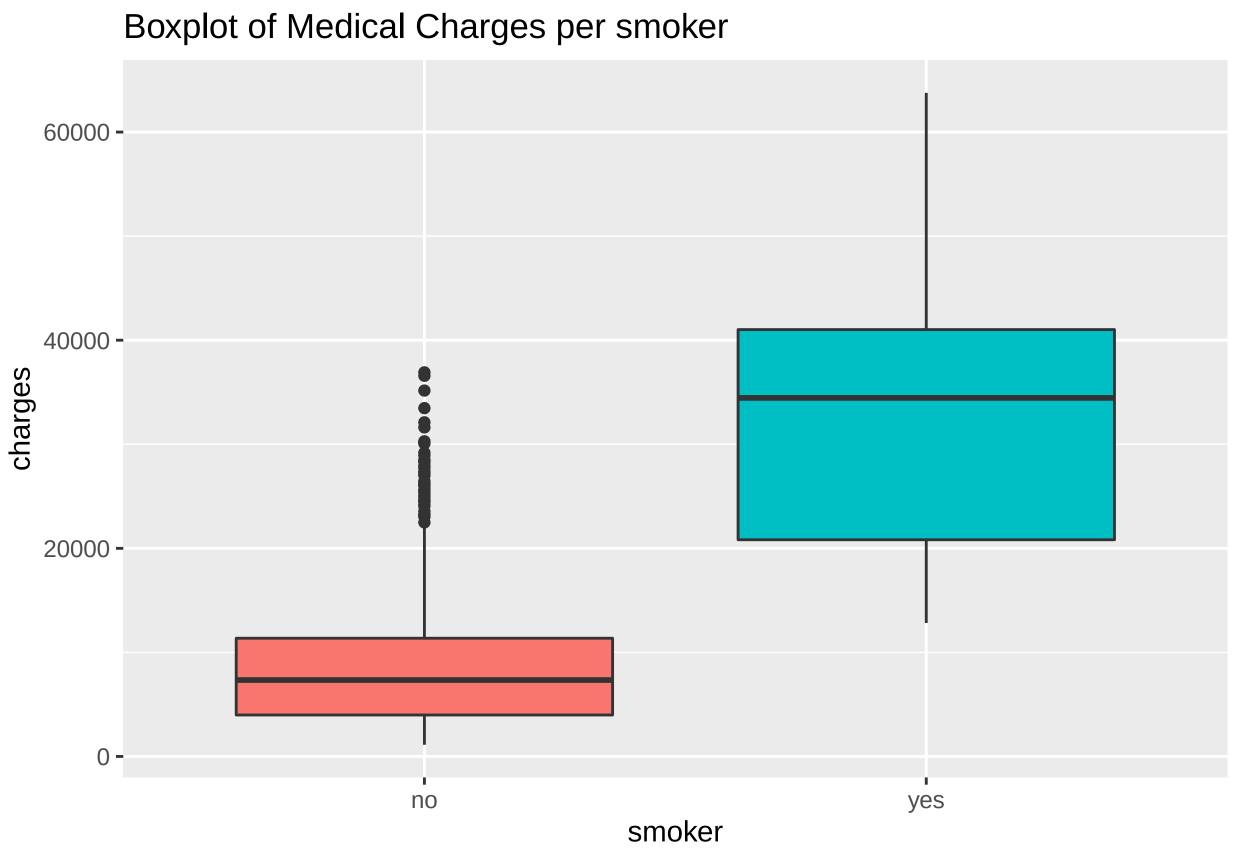
The distribution is right-skewed with a long tail to the right.The large majority of individuals in our data have yearly medical expenses between zero and $15,000, although the tail of the distribution extends far past these peaks. Because linear regression assumes a normal distribution for the dependent variable, this distribution is not ideal. In practice, the assumptions of linear regression are often violated. If needed, we may be able to correct this later on. There's a bump at around $40,000, perhaps another hidden distribution.

```r
for (col in c('sex', 'region', 'children', 'smoker')) {
  plot <- ggplot(data = insurance,
               aes_string(x = col, y = 'charges', group = col, fill = col)) +
            geom_boxplot(show.legend = FALSE) +
            ggtitle(glue::glue("Boxplot of Medical Charges per {col}"))
  print(plot)
}
```


Boxplot of Medical Charges per sex
charges
60000
40000
20000
0
female
male
sex


Boxplot of Medical Charges per region
charges
60000
40000
20000
0
northeast
northwest
southeast
southwest
region


Boxplot of Medical Charges per children
charges
60000
40000
20000
0
0
2
4
children

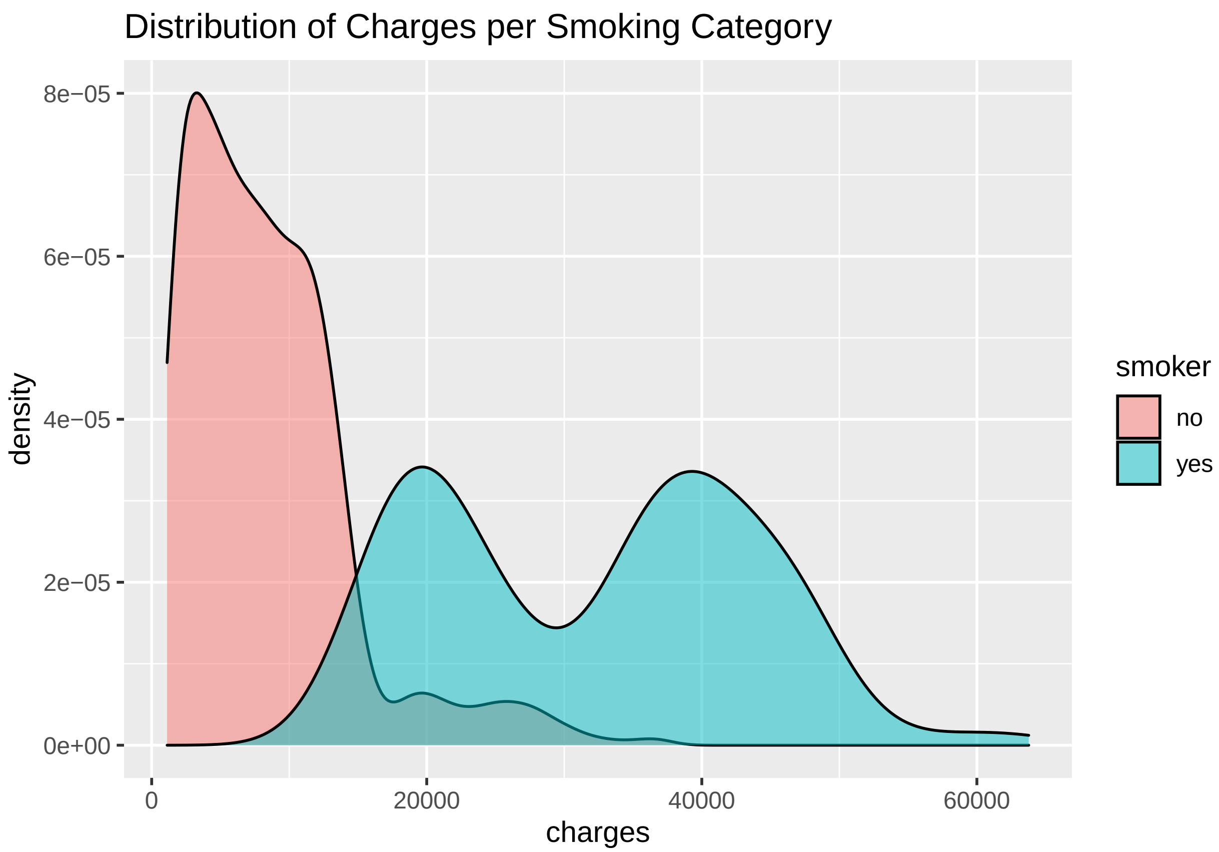


The plot above shows the boxplot of variable sex for insurance costs. The median costs for both sexes are pretty equal though there is more variance in insurnace costs for male. There's not a clear trend for variable region in relation with insurance costs. The insurance costs decreases slightly from east to west, however.

