Insurance Cost Prediction with Linear Regression Models - Group 1

Martin Hofwimmer, k01627988 - Mantas Bandonis, k01552238

June 2020

Abstract

The rising cost of health care is one of the world's most important problems. Accordingly, predicting such costs with accuracy is a significant first step in addressing this problem. Since the 1980s, there has been research on the predictive modeling of medical costs based on (health insurance) claims data using heuristic rules and regression methods.

A crisis such as Covid-19 affects all business sectors - but it especially puts a spotlight on insurers who can expect to be inundated with general inquiries and claims across multiple different lines, whether that be for health, life or non-life cover. Balancing the need for responding to this influx of activity in costs and remote workforce is an area that insurers are working to address.

We therefore, decided to do an analysis in order to find out how to better predict insurance charges for the general public, based on publicaly available data. We applied a simple OLS model with further improvements on the variables and found out that obesity and smokers are crucial parts of healthcare costs. With the additiona implications we were able to better predict the healthcare costs and therefore we say that insurance companies have to take significant factors into account. By predicting health care charges using linear regression methods, it is possible to impose different insurance premiums depending on the charge.

1. Introduction

This report documents the process and the results of our group project for the course Statistical Principles of Data Science. The task for the group project was to find a dataset with more than 500 observations and then to apply different statistical methods from the course.

In this report, first the dataset for our analysis is explained and the main objective of the analysis is described. In the methods section we describe how we proceed with our dataset and which methods we applied in order to predict the charge. Then the results of our analysis are presented. Finally, we summarize the main results and discuss the whole process.

For this group project we decided to work on the dataset "insurance.csv". This data set was found on kaggle.com and was uploaded in February 2018. The dataset consists of 1338 observations - thus it fulfils the requirement of a minimum of 500 observations.

The dataset consists of seven variables including different characteristics of insurance contractors as well as the charges they have to pay for the insurance. The goal of our analysis is to predict the charges based on the explanatory variables with a high accuracy.

The motivation for this work is that it is a real problem that insurance companies face in practice. Insurance companies invests a lot of time, effort, and money in creating models that accurately predicts health care costs. This is necessary because insurance companies must collect higher premium than the amount paid to the insured person in order to make profits.

2. Methods

In this group project we tried to predict the insurance charge with a high accuracy. Therfore we applied linear Regression and PCA. But before applying these methods data understanding is necessary and data cleaning and preparation must be done.

2.1 Data Understanding

The first step of our process is to import the dataset and analyse its structure. We do this with the head() and the str() command.

```
# Get first overview of data
head(my_data)
```

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

```
str(my_data)
```

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

The dataset contains 1338 observations of 7 variables. Two are stored as integers, two as numeric and the others are factors. At next we examine the data regarding data quality (accuracy, completeness, consistency, noise, and interpretability). This analysis is based on the guidelines in the course Data Mining (JKU Linz).

- **Accuracy: Do data reflect reality?** There are no typos and no impossible vaues, thus data might reflect the reality.
- **Completeness - Are all data available?** All data are available. There are no missing values in the dataset.
- **Consistency - Do data conform to (semantic) rules of the domain?** Yes, no violation of semantic rules of the domain could be identified.
- **Noise - Are there Outliers?** There are no values way out of the ranges.
- **Interpretability - Do I know what the data mean?** Yes a explanation for the variables is available. We know what the data means.

We can conclude that the data quality of this dataset is very high.

```
# Check for missings
colSums(is.na(my_data))
```

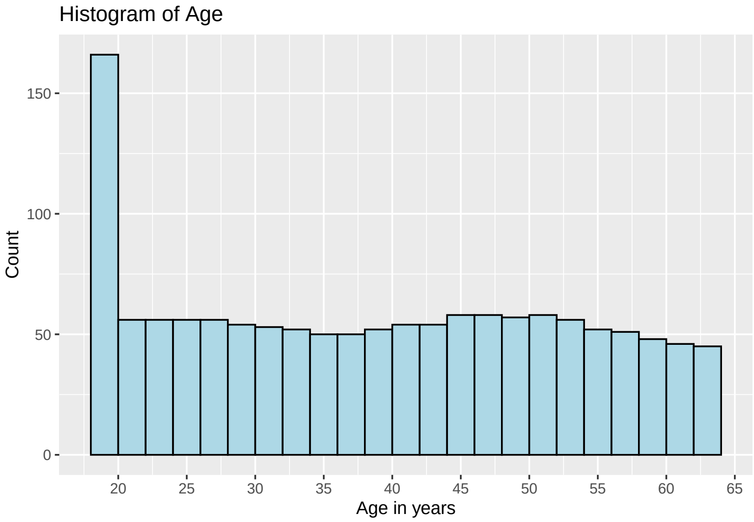
```
##      age      sex      bmi children   smoker   region  charges
##        0        0        0        0        0        0        0
```

2.2 Explanatory Data Analysis

In the next step Explanatory Data Analysis (EDA) is performed in order to get a better understanding of the variables. In the following section all variables are explained and analysed.

