Progetto Statistics for Finance and Insurance

Gerardo Santonicola, Simone Manzolillo

28/10/2021

Contents

PREDICTION OF THE HEALTH INSURANCE COST

==========================

Abstract

In this project work we use different kind of regression models to predict the health insurance cost of a sample of 1338 people. We are going to use the linear, ridge and lasso regression model, comparing the results of the prediction.

Specs: required libraries

```
#install.packages(c("knitr", "ggplot2", "corrplot", "glmnet", "jpeg", "ggpubr"))
library(knitr)
library(ggplot2)
library(corrplot)
library(glmnet)
library(jpeg)
library(ggpubr)
```

1.1 ***Exploratory data analysis***

The dataset is available on GitHub. It is composed by 7 variables:

Sex: gender of the insurance contractor.

BMI: body mass index of the contractor.

Smoker: dicotomic variable that indicates if the contractor is a smoker or not.

Region: residential area in the US, with 4 observations, northeast, southeast, northwest, southwest.

Children: number of children of the contractor covered by health insurance.

Charges:medical costs incurred by insurance.

We import the dataset in R and analyse the structure, apporting the required modifies to the variables.

```
setwd("/Users/geralt/Desktop/Progetto Amendola/Progetto statistics 2")
insurance <- read.csv("insurance.csv") #import the data.
```

Table 1: Insurance data

age	sex	bmi	children	smoker	region	charges
19	female	27.900	0	yes	southwest	16884.924
18	male	33.770	1	no	southeast	1725.552
28	male	33.000	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.471
32	male	28.880	0	no	northwest	3866.855
31	female	25.740	0	no	southeast	3756.622
46	female	33.440	1	no	southeast	8240.590
37	female	27.740	3	no	northwest	7281.506
37	male	29.830	2	no	northeast	6406.411
60	female	25.840	0	no	northwest	28923.137
25	male	26.220	0	no	northeast	2721.321
62	female	26.290	0	yes	southeast	27808.725
23	male	34.400	0	no	southwest	1826.843
56	female	39.820	0	no	southeast	11090.718
27	male	42.130	0	yes	southeast	39611.758
19	male	24.600	1	no	southwest	1837.237
52	female	30.780	1	no	northeast	10797.336
23	male	23.845	0	no	northeast	2395.172
56	male	40.300	0	no	southwest	10602.385
30	male	35.300	0	yes	southwest	36837.467

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

```r
summary(insurance)
```

```
##       age             sex                 bmi           children    
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000  
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000  
##  Median :39.00   Mode  :character   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  
##     smoker             region             charges     
##  Length:1338        Length:1338        Min.   : 1122  
##  Class :character   Class :character   1st Qu.: 4740  
##  Mode  :character   Mode  :character   Median : 9382  
##                                        Mean   :13270  
##                                        3rd Qu.:16640  
##                                        Max.   :63770  
```

We can see that 3 variables are seen from R in the wrong way, we have to modify the type of variables, changing "sex", "smoker" and "region" in factor.

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

summary(insurance)
```

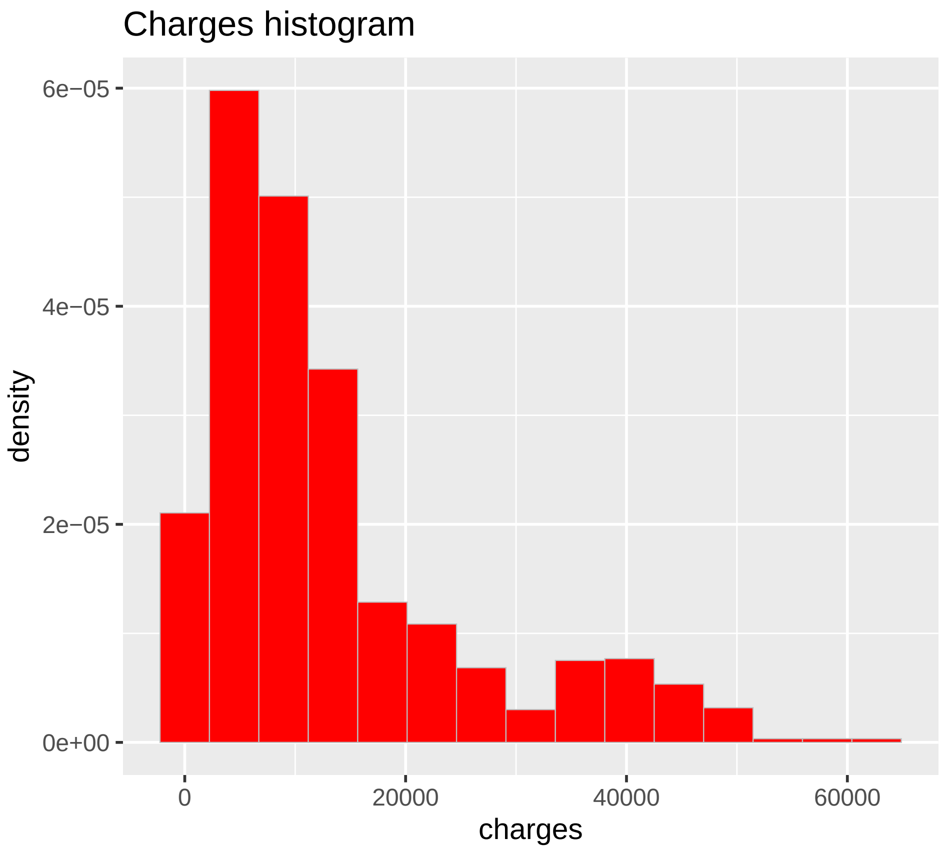
```
##       age             sex           bmi           children     smoker    
##  Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064  
##  1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274  
##  Median :39.00                Median :30.40   Median :1.000             
##  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  
```

1.2 *Visualization of the variables*

1.2.1 *Distribution of the response variable Charges*

First of all, we study the distribution of the response variable “Charges”. We use the ggplot library to represent the plots.