There's a clear trend here. Smokers have a much higher median insurance costs in comparison with non-smokers.

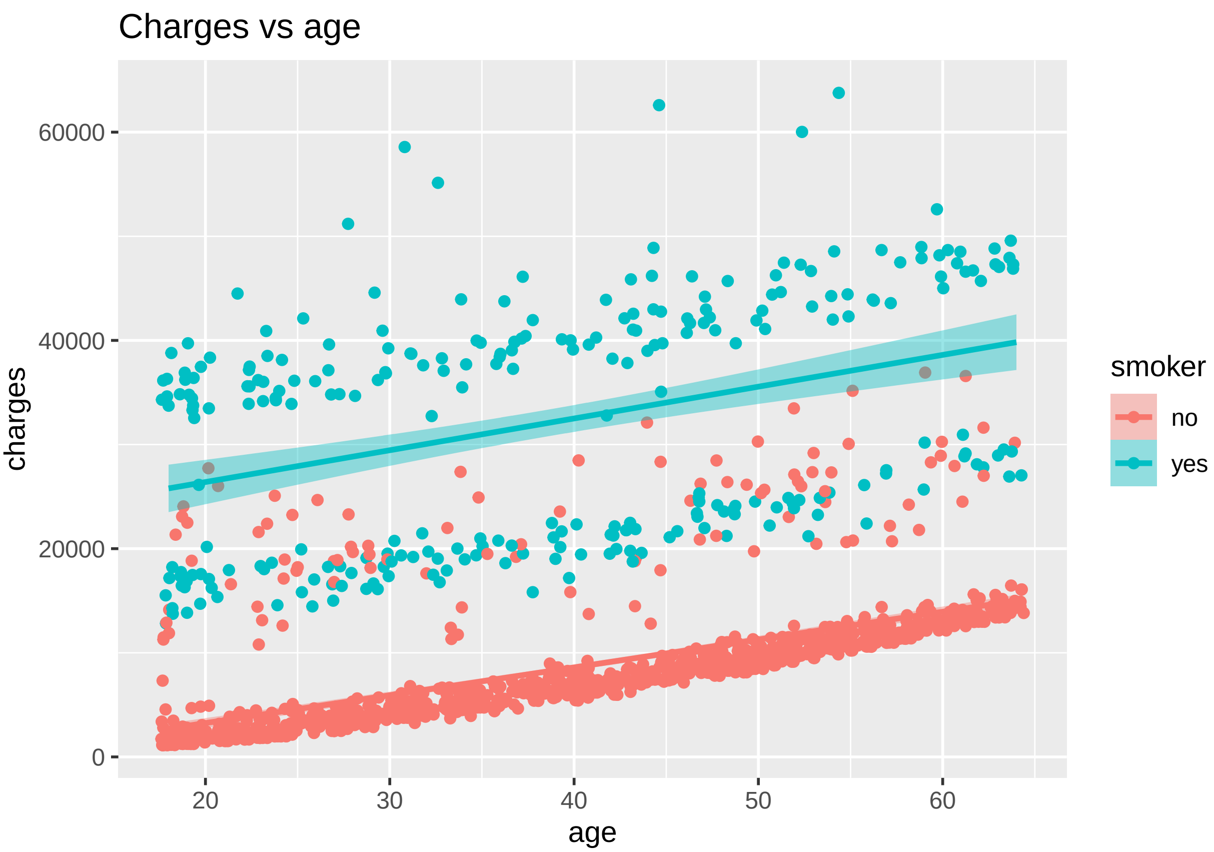
The median insurance costs start high for contractors with zero children then goes down for 1 children contractors. The median costs keep increasing but then decreases when a contractor has 5 children. This could be due to the insurance companies policy to start with a high default cost. They give discount for contractors with children at a small rate then give really high discount for contractors with more than 5 children. One thing to note is that the boxplots show there are many outliers in our categorical variables. The outliers have the potential to influence the model so we'll come back to address this issue if necessary. Lastly, smoker seems to make a significant difference to charges given by health insurance. Let's draw again the distribution of charges, now categorizing them into smoker.

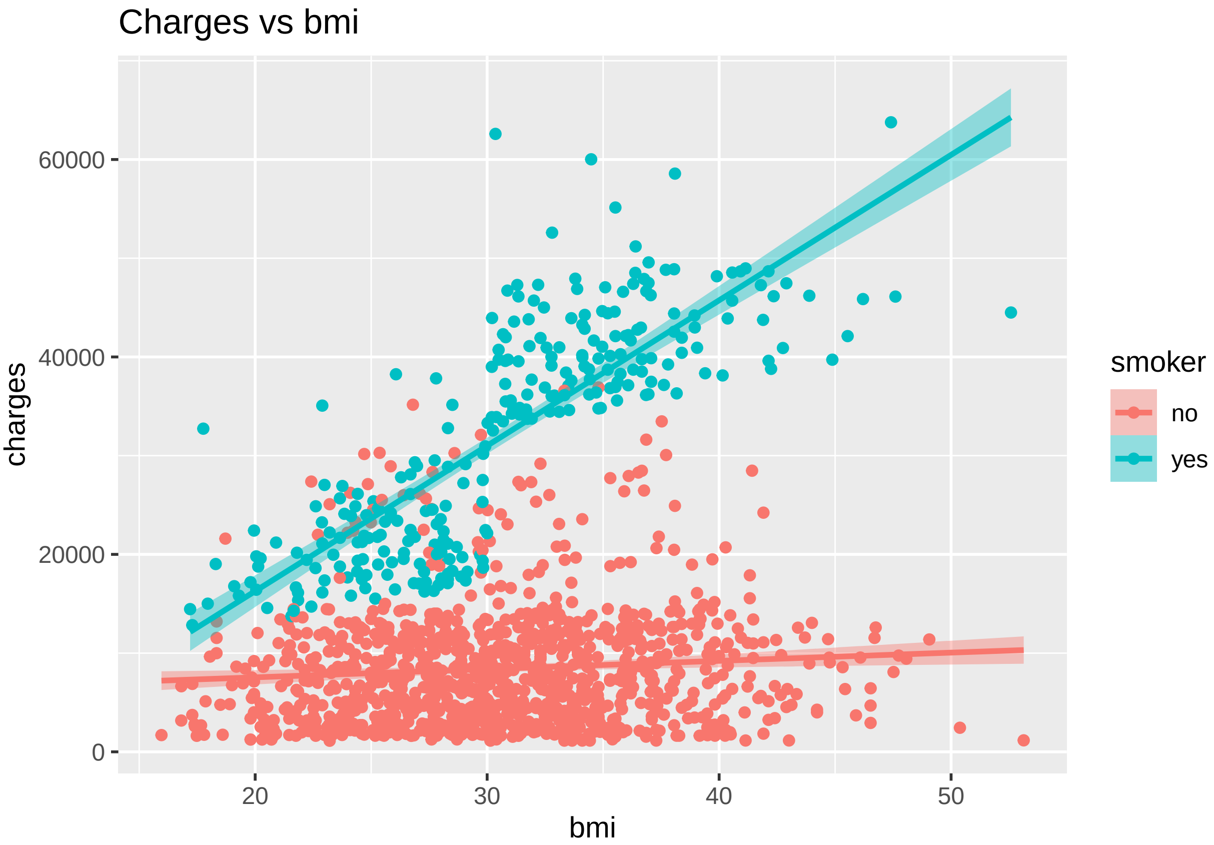
```r
ggplot(data = insurance, aes(x = charges, fill = smoker)) +
  geom_density(alpha = 0.5) +
  ggtitle("Distribution of Charges per Smoking Category")
```