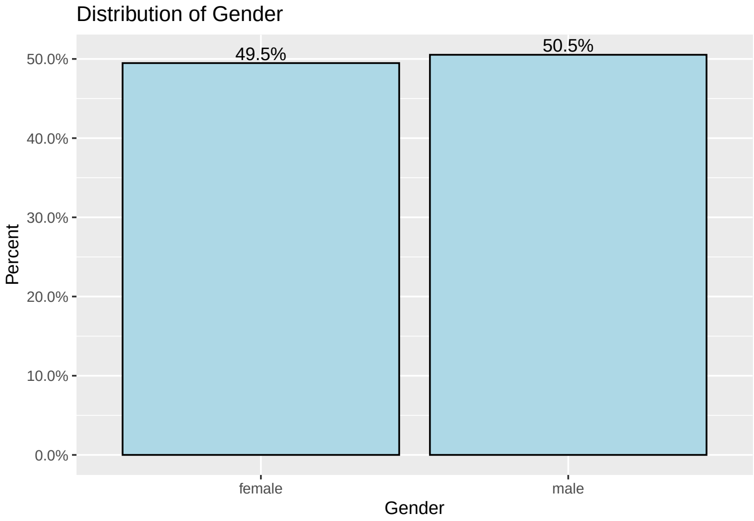
2.2.1 Age

The variable age indicates the age of the primary beneficiary in years. The ages in the sample ranges from 18 to 64 years. The mean age in the sample is 39.21, the median age is 39. As you can see in the histogram this variable is very equally distributed, except for the year 18 and 19. People of this age are overrepresented in the sample.



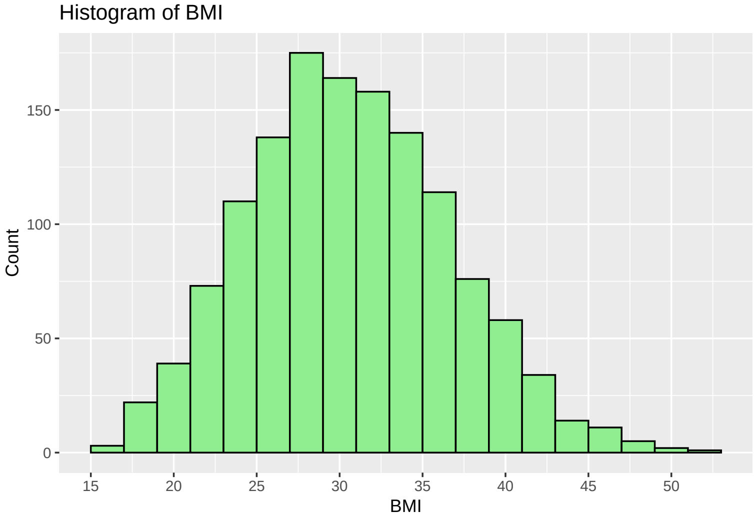
2.2.2 Sex

The variable sex indicates the gender of the insurance contractor. This variable can either be female or male. In the sample 49.5% of people are females, 50.5% are males.



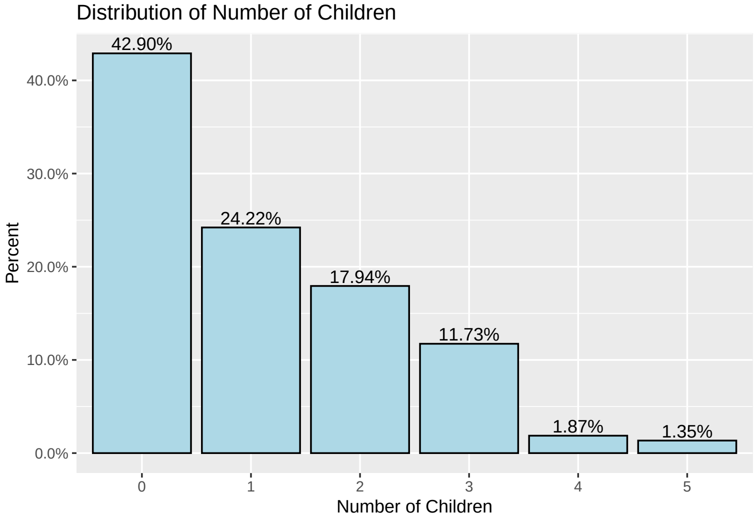
2.2.3 BMI

The variable bmi (Abbreviation for Body mass index) provides an understanding of the body. It is an objective index of body weight (unit kg / m ˆ 2) using the ratio of height to weight. Ideally the BMI lies between 18.5 and 24.9. The minimum BMI in the sample is 15.96, the maximum 53.13. The average Body mass index is 30.66, the Median is 30.4. This variable is normally distributed.



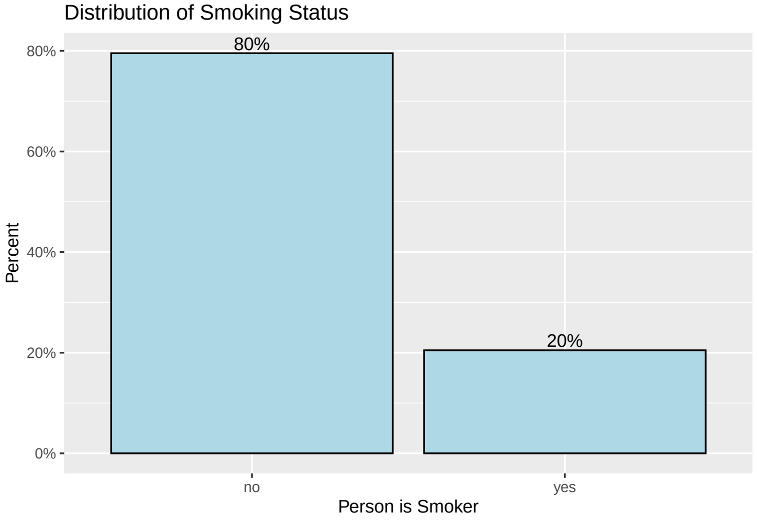
2.2.4 Children

The variable children shows the number of children that are covered by the health insurance (Number of dependents). In the dataset this number ranges from 0 to 5. For the majority of insurance contracters (42.9%) no children are covered by the health insurance. For 24.22% of insurance contractors one child is covered, for 17.94% two children are covered and for 11.73% three children are covered. Only for a small percentage 4 or 5 children are covered (1.87% and 1.35%)