```
ggplot(insurance, aes(x = charges, y = ..density..)) +
  geom_histogram(bins = 15, fill = "red", color = "grey", size = 0.2)  +
  labs(title = "Charges histogram")
```



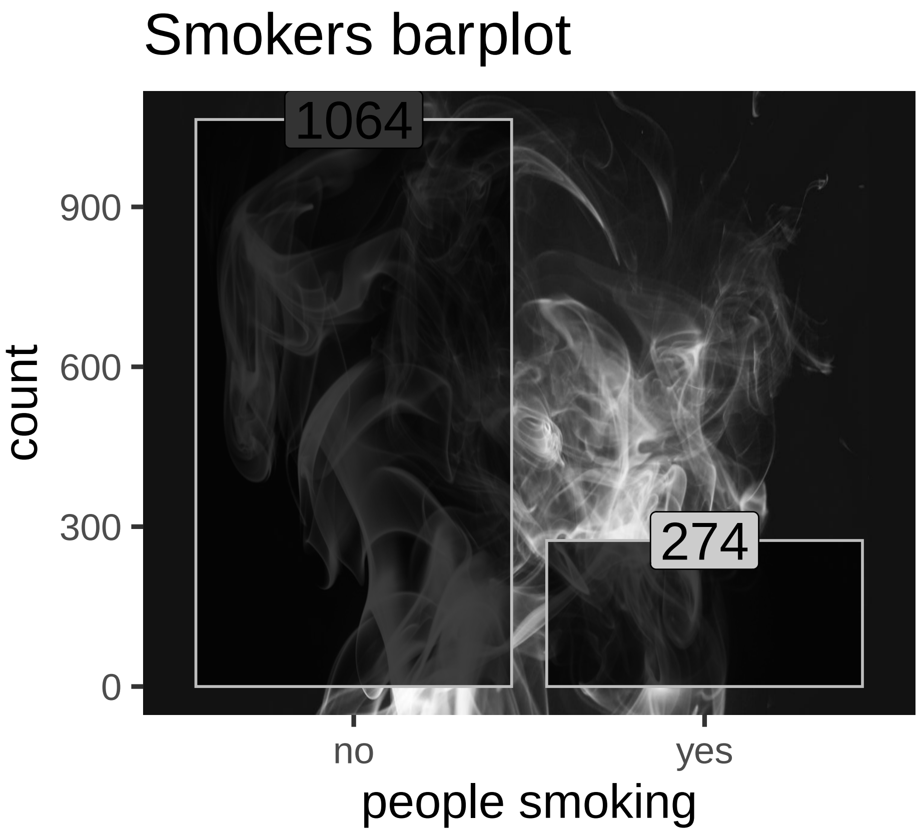
The histogram reveals that the charges variable is highly right skewed. This plot tells us that there are many outliers that have a greater insurance cost.

1.2.2 *Smoker variable*

Now we analyse the dependent variables of the dataset, starting from the smoker variable.

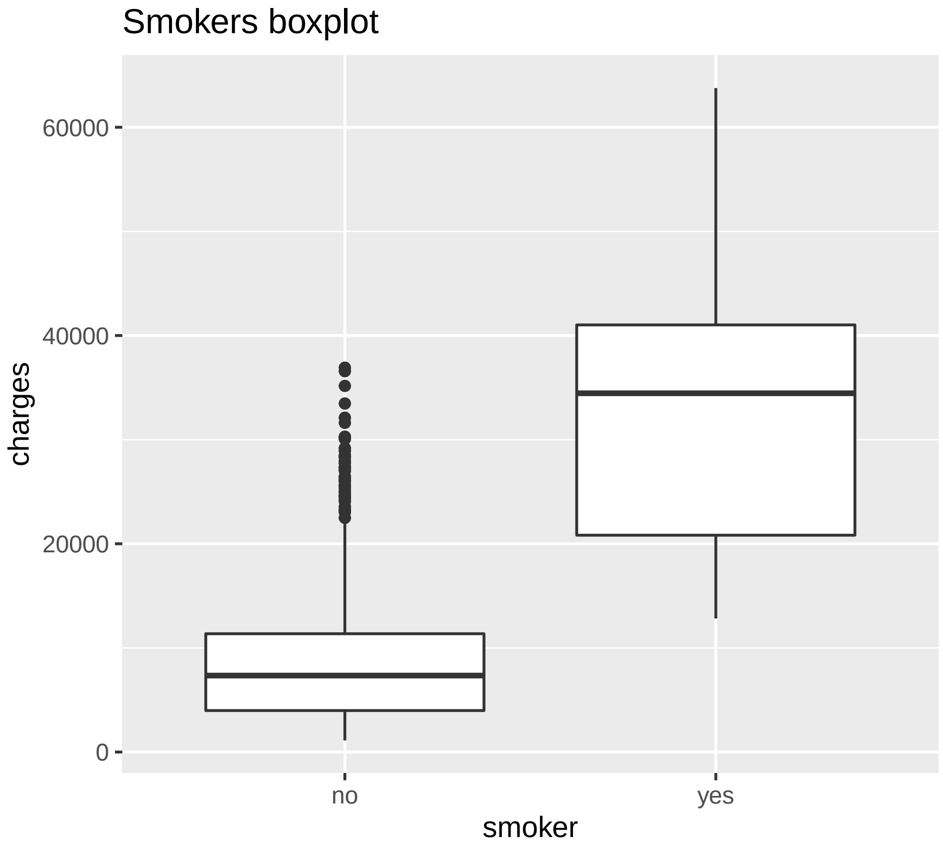
```
img.smoke <- readJPEG("liquid-smoke-06302016.jpeg") #import an image
#to use as wallpaper for the plot.