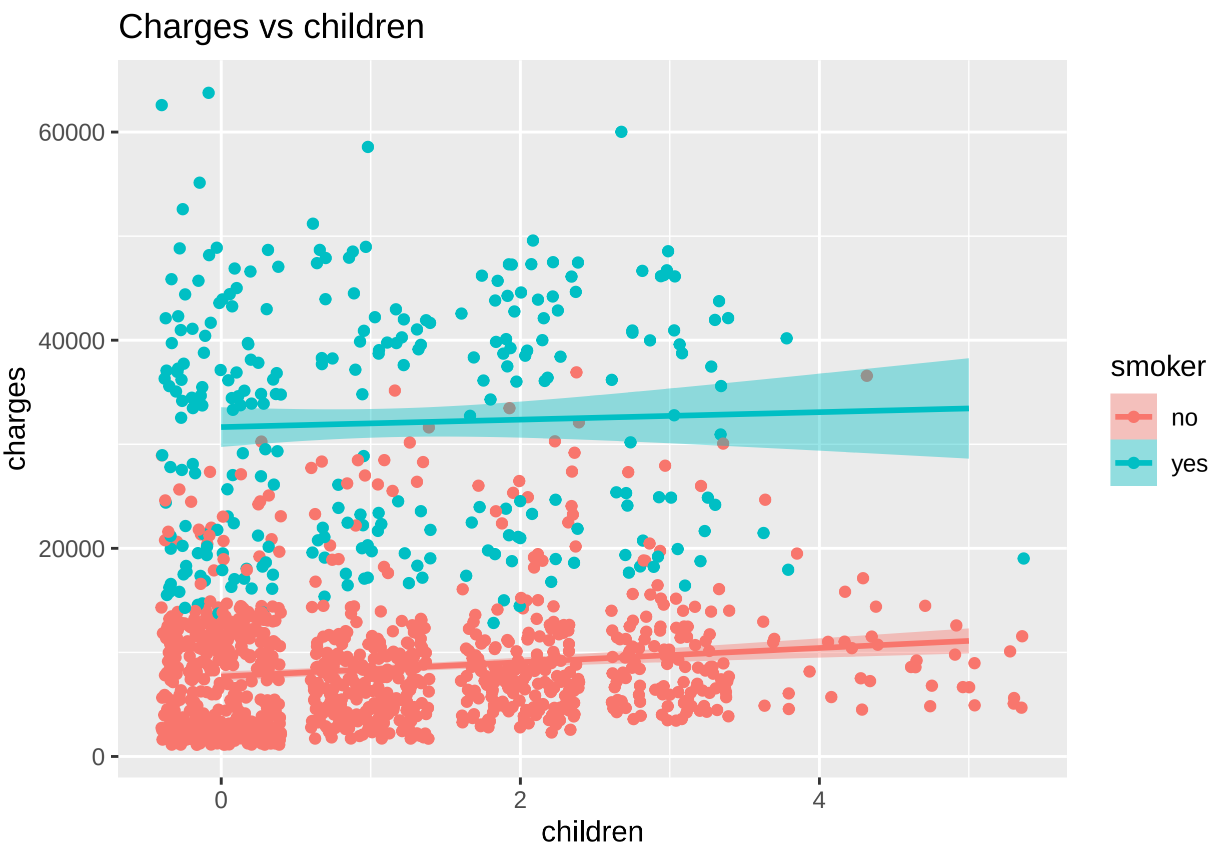


We see clearly that smokers have more charges than non-smokers.

```
for (feat in c('age', 'bmi', 'children')) {
  plot <- ggplot(data = insurance, aes_string(x = feat, y = 'charges', group = 'smoker', fill = 'smoker
    geom_jitter() +
    geom_smooth(method = 'lm') +
    ggtitle(glue::glue("Charges vs {feat}"))
  print(plot)
}
```


Charges vs age
charges
age
60000
40000
20000
0
20
30
40
50
60
smoker
no
yes


Charges vs bmi
charges
bmi
0
20000
40000
60000
20
30
40
50
smoker
no
yes



Smoker seems to have the highest impact on medical charges, even though the charges are growing with age, bmi, and children. Also, people who have more children generally smoke less.

Training a model

Splitting the dataset into the training set and the test set

```
set.seed(123)
## Obtain the training index
training_index <- sample(seq_len(nrow(insurance)), size = floor(0.7 * nrow(insurance)))
## Partition the data
training_set <- insurance[training_index, ]
test_set <- insurance[-training_index, ]
```

Linear Regression

Fit Multiple Linear Regression to the training set

```
model1 <- lm(charges ~., data = training_set)
step(model1,direction = "backward")
```

```
## Start:  AIC=16351.25
## charges ~ age + sex + bmi + children + smoker + region
##
```

```
##            Df  Sum of Sq        RSS   AIC
## - region    3 1.6207e+08 3.5622e+10 16350
## - sex       1 1.4189e+07 3.5474e+10 16350
## <none>                   3.5460e+10 16351
## - children  1 6.3404e+08 3.6094e+10 16366
## - bmi       1 4.3342e+09 3.9794e+10 16457
## - age       1 1.0236e+10 4.5696e+10 16587
## - smoker    1 8.9871e+10 1.2533e+11 17531
##
## Step:  AIC=16349.52
## charges ~ age + sex + bmi + children + smoker
##
##            Df  Sum of Sq        RSS   AIC
## - sex       1 1.2082e+07 3.5634e+10 16348
## <none>                   3.5622e+10 16350
## - children  1 6.2869e+08 3.6251e+10 16364
## - bmi       1 4.3795e+09 4.0002e+10 16456
## - age       1 1.0300e+10 4.5923e+10 16585
## - smoker    1 9.0232e+10 1.2585e+11 17529
##
## Step:  AIC=16347.84
## charges ~ age + bmi + children + smoker
##
##            Df  Sum of Sq        RSS   AIC
## <none>                   3.5634e+10 16348
## - children  1 6.2421e+08 3.6258e+10 16362
## - bmi       1 4.3732e+09 4.0007e+10 16454
## - age       1 1.0323e+10 4.5957e+10 16584
## - smoker    1 9.0854e+10 1.2649e+11 17532
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker, data = training_set)
##
## Coefficients:
## (Intercept)          age          bmi     children    smokeryes
##    -12811.4        240.9        354.1        667.7      23925.1
```

```r
linear_reg <- lm(formula = charges ~ age + bmi + children + smoker, data = training_set)
y_pred <- predict(linear_reg, test_set)
mae <- MAE(y_pred, test_set$charges)
rmse <- RMSE(y_pred, test_set$charges)
lin_reg <- cbind("MAE" = mae, "RMSE" = rmse)
lin_reg
```

```
##           MAE     RMSE
## [1,] 3997.499 5829.191
```

Polynomial Regression

We can improve our model by feature engineering, specifically, by making new features that capture the interactions between existing features. This is called polynomial regression. The idea is to generate a new feature matrix consisting of all polynomial combinations of the features with degrees less than or equal to the specified degree. For example, if an input sample is two-dimensional and of the form [a, b], the degree-2 polynomial features are [1, a, b, a^2, ab, b^2]. We will use degree 2. We don't want charges to be included in the process of generating the polynomial combinations, so we take out charges from train and test and save them as y_train and y_test, respectively.

```
y_train <- training_set$charges
y_test <- test_set$charges
```

From EDA we know that sex and region have no correlation with charges. We can drop them. Also, since polynomial combinations don't make sense to categorical features, we mutate smoker as numeric.

```
X_train <- training_set %>%
  select(-c(charges, sex, region)) %>%
  mutate(smoker = as.numeric(smoker))
X_test <- test_set %>%
  select(-c(charges, sex, region)) %>%
  mutate(smoker = as.numeric(smoker))
```