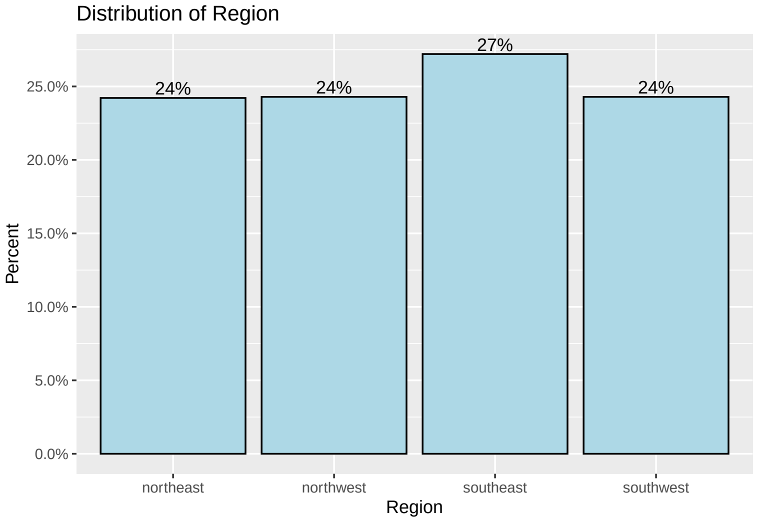
2.2.5 Smoker

This variable indicates whether the person is a dependent smoker or not. In the sample 20% of people are dependent smokers.



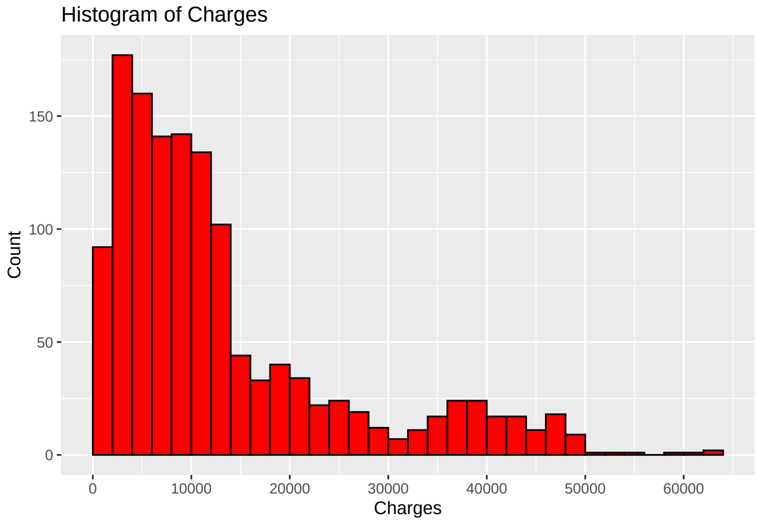
2.2.6 Region

The variable region indicates the beneficiary's residential area in the US. This variable can take the values northeast, southeast, southwest and northwest. The individuals are very equally distributed among these four regions. 24% live in northeast, northwest and southwest. A slightly larger percentage of 27% lives in southeast.



2.2.7 Charges

The variable charges describes the individual medical costs billed by the health insurance. This variable is of interest in our analysis (response variable in regression model). The minimum charge in the sample is USD 1,122, the maximum charge is USD 63,770. The average charge is USD 13,270. The median charge is USD 9,382 which means 50% of people have an charge lower, 50% higher than USD 9,382. It can be seen in the histogram that the distribution is skewed to the right.

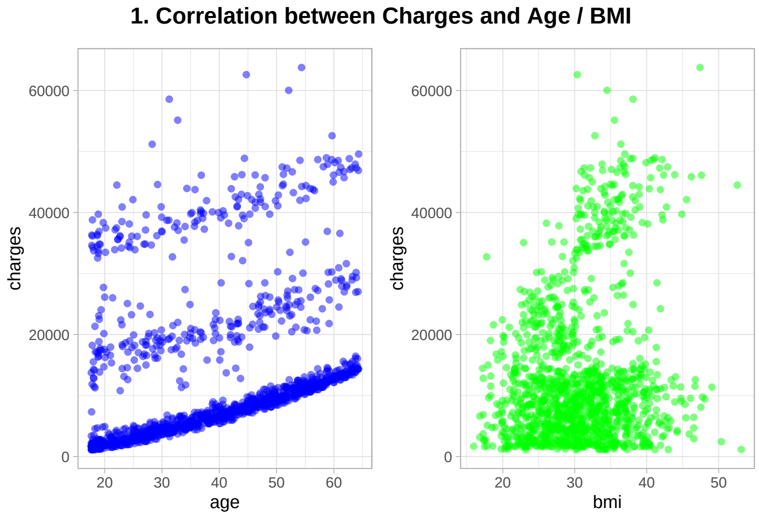


2.3 Analysing Associations of Charge

In the next step we want to analyse which variables influence the dependent variable charge. Therefore we analyse the joint distribution of charge with all other variables.