ggplot(insurance,aes(x=smoker,fill=smoker))+  background_image(img.smoke)+
  geom_bar(stat = 'count', alpha=0.75, colour="gray", fill="black")+
  labs(x = 'people smoking') +
  geom_label(stat='count',aes(label=..count..), size=7) +
  theme_grey(base_size = 18) + scale_fill_grey()+labs(title = "Smokers barplot") +
  theme ( legend.position = "none")
```



```
 #barplot that split the number of smokers from the non-smokers.
```

```
ggplot(insurance,aes(x=smoker,y=charges))+ geom_boxplot() +
  scale_fill_grey() + labs(title= "Smokers boxplot") +
  theme ( legend.position = "none") #boxplot indicating insurance costs based
```

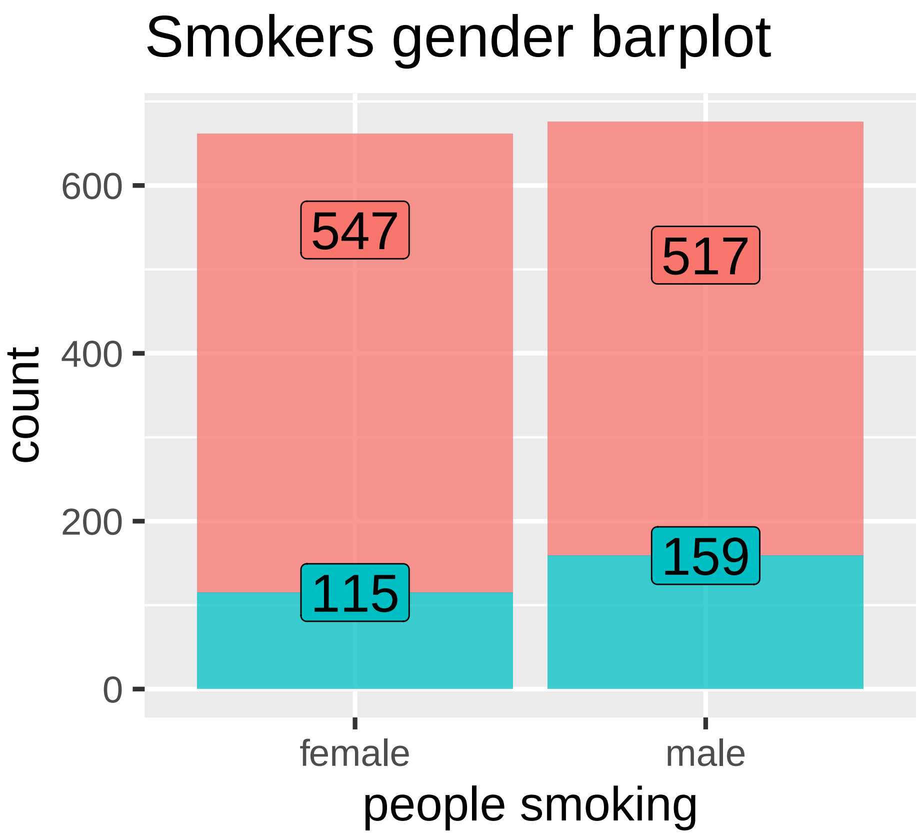