We use the formula below to apply polynomial combinations.

```
formula <- as.formula(
  paste(
    ' ~ .^2 + ',
    paste('poly(', colnames(X_train), ', 2, raw=TRUE)[, 2]', collapse = ' + ')
  )
)
```

Then, insert y_train and y_test back to the new datasets.

```
train_poly <- as.data.frame(model.matrix(formula, data = X_train))
test_poly <- as.data.frame(model.matrix(formula, data = X_test))
train_poly$charges <- y_train
test_poly$charges <- y_test
colnames(train_poly)
```

```
##  [1] "(Intercept)"                         "age"
##  [3] "bmi"                                 "children"
##  [5] "smoker"                              "poly(age, 2, raw = TRUE)[, 2]"
##  [7] "poly(bmi, 2, raw = TRUE)[, 2]"       "poly(children, 2, raw = TRUE)[, 2]"
##  [9] "poly(smoker, 2, raw = TRUE)[, 2]"    "age:bmi"
## [11] "age:children"                        "age:smoker"
## [13] "bmi:children"                        "bmi:smoker"
## [15] "children:smoker"                     "charges"
```

We can see that our new datasets train_poly and test_poly now have 16 columns:

1.(Intercept) is a column consists of constant 1, this is the constant term in the polynomial.

2.age , bmi , children , smoker are the original features.

3.age^2 , bmi^2 , children^2 , smoker^2 are the square of the original features.

4.age x bmi, age x children , age x smoker , bmi x children , bmi x smoker , children x smoker are six interactions between pairs of four features.

5.charges is the target feature.