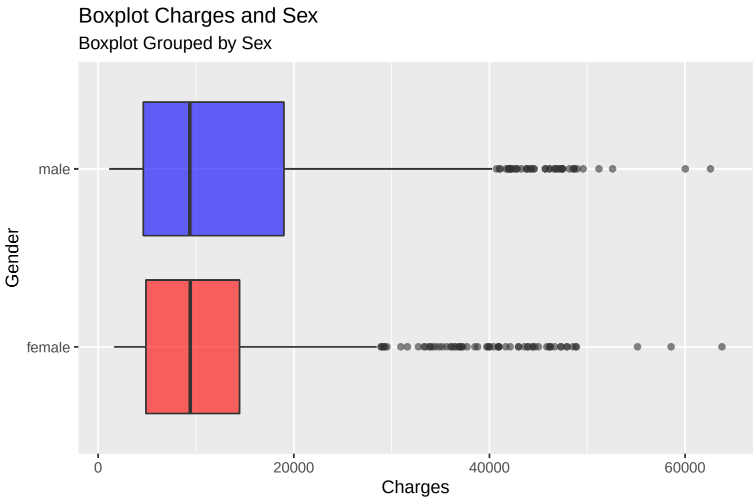
2.3.1 Charges and BMI/Age

The following plots show the joint distribution of the variables age and bmi with charge. It is observeable that there is a positive correlation between age and charges. Furthermore, it can be seen that the datapoints are not randomly distributed but there is a pattern. There is also a positive correlation between the body mass index and charge.



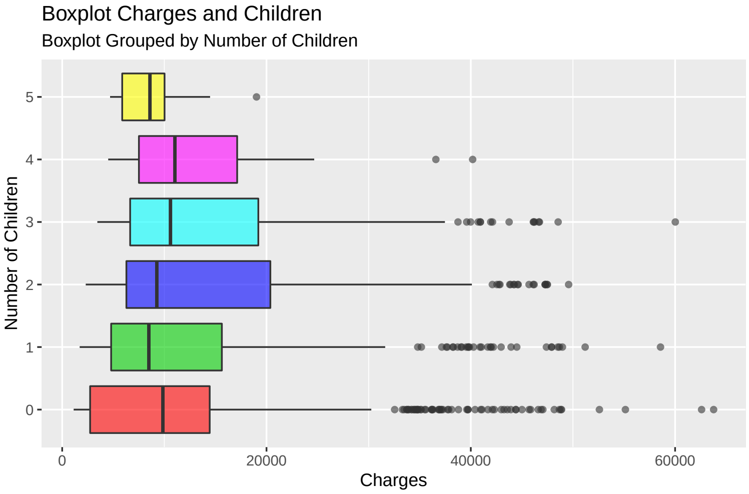
2.3.2 Charges and Sex

Analysing the distribution of charges grouped by sex it can be seen that they are very similar. However, the 75% quantile is lower for females. Furthermore, the interquartile range is higher for males.



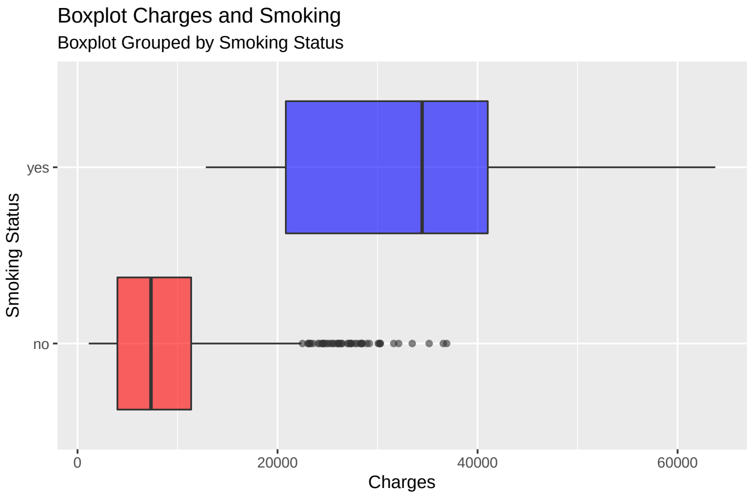
2.3.3 Charges and Number of Children

The number of children shows no obvious influence on charge.



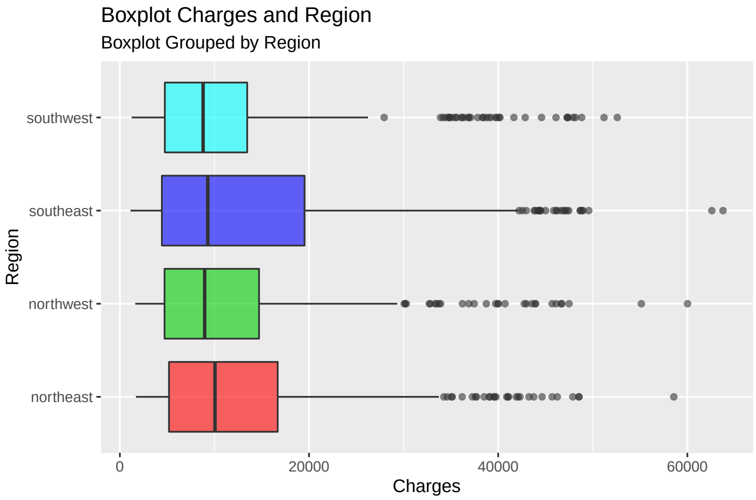
2.3.4 Charges and Smoking Status

As observable in the graph the smoking status has a large influence on the charges. Non-Smoker have an average charge of USD 8,437.27, Smoker an average charge of USD 32,050.23.