```
#on whether the policyholder is a smoker or not.
```

From the plots above we can see that, obviously, the charges for a smoker are higher than the insurance cost fort the non-smoker. In addiction we can see that smokers are less numerous than non-smokers.

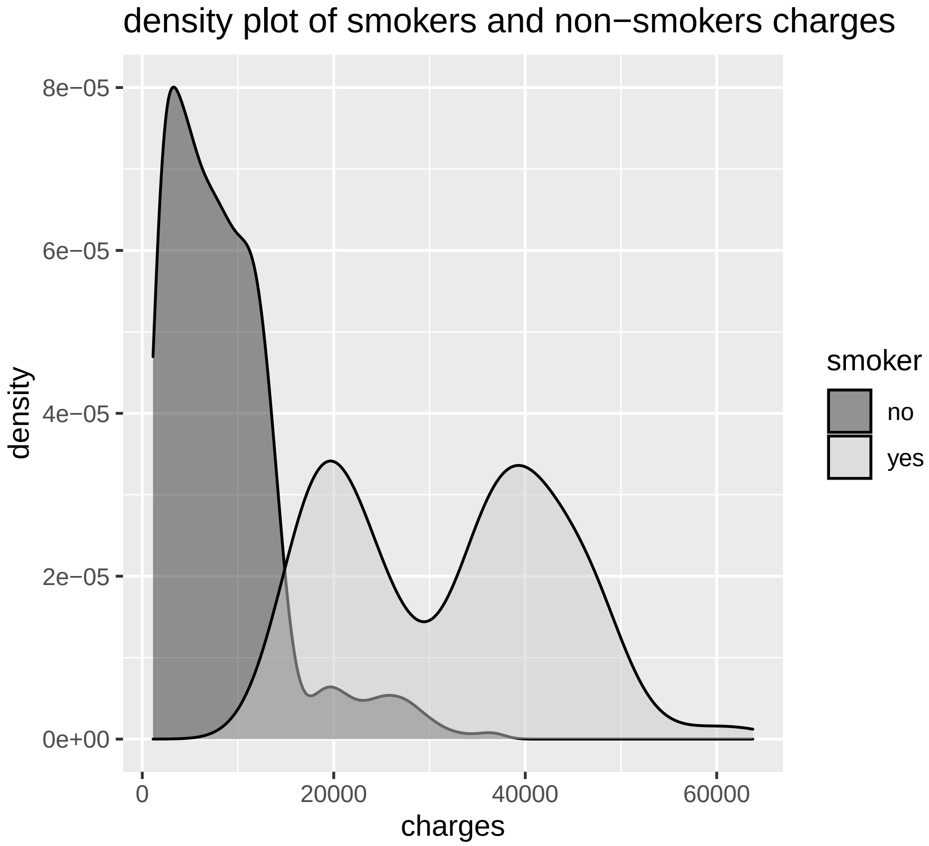
Now we compare the smoking males against females.

```r
ggplot(insurance,aes(x=sex,fill=smoker))+geom_bar(stat = 'count', alpha=0.75)+
  labs(x = 'people smoking') +
  geom_label(stat='count',aes(label=..count..), size=7) +
  theme_grey(base_size = 18)  + labs(title= "Smokers gender barplot")+
  theme ( legend.position = "none")
```



Now we can understand from the charges if a policyholder is a smoker or not with the plot below:

```
ggplot(insurance,aes(x=charges,fill=smoker))+
  geom_density(alpha=0.5, aes(fill=factor(smoker))) +
  labs(title="smoker")  + theme_grey() + scale_fill_grey() +
  labs(title= "density plot of smokers and non-smokers charges")
```

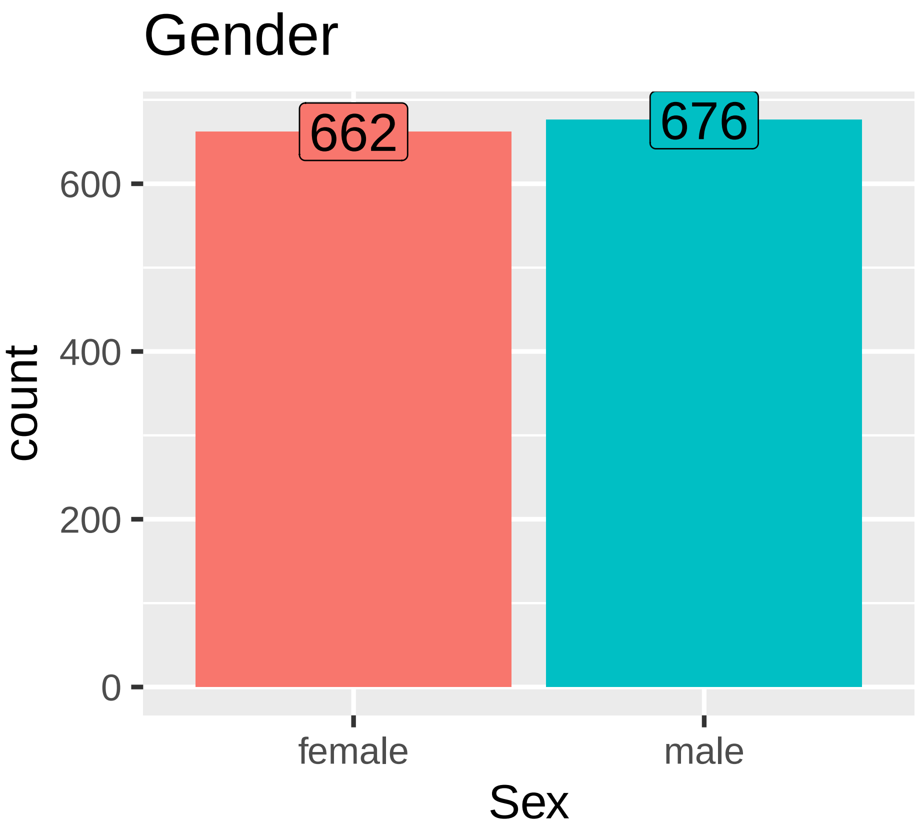


The first plot confirms that non-smokers are more numerous than smokers, but we don't see a relevant difference between genders. We will analyze the sex variable later, now the second plot shows that policyholders who have an insurance charge of more than 1800 $ are almost all smokers.

1.2.3 *Sex variable*

We represent the gender variable to verify the significance of sex in relation to insurance costs.

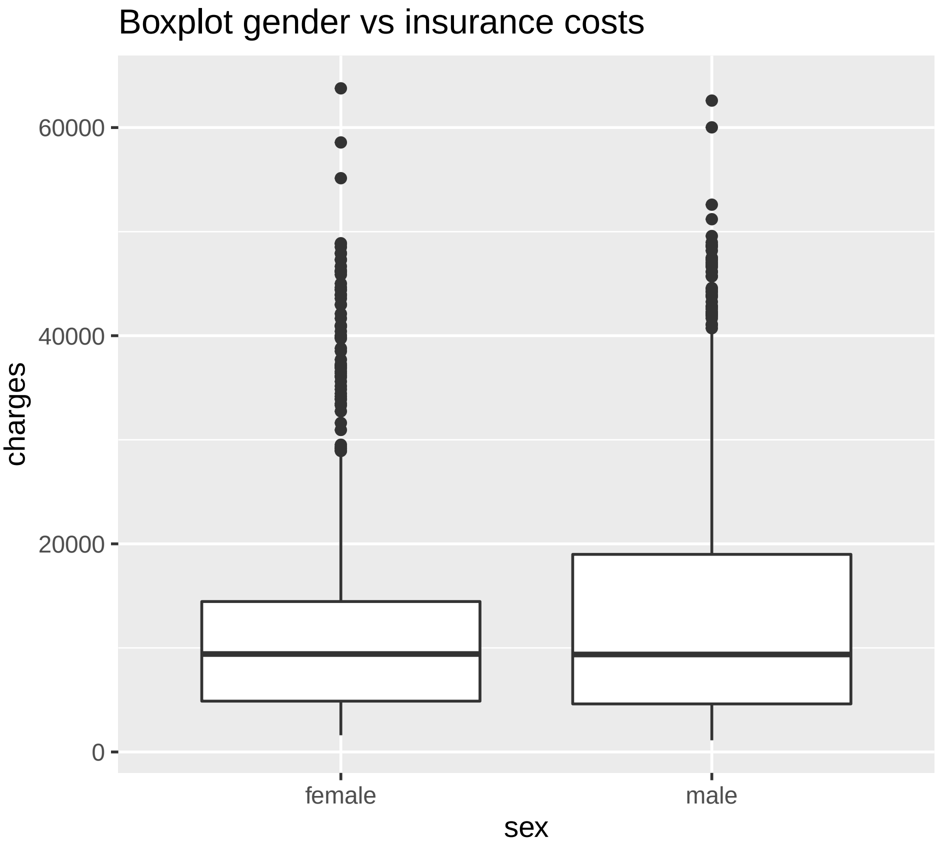
```
ggplot(insurance,aes(x=sex,fill=sex))+geom_bar(stat = 'count')+
  labs(x = 'Sex', title= 'Gender') +
  geom_label(stat='count',aes(label=..count..), size=7) +
  theme_grey(base_size = 18)+
  theme ( legend.position = "none") #barplot that shows the number of males
```