We start with all features and work our way down using backward elimination.

```
polynom_regr <- lm(formula = charges ~ ., data = train_poly)
step(polynom_regr)
```

```
## Start:  AIC=15909.7
## charges ~ ‘(Intercept)‘ + age + bmi + children + smoker + ‘poly(age, 2, raw = TRUE)[, 2]‘ +
##     ‘poly(bmi, 2, raw = TRUE)[, 2]‘ + ‘poly(children, 2, raw = TRUE)[, 2]‘ +
##     ‘poly(smoker, 2, raw = TRUE)[, 2]‘ + ‘age:bmi‘ + ‘age:children‘ +
##     ‘age:smoker‘ + ‘bmi:children‘ + ‘bmi:smoker‘ + ‘children:smoker‘
##
##
## Step:  AIC=15909.7
## charges ~ ‘(Intercept)‘ + age + bmi + children + smoker + ‘poly(age, 2, raw = TRUE)[, 2]‘ +
##     ‘poly(bmi, 2, raw = TRUE)[, 2]‘ + ‘poly(children, 2, raw = TRUE)[, 2]‘ +
##     ‘age:bmi‘ + ‘age:children‘ + ‘age:smoker‘ + ‘bmi:children‘ +
##     ‘bmi:smoker‘ + ‘children:smoker‘
##
##
## Step:  AIC=15909.7
## charges ~ age + bmi + children + smoker + ‘poly(age, 2, raw = TRUE)[, 2]‘ +
##     ‘poly(bmi, 2, raw = TRUE)[, 2]‘ + ‘poly(children, 2, raw = TRUE)[, 2]‘ +
##     ‘age:bmi‘ + ‘age:children‘ + ‘age:smoker‘ + ‘bmi:children‘ +
##     ‘bmi:smoker‘ + ‘children:smoker‘
##
##                                         Df  Sum of Sq        RSS   AIC
## - ‘age:children‘                         1 2.2613e+05 2.1889e+10 15908
## - ‘age:smoker‘                           1 4.0077e+05 2.1889e+10 15908
## - ‘age:bmi‘                              1 2.3313e+07 2.1912e+10 15909
## - ‘children:smoker‘                      1 4.0453e+07 2.1929e+10 15909
## <none>                                                2.1889e+10 15910
## - age                                    1 5.3999e+07 2.1943e+10 15910
## - ‘bmi:children‘                         1 5.5391e+07 2.1944e+10 15910
## - ‘poly(children, 2, raw = TRUE)[, 2]‘  1 6.7860e+07 2.1957e+10 15911
## - children                               1 2.0242e+08 2.2091e+10 15916
## - ‘poly(bmi, 2, raw = TRUE)[, 2]‘       1 2.0624e+08 2.2095e+10 15916
## - bmi                                    1 4.0248e+08 2.2291e+10 15925
## - ‘poly(age, 2, raw = TRUE)[, 2]‘       1 4.1842e+08 2.2307e+10 15925
## - smoker                                 1 1.9224e+09 2.3811e+10 15986
## - ‘bmi:smoker‘                           1 1.2921e+10 3.4810e+10 16342
##
## Step:  AIC=15907.71
## charges ~ age + bmi + children + smoker + ‘poly(age, 2, raw = TRUE)[, 2]‘ +
##     ‘poly(bmi, 2, raw = TRUE)[, 2]‘ + ‘poly(children, 2, raw = TRUE)[, 2]‘ +
##     ‘age:bmi‘ + ‘age:smoker‘ + ‘bmi:children‘ + ‘bmi:smoker‘ +
##     ‘children:smoker‘
##
##                                         Df  Sum of Sq        RSS   AIC
## - ‘age:smoker‘                           1 3.8654e+05 2.1890e+10 15906
## - ‘age:bmi‘                              1 2.3186e+07 2.1912e+10 15907
## - ‘children:smoker‘                      1 4.0273e+07 2.1929e+10 15907
## <none>                                                2.1889e+10 15908
## - age                                  1 5.4444e+07 2.1944e+10 15908
## - ‘bmi:children‘                       1 5.6605e+07 2.1946e+10 15908
## - ‘poly(children, 2, raw = TRUE)[, 2]‘  1 6.7639e+07 2.1957e+10 15909
## - ‘poly(bmi, 2, raw = TRUE)[, 2]‘       1 2.0603e+08 2.2095e+10 15914
## - children                             1 2.2844e+08 2.2118e+10 15915
## - bmi                                  1 4.0289e+08 2.2292e+10 15923
## - ‘poly(age, 2, raw = TRUE)[, 2]‘       1 4.1835e+08 2.2307e+10 15923
## - smoker                               1 1.9223e+09 2.3811e+10 15984
## - ‘bmi:smoker‘                         1 1.2921e+10 3.4810e+10 16340
##
## Step:  AIC=15905.73
## charges ~ age + bmi + children + smoker + ‘poly(age, 2, raw = TRUE)[, 2]‘ +
##     ‘poly(bmi, 2, raw = TRUE)[, 2]‘ + ‘poly(children, 2, raw = TRUE)[, 2]‘ +
##     ‘age:bmi‘ + ‘bmi:children‘ + ‘bmi:smoker‘ + ‘children:smoker‘
##
##                                         Df  Sum of Sq        RSS   AIC
## - ‘age:bmi‘                            1 2.3160e+07 2.1913e+10 15905
## - ‘children:smoker‘                    1 3.9935e+07 2.1929e+10 15905
## <none>                                              2.1890e+10 15906
## - ‘bmi:children‘                       1 5.6410e+07 2.1946e+10 15906
## - age                                  1 5.8123e+07 2.1948e+10 15906
## - ‘poly(children, 2, raw = TRUE)[, 2]‘  1 6.7744e+07 2.1957e+10 15907
## - ‘poly(bmi, 2, raw = TRUE)[, 2]‘       1 2.0609e+08 2.2096e+10 15912
## - children                             1 2.2805e+08 2.2118e+10 15913
## - bmi                                  1 4.0322e+08 2.2293e+10 15921
## - ‘poly(age, 2, raw = TRUE)[, 2]‘       1 4.1820e+08 2.2308e+10 15921
## - smoker                               1 2.4315e+09 2.4321e+10 16002
## - ‘bmi:smoker‘                         1 1.2929e+10 3.4818e+10 16338
##
## Step:  AIC=15904.72
## charges ~ age + bmi + children + smoker + ‘poly(age, 2, raw = TRUE)[, 2]‘ +
##     ‘poly(bmi, 2, raw = TRUE)[, 2]‘ + ‘poly(children, 2, raw = TRUE)[, 2]‘ +
##     ‘bmi:children‘ + ‘bmi:smoker‘ + ‘children:smoker‘
##
##                                         Df  Sum of Sq        RSS   AIC
## - age                                  1 3.5938e+07 2.1949e+10 15904
## - ‘children:smoker‘                    1 3.8834e+07 2.1952e+10 15904
## <none>                                              2.1913e+10 15905
## - ‘bmi:children‘                       1 5.5577e+07 2.1968e+10 15905
## - ‘poly(children, 2, raw = TRUE)[, 2]‘  1 6.9525e+07 2.1982e+10 15906
## - ‘poly(bmi, 2, raw = TRUE)[, 2]‘       1 1.9787e+08 2.2111e+10 15911
## - children                             1 2.2751e+08 2.2140e+10 15912
## - bmi                                  1 3.8112e+08 2.2294e+10 15919
## - ‘poly(age, 2, raw = TRUE)[, 2]‘       1 4.4362e+08 2.2356e+10 15922
## - smoker                               1 2.4186e+09 2.4331e+10 16001
## - ‘bmi:smoker‘                         1 1.2906e+10 3.4819e+10 16336
##
## Step:  AIC=15904.25
## charges ~ bmi + children + smoker + ‘poly(age, 2, raw = TRUE)[, 2]‘ +
##     ‘poly(bmi, 2, raw = TRUE)[, 2]‘ + ‘poly(children, 2, raw = TRUE)[, 2]‘ +
##     ‘bmi:children‘ + ‘bmi:smoker‘ + ‘children:smoker‘
##
##                                         Df  Sum of Sq        RSS   AIC
## - ‘children:smoker‘                    1 3.6995e+07 2.1986e+10 15904
## <none>                                              2.1949e+10 15904
## - ‘poly(children, 2, raw = TRUE)[, 2]‘  1 5.3253e+07 2.2002e+10 15904
## - ‘bmi:children‘                        1 5.3726e+07 2.2002e+10 15904
## - ‘poly(bmi, 2, raw = TRUE)[, 2]‘       1 1.9470e+08 2.2143e+10 15910
## - children                              1 2.0651e+08 2.2155e+10 15911
## - bmi                                   1 3.8752e+08 2.2336e+10 15919
## - smoker                                1 2.4430e+09 2.4392e+10 16001
## - ‘poly(age, 2, raw = TRUE)[, 2]‘       1 1.1763e+10 3.3712e+10 16304
## - ‘bmi:smoker‘                          1 1.2957e+10 3.4906e+10 16336
##
## Step:  AIC=15903.83
## charges ~ bmi + children + smoker + ‘poly(age, 2, raw = TRUE)[, 2]‘ +
##     ‘poly(bmi, 2, raw = TRUE)[, 2]‘ + ‘poly(children, 2, raw = TRUE)[, 2]‘ +
##     ‘bmi:children‘ + ‘bmi:smoker‘
##
##                                       Df  Sum of Sq        RSS   AIC
## - ‘poly(children, 2, raw = TRUE)[, 2]‘  1 4.2748e+07 2.2028e+10 15904
## <none>                                              2.1986e+10 15904
## - ‘bmi:children‘                        1 5.1059e+07 2.2037e+10 15904
## - children                              1 1.7098e+08 2.2157e+10 15909
## - ‘poly(bmi, 2, raw = TRUE)[, 2]‘       1 2.0714e+08 2.2193e+10 15911
## - bmi                                   1 3.7270e+08 2.2358e+10 15918
## - smoker                                1 2.5971e+09 2.4583e+10 16006
## - ‘poly(age, 2, raw = TRUE)[, 2]‘       1 1.1734e+10 3.3719e+10 16302
## - ‘bmi:smoker‘                          1 1.2922e+10 3.4908e+10 16335
##
## Step:  AIC=15903.65
## charges ~ bmi + children + smoker + ‘poly(age, 2, raw = TRUE)[, 2]‘ +
##     ‘poly(bmi, 2, raw = TRUE)[, 2]‘ + ‘bmi:children‘ + ‘bmi:smoker‘
##
##                                  Df  Sum of Sq        RSS   AIC
## - ‘bmi:children‘                   1 4.0172e+07 2.2069e+10 15903
## <none>                                         2.2028e+10 15904
## - children                         1 1.2823e+08 2.2157e+10 15907
## - ‘poly(bmi, 2, raw = TRUE)[, 2]‘  1 1.9649e+08 2.2225e+10 15910
## - bmi                              1 3.9032e+08 2.2419e+10 15918
## - smoker                           1 2.5852e+09 2.4614e+10 16006
## - ‘poly(age, 2, raw = TRUE)[, 2]‘  1 1.1720e+10 3.3748e+10 16301
## - ‘bmi:smoker‘                     1 1.2913e+10 3.4941e+10 16334
##
## Step:  AIC=15903.35
## charges ~ bmi + children + smoker + ‘poly(age, 2, raw = TRUE)[, 2]‘ +
##     ‘poly(bmi, 2, raw = TRUE)[, 2]‘ + ‘bmi:smoker‘
##
##                                  Df  Sum of Sq        RSS   AIC
## <none>                                         2.2069e+10 15903
## - ‘poly(bmi, 2, raw = TRUE)[, 2]‘  1 1.9459e+08 2.2263e+10 15910
## - bmi                              1 4.1728e+08 2.2486e+10 15919
## - children                         1 7.6526e+08 2.2834e+10 15933
## - smoker                           1 2.5463e+09 2.4615e+10 16004
## - ‘poly(age, 2, raw = TRUE)[, 2]‘  1 1.1711e+10 3.3779e+10 16300
## - ‘bmi:smoker‘                     1 1.2911e+10 3.4980e+10 16332


##
```

```
## Call:
## lm(formula = charges ~ bmi + children + smoker + ‘poly(age, 2, raw = TRUE)[, 2]‘ +
##     ‘poly(bmi, 2, raw = TRUE)[, 2]‘ + ‘bmi:smoker‘, data = train_poly)
##
## Coefficients:
##                     (Intercept)                              bmi
##                       13171.245                         -871.030
##                        children                           smoker
##                         739.339                       -19907.206
## ‘poly(age, 2, raw = TRUE)[, 2]‘  ‘poly(bmi, 2, raw = TRUE)[, 2]‘
##                           3.214                           -8.633
##                    ‘bmi:smoker‘
##                        1428.570
```

```
lm_poly <- lm(formula = charges ~ bmi + children + smoker + `poly(age, 2, raw = TRUE)[, 2]` +
    `poly(bmi, 2, raw = TRUE)[, 2]` + `poly(children, 2, raw = TRUE)[, 2]` +
    `bmi:smoker` + `children:smoker`, data = train_poly)
y_pred <- predict(lm_poly, test_poly)
mae1 <- MAE(y_pred, test_set$charges)
rmse1 <- RMSE(y_pred, test_set$charges)


poly_reg <- cbind("MAE" = mae1, "RMSE" = rmse1)
poly_reg
```

```
##           MAE    RMSE
## [1,] 2767.815 4680.93
```

Summary of the two models

```
summary(linear_reg)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker, data = training_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.4  -3089.2   -897.7   1721.3  29887.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12811.39    1164.39 -11.003  < 2e-16 ***
## age            240.89      14.67  16.423  < 2e-16 ***
## bmi            354.14      33.13  10.689  < 2e-16 ***
## children       667.68     165.34   4.038 5.83e-05 ***
## smokeryes    23925.08     491.07  48.721  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 6187 on 931 degrees of freedom
## Multiple R-squared:  0.7465, Adjusted R-squared:  0.7455
## F-statistic: 685.6 on 4 and 931 DF,  p-value: < 2.2e-16
```