2.3.5 Charges and Region

The region seem to have no strong influence on the charge.



3. Data Preprocessing

Three categorical variables sex, smoker and region needed to be converted to numeric ones, the first two variables sex and smoker have binary categorical values so it could be converted directly, while the third variable 'region' has multi categorical values, so we needed to specify a specific numeric value for each categorical value, this will result in a new numeric column which will replace the categorical one, this result in a full numeric data set.

```r
#save cleaned/prepared dataset
write.csv(my_data_num,"insurance_cleaned.csv")
```

4. Regression

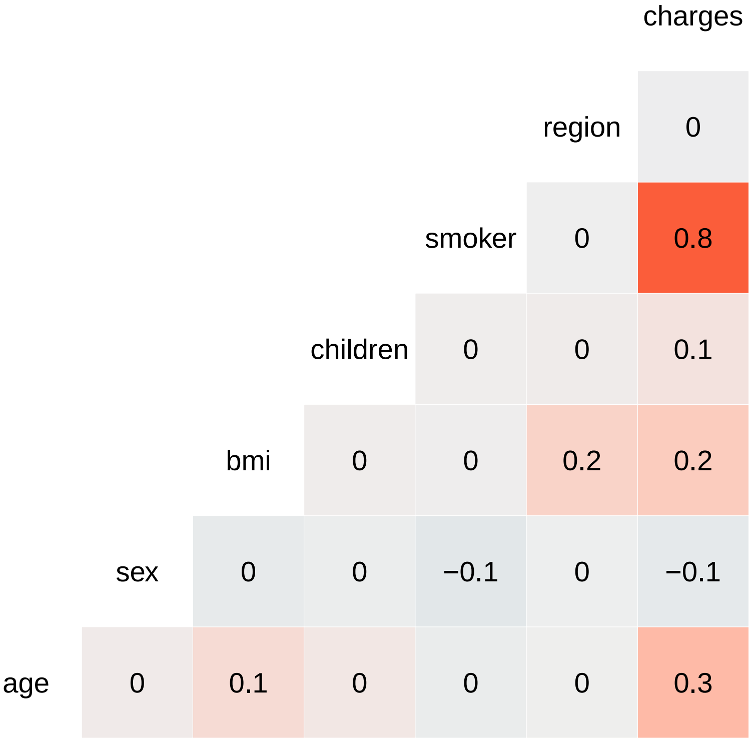
The regression was mostly done in Python.

4.1 Correlation Analysis

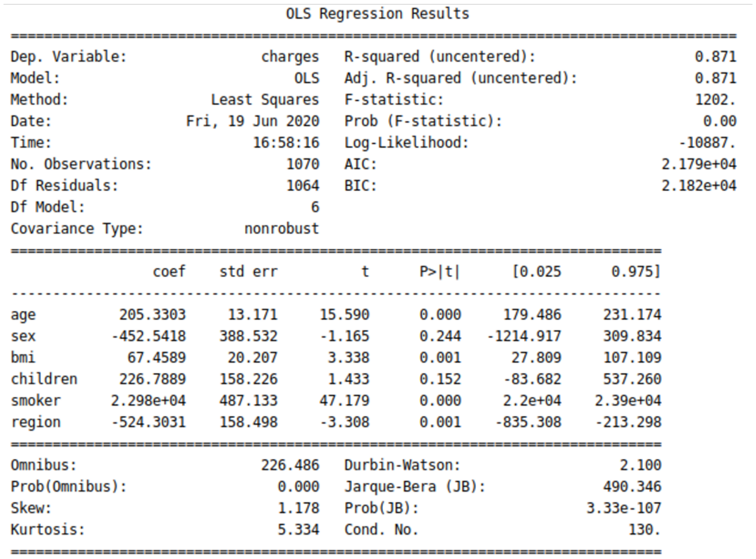
```
##          age          sex          bmi     children       smoker       region
##  0.299008193 -0.057292062  0.198340969  0.067998227  0.787251430 -0.006208235
##      charges
##  1.000000000
```

4.2 Linear Regression Model

Based on the graph, a strong correlation is observed only with the fact of smoking the patient, and by checking the correlation between the smoker variable and the other variables, we find there is not a considered correlation value, So there is no Collinearity, as a result we will use the all independent variables to train our model specially we have only 7 features in our data-set then we will check the R squared