```
#and females
```

We can see the number of males and females from the barplot above.

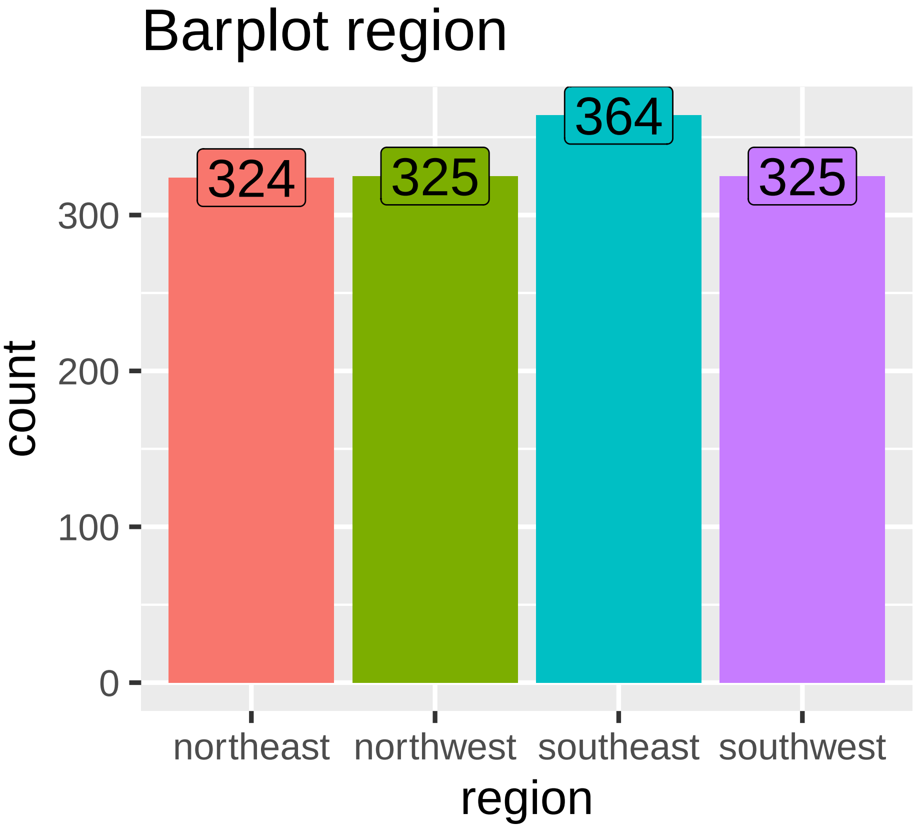
```
ggplot(insurance, aes(x=sex,y=charges))+
  geom_boxplot() + labs(title="Boxplot gender vs insurance costs") #boxplot
```



From the box-plot we can see that bviously the gender of the policyholder doesn't have a relevance on insurance charges. So there is no discrimination according to you are female or male.

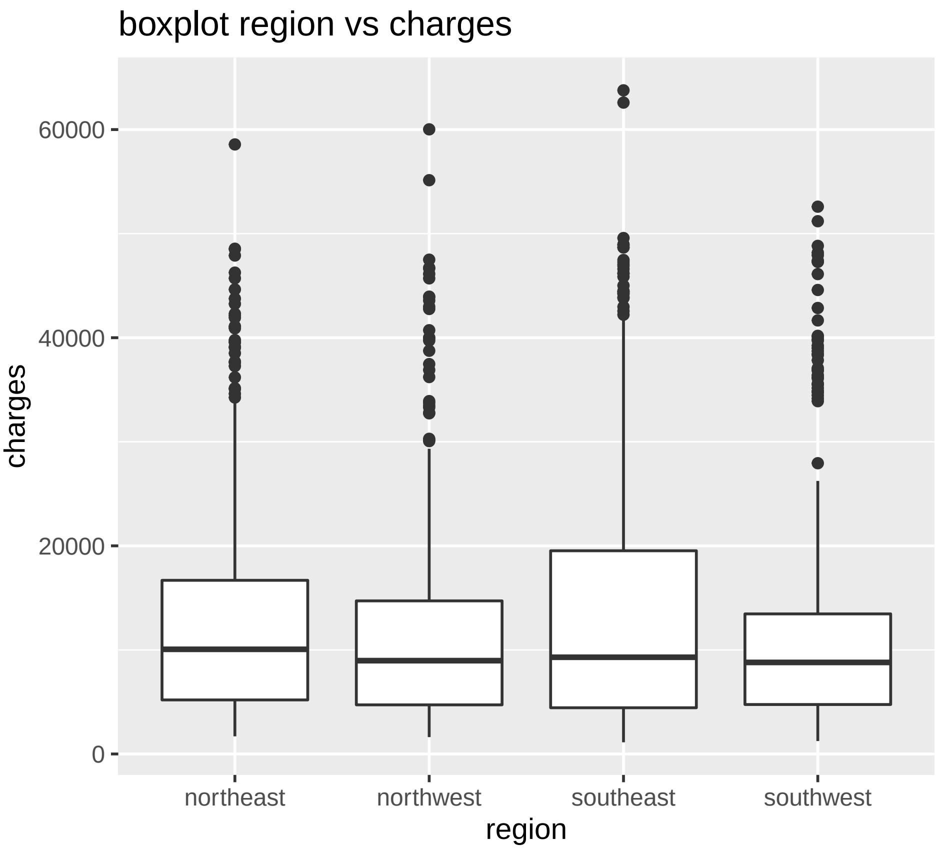
1.2.4 *Region*