We have four features, all of which are significant (has a real effect, not due to random chance and sampling) on charges. From the coefficients, we know that a non-smoker zero years old who has no children and zero BMI will be charged -$12811 by health insurance (which we know this scenario is impossible). Also, since smoker has the biggest coefficient of all features, a unit change in smoker gives a bigger change in charges than a unit change in other features give, given all other features are fixed. In this case, given all other features are fixed, a non-smoker would have less charge than a smoker by $23,925, which makes sense. This model also has 0.7455 adjusted R-squared, which means the model with its features explains 74% of the total variation in charges.

```
summary(lm_poly)
```

```
##
## Call:
## lm(formula = charges ~ bmi + children + smoker + ‘poly(age, 2, raw = TRUE)[, 2]‘ +
##     ‘poly(bmi, 2, raw = TRUE)[, 2]‘ + ‘poly(children, 2, raw = TRUE)[, 2]‘ +
##     ‘bmi:smoker‘ + ‘children:smoker‘, data = train_poly)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11907.7  -1854.0  -1218.0   -445.9  30166.6
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           1.269e+04  3.876e+03   3.273  0.00110 **
## bmi                                  -8.773e+02  2.086e+02  -4.206 2.85e-05 ***
## children                              1.635e+03  5.579e+02   2.930  0.00347 **
## smoker                               -1.955e+04  1.946e+03 -10.046  < 2e-16 ***
## ‘poly(age, 2, raw = TRUE)[, 2]‘       3.221e+00  1.448e-01  22.249  < 2e-16 ***
## ‘poly(bmi, 2, raw = TRUE)[, 2]‘      -8.596e+00  3.027e+00  -2.839  0.00462 **
## ‘poly(children, 2, raw = TRUE)[, 2]‘ -1.214e+02  9.286e+01  -1.307  0.19156
## ‘bmi:smoker‘                          1.431e+03  6.130e+01  23.343  < 2e-16 ***
## ‘children:smoker‘                    -4.079e+02  3.392e+02  -1.203  0.22943
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4872 on 927 degrees of freedom
## Multiple R-squared:  0.8435, Adjusted R-squared:  0.8422
## F-statistic: 624.6 on 8 and 927 DF,  p-value: < 2.2e-16
```

We have eight features, all of which are significant on charges, except for children:smoker. From the coefficients, we know that a non-smoker zero years old who has no children and zero BMI will be charged $12690 by health insurance (which we know this scenario is impossible). Also, since smoker has the biggest coefficient of all features, a unit change in smoker gives a bigger change in charges than a unit change in other features give, given all other features are fixed. In this case, given all other features are fixed, a non-smoker would have more charge than a smoker by $19950. The adjusted R-squared of this model is 0.8422, which means the model with its features explains 84% of the total variation in charges. In other words, this Polynomial Regression model captures more variance of charges than the earlier Linear Regression model.

Improving the model

The model can be improved by looking deeper into several features. For example, assuming that the increase in age and the expenses is not in a linear fashion; the older one gets, an even larger amount of expenses for medical is needed when compared to a year prior, so a square term of the age, age^2 is added.

Also, thinking that the bmi is affecting the medical expenses when one's bmi passes certain threshold value is considered being obese. Therefore, another new term, bmi30, is added by categorize the numerical value bmi feature into two portions; 0 for bmi below 30 and 1 for bmi above.

Finally, realizing that the increase in medical expenses is much higher for smoker than those with each unit incrase of bmi from the previous multiple regression model makes it reasonable to assume that the effect of smoking on medical expenses is a lot more and an obese smoker is spending even more on medical than it were for individual with obesity or is a smoker alone. Because individual is more proned to getting various healthy related issues when s/he is obese and smoking together. So an interaction term of bmi30*smoker is added in the improved model.

```
insurance$age2 <- insurance$age^2
## Add an indicator for BMI
insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)
## Partition the data again with the additional columns but using the same index
training_set_new <- insurance[training_index, ]
test_set_new <- insurance[-training_index, ]
## Create the final model
```

```
model_improv <- lm(charges~ age + bmi + children + smoker+age2+bmi30+bmi30*smoker, data=training_set_new
summary(model_improv)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + age2 +
##     bmi30 + bmi30 * smoker, data = training_set_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4517.8 -1725.9 -1194.8  -626.4 23563.8
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        222.7848  1642.1011   0.136  0.89211
## age                -72.7589    74.8051  -0.973  0.33098
## bmi                111.5999    39.3586   2.835  0.00468 **
## children           768.3147   126.7209   6.063 1.94e-09 ***
## smokeryes        13624.4469   511.7088  26.625  < 2e-16 ***
## age2                 4.1467     0.9342   4.439 1.01e-05 ***
## bmi30             -974.6593   504.4265  -1.932  0.05364 .
## smokeryes:bmi30 20058.0551   710.4371  28.233  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4463 on 928 degrees of freedom
## Multiple R-squared:  0.8685, Adjusted R-squared:  0.8675
## F-statistic: 875.8 on 7 and 928 DF,  p-value: < 2.2e-16
```

```
y_pred <- predict(model_improv, test_set_new)
mae2 <- MAE(y_pred, test_set_new$charges)
rmse2 <- RMSE(y_pred, test_set_new$charges)
lin_reg2 <- cbind("MAE" = mae2, "RMSE" = rmse2)
lin_reg2
```

```
##           MAE     RMSE
## [1,] 2454.077 4506.261
```

Relative to our previous models, the R-squared value has improved from 0.74 to 0.84 and now about 0.87. Our model is now explaining 86.75 percent of the variation in medical treatment costs. Additionally, our theories about the model's functional form seem to be validated. The higher-order age2 term is statistically significant, as is the obesity indicator, bmi30. The interaction between obesity and smoking suggests a massive effect; in addition to the increased costs of over $13,624 for smoking alone, obese smokers spend another $20058 per year. This may suggest that smoking exacerbates diseases associated with obesity.

Comparing models

```
compare <- cbind(c(lin_reg,lin_reg2,poly_reg))
compare
```

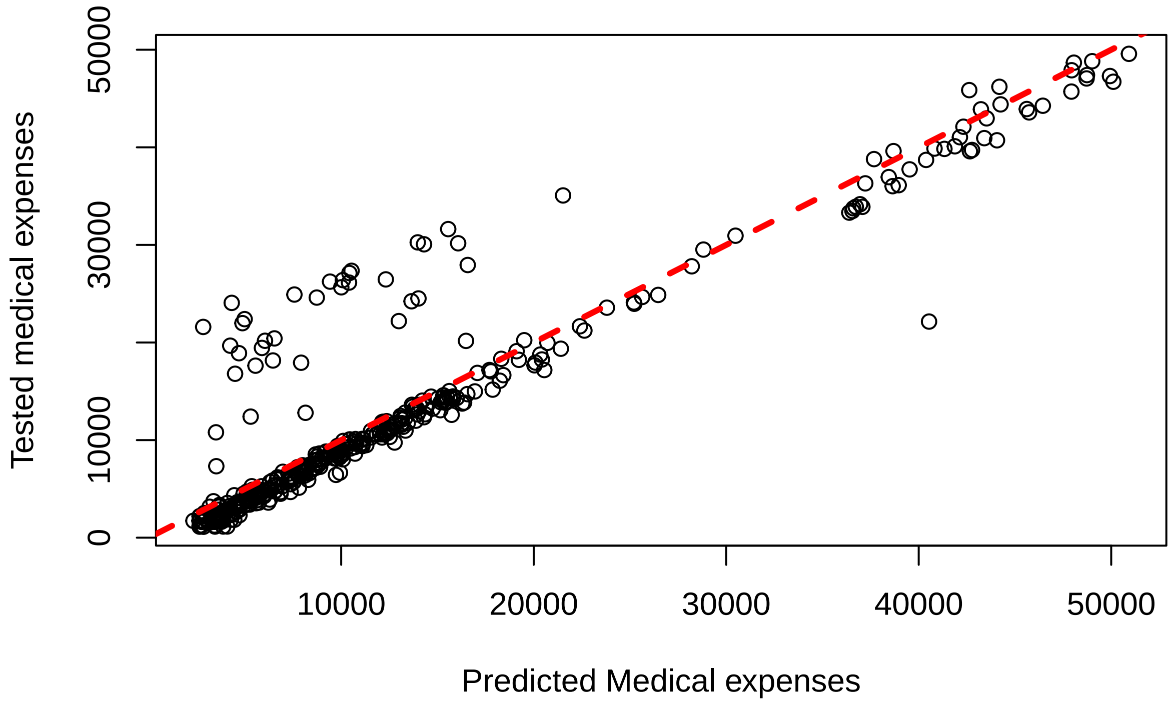
```
##          [,1]
## [1,] 3997.499
## [2,] 5829.191
## [3,] 2454.077
## [4,] 4506.261
## [5,] 2767.815
## [6,] 4680.930
```