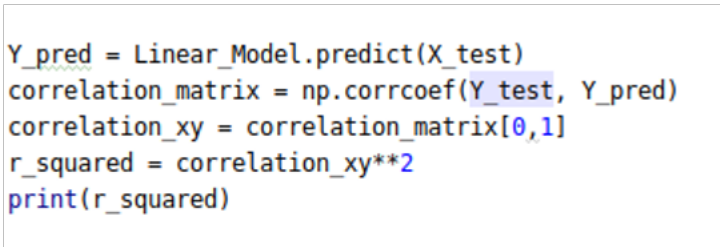
80% of the data into the training data and the rest should be the test data and the summary:



```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                charges   R-squared (uncentered):                   0.871
Model:                            OLS   Adj. R-squared (uncentered):              0.871
Method:                 Least Squares   F-statistic:                              1202.
Date:                Fri, 19 Jun 2020   Prob (F-statistic):                        0.00
Time:                        16:58:16   Log-Likelihood:                         -10887.
No. Observations:                1070   AIC:                                  2.179e+04
Df Residuals:                    1064   BIC:                                  2.182e+04
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
age          205.3303     13.171     15.590      0.000     179.486     231.174
sex         -452.5418    388.532     -1.165      0.244   -1214.917     309.834
bmi           67.4589     20.207      3.338      0.001      27.809     107.109
children     226.7889    158.226      1.433      0.152     -83.682     537.260
smoker      2.298e+04    487.133     47.179      0.000      2.2e+04    2.39e+04
region      -524.3031    158.498     -3.308      0.001    -835.308    -213.298
==============================================================================
Omnibus:                      226.486   Durbin-Watson:                   2.100
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              490.346
Skew:                           1.178   Prob(JB):                    3.33e-107
Kurtosis:                       5.334   Cond. No.                         130.
==============================================================================
```

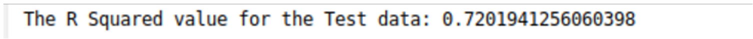
By checking the model summary we find:

The R squared value: As the R squared represents how close the data are to the fitted regression line, we find that our model achieves 87.1% in case of the training data.



```
Y_pred = Linear_Model.predict(X_test)
correlation_matrix = np.corrcoef(Y_test, Y_pred)
correlation_xy = correlation_matrix[0,1]
r_squared = correlation_xy**2
print(r_squared)
```

When we used the test data, the model achieves 72.01%.



```
The R Squared value for the Test data: 0.7201941256060398
```

The Coefficients values : The sign of a regression coefficient tells you whether there is a positive or negative correlation between each independent variable the dependent variable. A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase. A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease. in our summary, we found two negative coefficients, so it's recommended to eliminate their relative variables to enhance the results.

P-value : A p-value less than 0.05 (typically < 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, we reject the null hypothesis, and accept the alternative hypothesis. A p-value higher than 0.05 (> 0.05) is not statistically significant and indicates strong evidence for the null hypothesis. This means we retain the null hypothesis and reject the alternative hypothesis. You should note that you cannot accept the null hypothesis, we can only reject the null or fail to reject it. In our result, we got 4 variables $>$ .05 while 2 are greater than .05 and achieve the Null Hypothesis, in other words, they have no effect on the dependent variable.

4.2.1 Results

On our train data we got an RMSE value of 6213.810374050633.

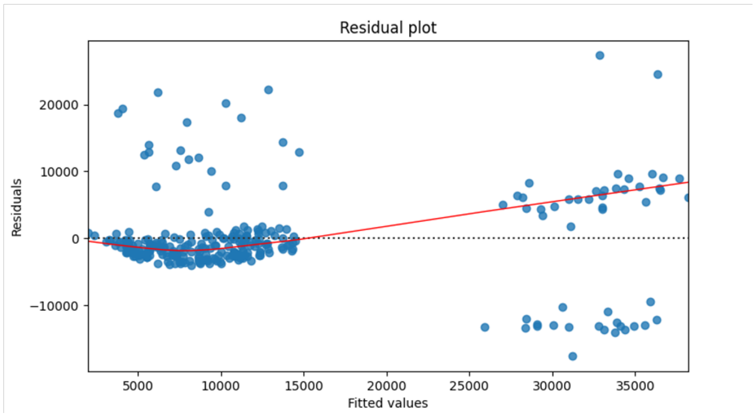
For test data wie achieved 7015.6140505048825.

RMSE measures the differences between values predicted by a hypothetical model and the observed values. In other words, it measures the quality of the fit between the actual data and the predicted model. RMSE is one of the most frequently used measures of the goodness of fit of generalized regression models. According to our RMSE results for both the training and the test data: we will notice that there is not a big difference in the result value, which means our model isn't over fitted, There might be a case of overfitting where you might get very low RMSE in training data but high RMSE in test data.

4.3 Model Validation

In order to validated model we need to check few assumption of linear regression model. The common assumption for Linear Regression model are the following:

Residual versus fitted value

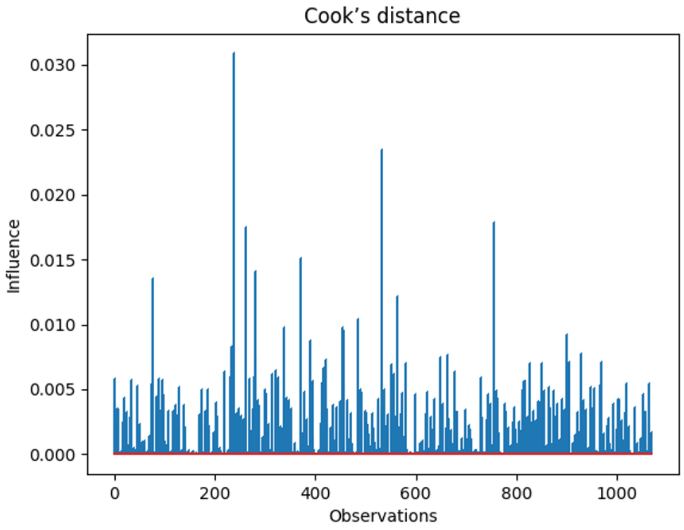


Residuals versus fits plot is the most frequently created plot. It is a scatter plot of residuals on the y axis and fitted values (estimated responses) on the x axis. The plot is used to detect non-linearity, unequal error variances, and outliers. In this plot (above) each point is one insurance value, where the prediction made by the model is on the x-axis, and the accuracy of the prediction is on the y-axis. The distance from the line at 0 is how bad the prediction was for that value.