```
ggplot(insurance,aes(x=region, fill=region))+geom_bar(stat = 'count')+
  labs(x = 'region', title='Barplot region') +
  geom_label(stat='count',aes(label=..count..), size=7) +
  theme_grey(base_size = 18) + theme ( legend.position = "none") #barplot region
```



Plot above shows us the number of policyholders living in the four regions.

```
ggplot(insurance,aes(x=region,y=charges))+geom_boxplot() +
  labs(title="boxplot region vs charges")
```



From the box-plot above we can argue that there is no significant difference between the region where you live and the insurance charges.

1.2.5 *Numeric variables*

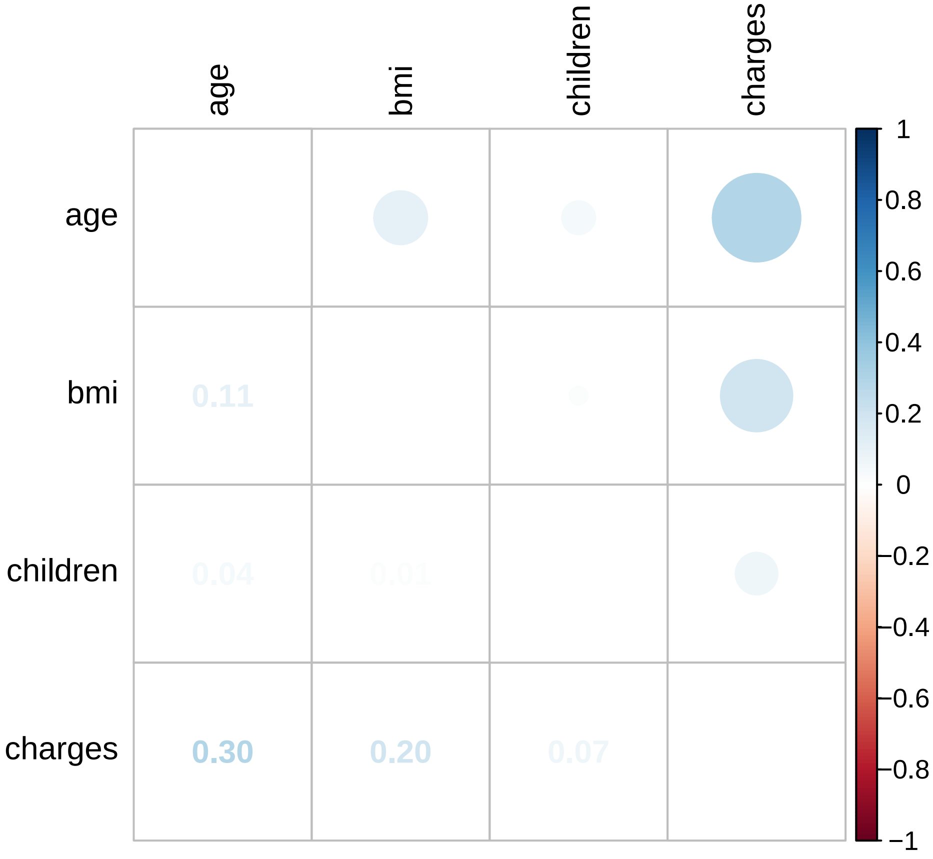
Now we study the correlation between the numeric variables and the response variable. We have four numeric variables. We create the correlation matrix and plot the correlations.

Correlation plot