Predictions

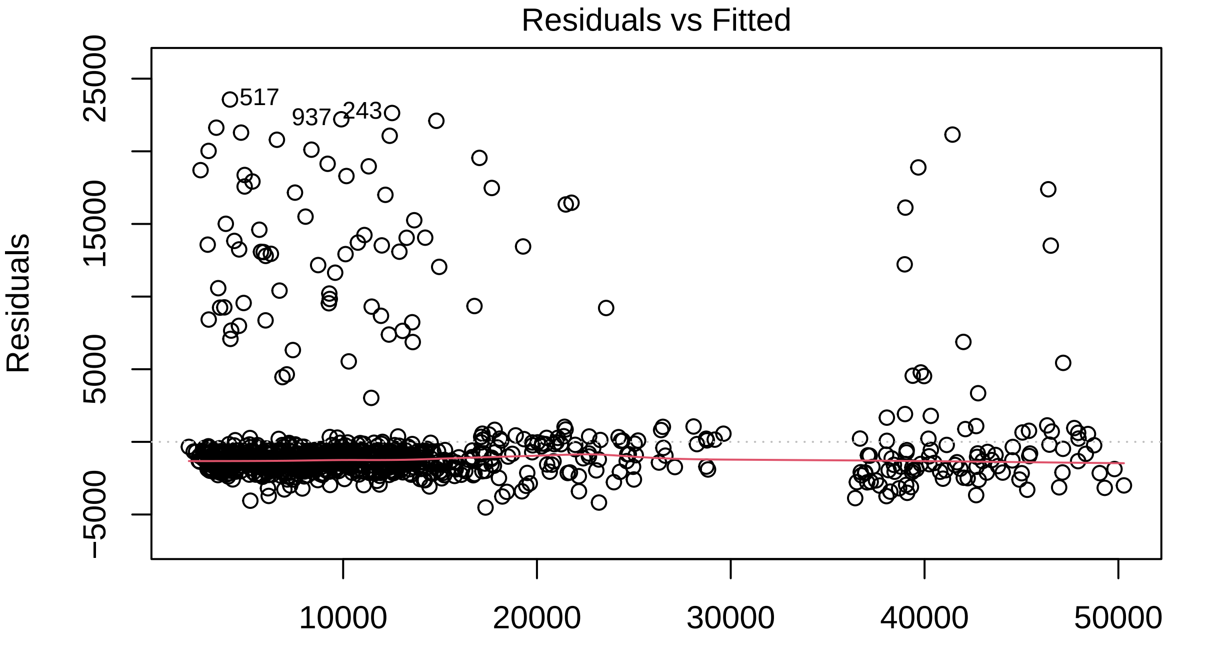
```
predictions = predict(model_improv, newdata = test_set_new)
cor(predictions, test_set_new$charges)
```

```
## [1] 0.9242925
```

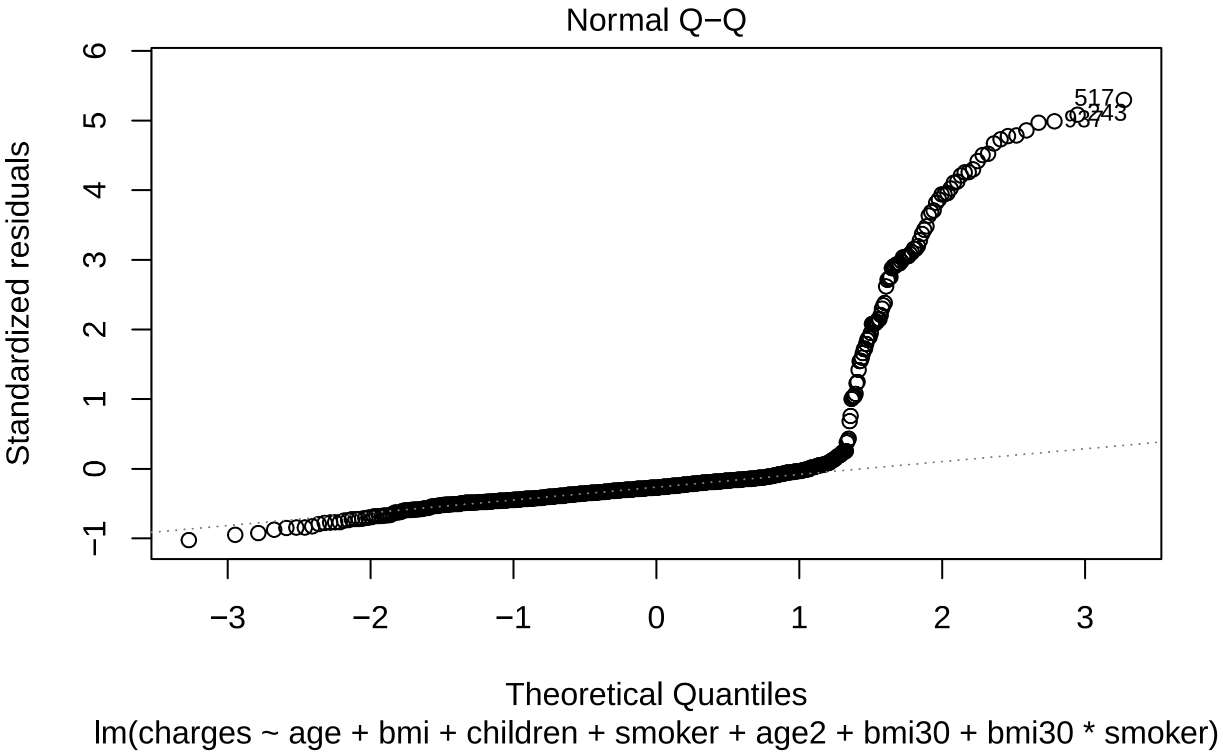
```
plot(predictions, test_set_new$charges,xlab="Predicted Medical expenses",ylab = "Tested medical expense
abline(a = 0, b = 1, col = 'red', lwd = 3, lty = 2)
```