Since... Residual = Observed − Predicted

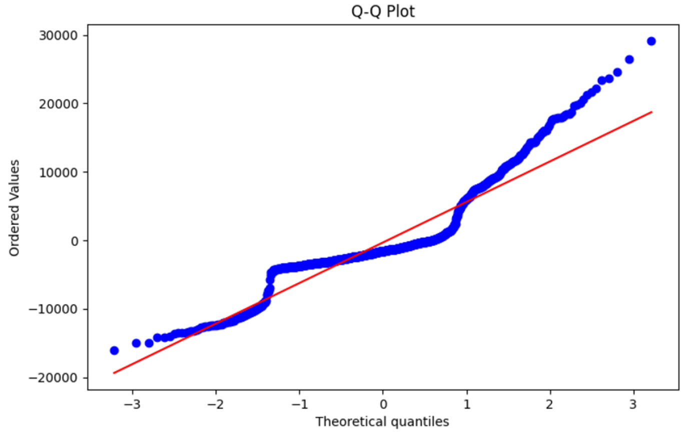
positive values for the residual (on the y-axis) mean the prediction was too low, and negative values mean the prediction was too high; 0 means the guess was exactly correct. According to the graph above, the homoscedasticity is not achieved, the variance in the error is not fixed according to the X axis, but it increases, which causes heteroscedasticity. Solution: we can transform the output variable (charges) by using the log to the current values, and repeat the process again.

Cook's distance



To find the influential outliers in a set of predictor variables. In other words, it's a way to identify points that negatively affect your regression model. The measurement is a combination of each observation's leverage and residual values; the higher the leverage and residuals, the higher the Cook's distance. To detect the outliers: Detect the threshold, which is used as an indicator to detect if the current point is an outlier or not. With the given equation: Threshold = 4 / n-k-1 , n: is the total number of observations, k is the explanatory variables: Threshold = 4 / (4/len(X_train) − 6 - 1) = 0.0038 The cut-off here is around 0.0038, any points above that cut-off might be considered as an outlier point.

Quantile-Quantile Plot



Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of your sample data. While Normal Q-Q Plots are the ones most often used in practice due to so many statistical methods assuming normality, Q-Q Plots can actually be created for any distribution, Also, it helps to determine if two data sets come from populations with a common distribution. In our graph, in some parts we can see that Y-values > X-values, which means that the x-quantiles are lower than the y-quantiles, and in other parts we can see that Y-values < X-values , which means that the y-quantiles are lower than the x-quantiles, in other wards, they come from different distribution.

In our next method PCA, we will try to get better a solution comparing to the current one which we got with the OLS.

4.4 Further improvements

In order to achieve better results it might be necessary to take other variables into account and combine them. We introduced an additonal variable which categorizes all people with a higher BMI than 30 into obeses “yes” and all people below “no”.

```
my_data$obese <- as.factor(ifelse(my_data$bmi >=30, "yes", "no"))
```

We split our data into train and test data, holding out 20% on the test set.

```
n_train <- round(0.8 * nrow(my_data))
train_indices <- sample(1:nrow(my_data), n_train)
Data_train <- my_data[train_indices, ]
Data_test <- my_data[-train_indices, ]

formula_0 <- as.formula("charges ~ age + sex + bmi + children + smoker + region")
```

Furthermore, we want to evaluate the model in a better way and look wich variables are important.

```
model1 <- lm(formula_0, data=Data_train)
summary(model1)
```

```
## 
## Call:
## lm(formula = formula_0, data = Data_train)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -11336.4  -2882.6   -992.6   1374.7  29854.6 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      -11418.21    1111.51 -10.273  < 2e-16 ***
## age                 251.97      13.39  18.821  < 2e-16 ***
## sexmale            -161.74     376.25  -0.430  0.66738    
## bmi                 336.17      31.99  10.509  < 2e-16 ***
## children            425.58     153.10   2.780  0.00554 ** 
## smokeryes         23945.72     466.73  51.305  < 2e-16 ***
## regionnorthwest    -540.39     533.79  -1.012  0.31159    
## regionsoutheast   -1172.06     539.62  -2.172  0.03008 *  
## regionsouthwest   -1254.67     544.56  -2.304  0.02142 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6117 on 1061 degrees of freedom
## Multiple R-squared:  0.749,  Adjusted R-squared:  0.7471 
## F-statistic: 395.8 on 8 and 1061 DF,  p-value: < 2.2e-16
```

Smokers increases health care costs (charges) by 23k per year. As the number of children increases, helath care costs(charges) can be increased by 0.5k. Also the BMI plays a significant role in the healthcare costs. We guess that the increase in dependents can increase the cost of care such as hospital care. How do you make a model if you want to give a higher penalty to an obese & smoke person? We also see that we have sim clearliny non-significant variables like sex and region, which dont affect the charges.

```
#Saving R-squared
r_sq_0 <- summary(model1)$r.squared
#predict data on test set
prediction_0 <- predict(model1, newdata = Data_test)
```

```
#calculating the residuals
residuals_0 <- Data_test$charges - prediction_0
#calculating Root Mean Squared Error
rmse_0 <- sqrt(mean(residuals_0^2))
```

```
print(paste0("RMSE for first model: ", round(rmse_0, 2)))
```

```
## [1] "RMSE for first model: 5849.8"
```