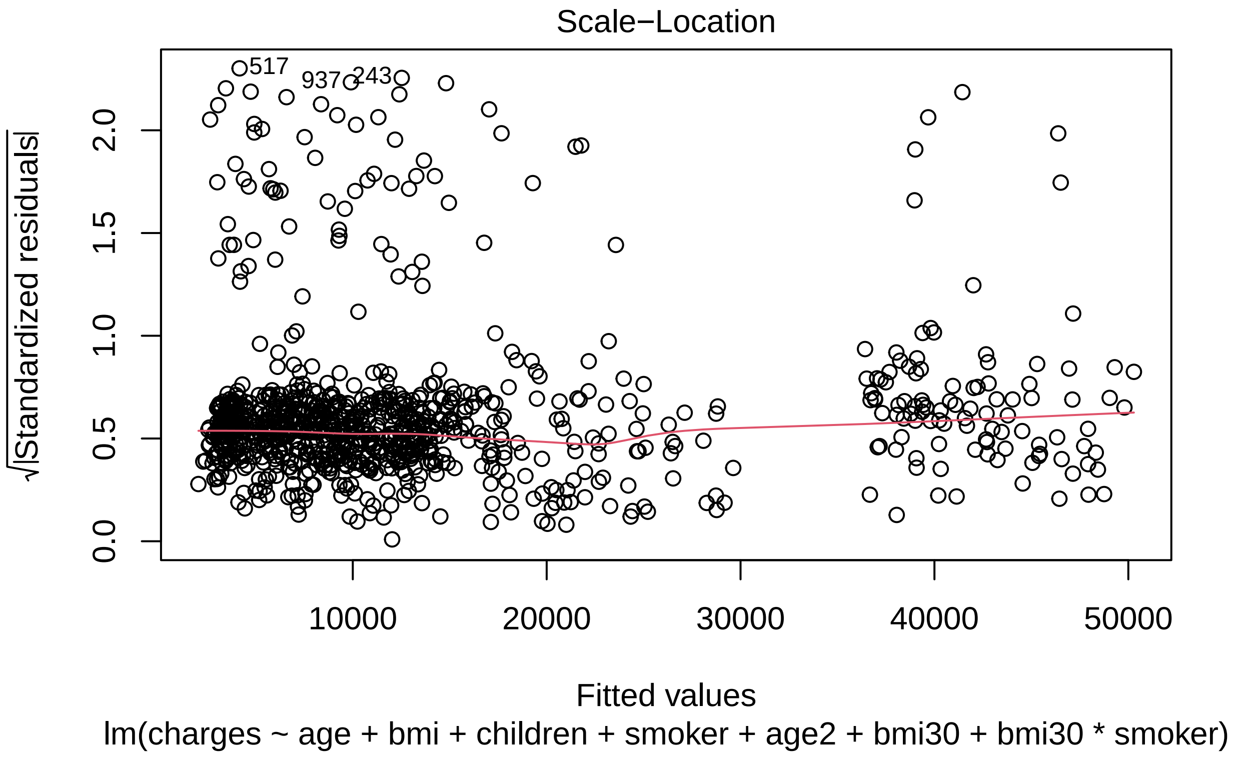


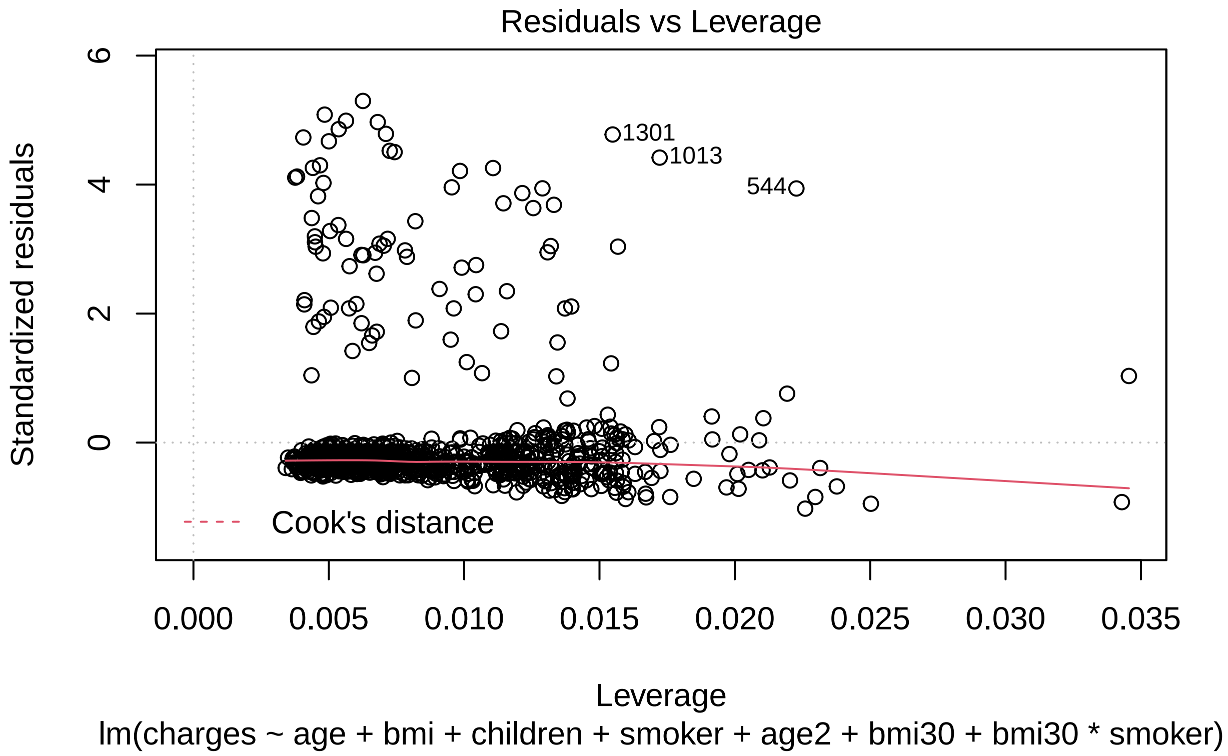
```
plot(model_improv)
```



lm(charges ~ age + bmi + children + smoker + age2 + bmi30 + bmi30 * smoker)


Normal Q−Q
Standardized residuals
−1
0
1
2
3
4
5
6
517
243
937
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(charges ~ age + bmi + children + smoker + age2 + bmi30 + bmi30 * smoker)


Scale−Location
517
937
243
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
10000
20000
30000
40000
50000
Fitted values
lm(charges ~ age + bmi + children + smoker + age2 + bmi30 + bmi30 * smoker)



We see that mosto of our graphs looks normal except the normality one with the Q-Q plot when we see the impact of the outliers which they are seem to be far away from the line.

```
predict(model_improv,
        data.frame(age = 18, age2 = 18^2, children = 0,
                   bmi = 42, sex = "male", bmi30 = 1,
                   smoker = "no", region = "northeast"))
```

```
##        1
## 3969.183
```

```
predict(model_improv,
        data.frame(age = 18, age2 = 18^2, children = 0,
                   bmi = 42, sex = "male", bmi30 = 1,
                   smoker = "yes", region = "northeast"))
```

```
##        1
## 37651.69
```

```
predict(model_improv,
        data.frame(age = 40, age2 = 40^2, children = 2,
                   bmi = 35, sex = "female", bmi30 = 1,
                   smoker = "no", region = "northeast"))
```

```
##        1
## 8415.069
```

```
predict(model_improv,
        data.frame(age = 40, age2 = 40^2, children = 2,
                   bmi = 35, sex = "female", bmi30 = 1,
                   smoker = "yes", region = "northeast"))
```

```
##        1
## 42097.57
```

```
predict(model_improv,
        data.frame(age = 80, age2 = 80^2, children = 4,
                   bmi = 35, sex = "male", bmi30 = 1,
                   smoker = "no", region = "northeast"))
```

```
##        1
## 26945.36
```

```
predict(model_improv,
        data.frame(age = 80, age2 = 80^2, children = 4,
                   bmi = 35, sex = "male", bmi30 = 1,
                   smoker = "yes", region = "northeast"))
```

```
##        1
## 60627.86
```

We predict the charges for 3 persons changing the parameter of smoking to "no" or "yes" and we can see the difference that is really significant .