Furthermore, we build a model for the charges based on the variables of obesity and smoking.

```
model2 <- lm(charges ~ obese * smoker, data=Data_train)
summary(model2)
```

```
##
## Call:
## lm(formula = charges ~ obese * smoker, data = Data_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19536.6  -4238.1   -987.9   2904.6  27188.5
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7971.7      295.8  26.952   <2e-16 ***
## obeseyes               972.6      406.3   2.394   0.0168 *
## smokeryes            13453.2      656.2  20.503   <2e-16 ***
## obeseyes:smokeryes   19283.2      898.4  21.463   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5915 on 1066 degrees of freedom
## Multiple R-squared:  0.7642, Adjusted R-squared:  0.7635
## F-statistic:  1152 on 3 and 1066 DF,  p-value: < 2.2e-16
```

Obesity increases health care costs by about 1k Dollar, and smoking increases health care costs by 13,558 Dollar. But the both components are applied, (if smoking and obesity are together). It can be expected that medical expenses will increase the most with about 19k Dollar. Furthermore, we examine that the variable obese alone is not significant but plays a major role when combined with smoking. As a result of the model comparison above, by using * rather than +, the prediction of the model became more similar to reality.

```
r_sq_1 <- summary(model2)$r.squared

prediction_1 <- predict(model2, newdata = Data_test)

residuals_1 <- Data_test$charges - prediction_1
rmse_1 <- sqrt(mean(residuals_1^2))
```

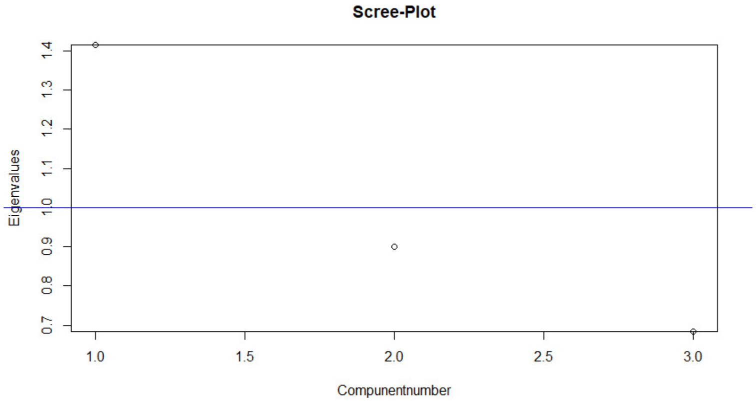
We can furthermore see an improvement in our RMSE score.

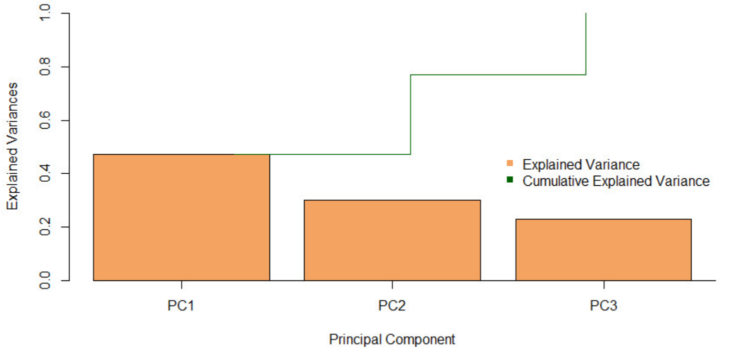
```
print(paste0("RMSE for new model: ", round(rmse_1, 2)))
```

```
## [1] "RMSE for new model: 5866.06"
```

5. PCA

The first two principal components account for 77.18% of the total variance. A scree graph of the eigenvalues can be plotted to visualize the proportion of variance explained by each subsequential eigenvalue. We're going to take PCA 1 and PCA 2.

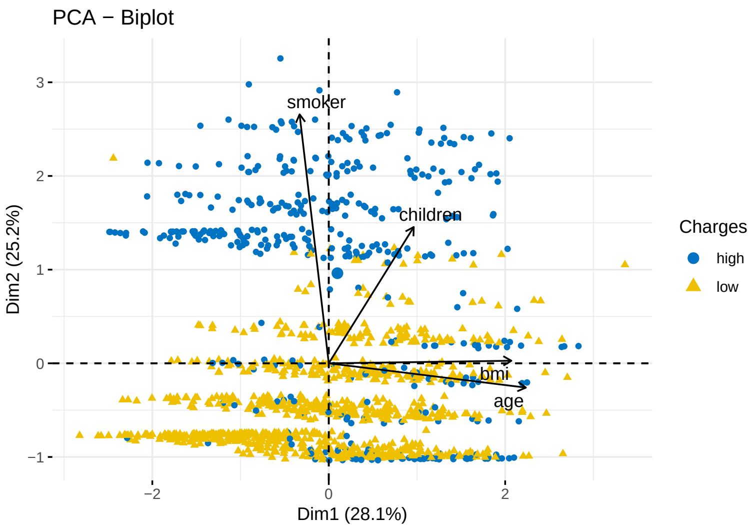




5.1 Further PCA consideration

We introduce an additional variable group, which categorizes the charges in high and low. Everything above the mean is considered as high and below as low.

We apply PCA with the features smoker, children, bmi and age. We can see both Principal components explaining about 50% of the data. Especically, the variable smoker is very extended in the second component pointing towards the high charges category.



6. Discussion

????