```
numericvariables <- which(sapply(insurance, is.numeric))
#select the numeric variablesfrom the dataset

insurance.numeric <- insurance[, numericvariables] #creating the matrix
#with only numeric variables.

correlation.insurance <- cor(insurance.numeric) #creating di correlation matrix
corrplot.mixed(correlation.insurance, tl.col="black", tl.pos = "lt")
```

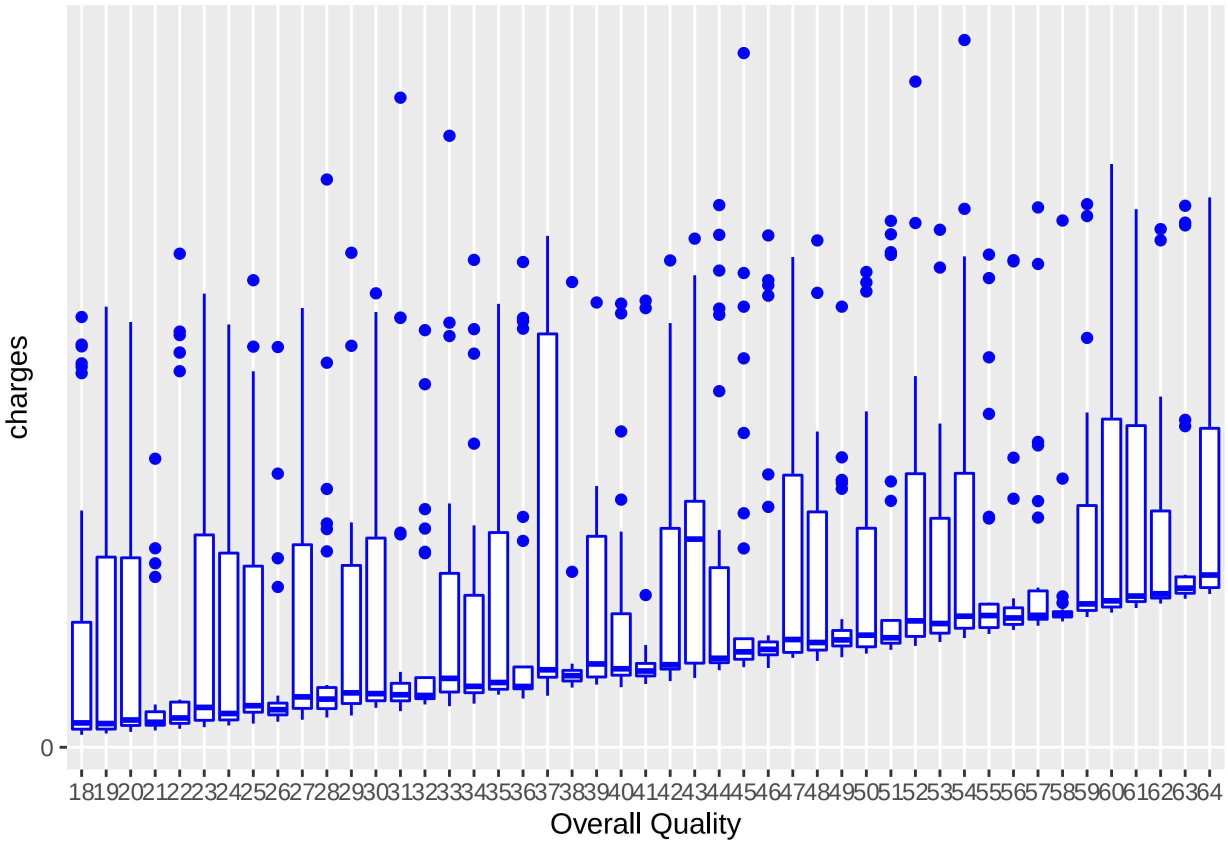


```
#plot of the correlation matrix
```

We found that all these numeric variables are weakly correlated with variable insurance charges.

But we have to focus on the age variable:

```
ggplot(insurance, aes(x=as.factor(insurance$age), y=charges))+
        geom_boxplot(col='blue') + labs(x='Overall Quality') +
        scale_y_continuous(breaks= seq(0, 800000, by=100000))
```



We can see that this positive correlation between charges and ages is verified because the insurance charges increase when the person becomes more and more senior.

2.1 *Modelling*

2.1.1 *Ridge regression*

We improve the linear model introducing some additional fitting procedures that give a better accuracy and model interpretability. In linear model, OLS estimator is decomposed in bias and variance, where bias is very low and variance is high. In ridge regression bias is higher and variance is lower. We extend the OLS with shrinkage approach. Ridge regression is a shrinkage-type estimator, it's a linear regression with a shrinkage penalty term. The RSS for linear model is:

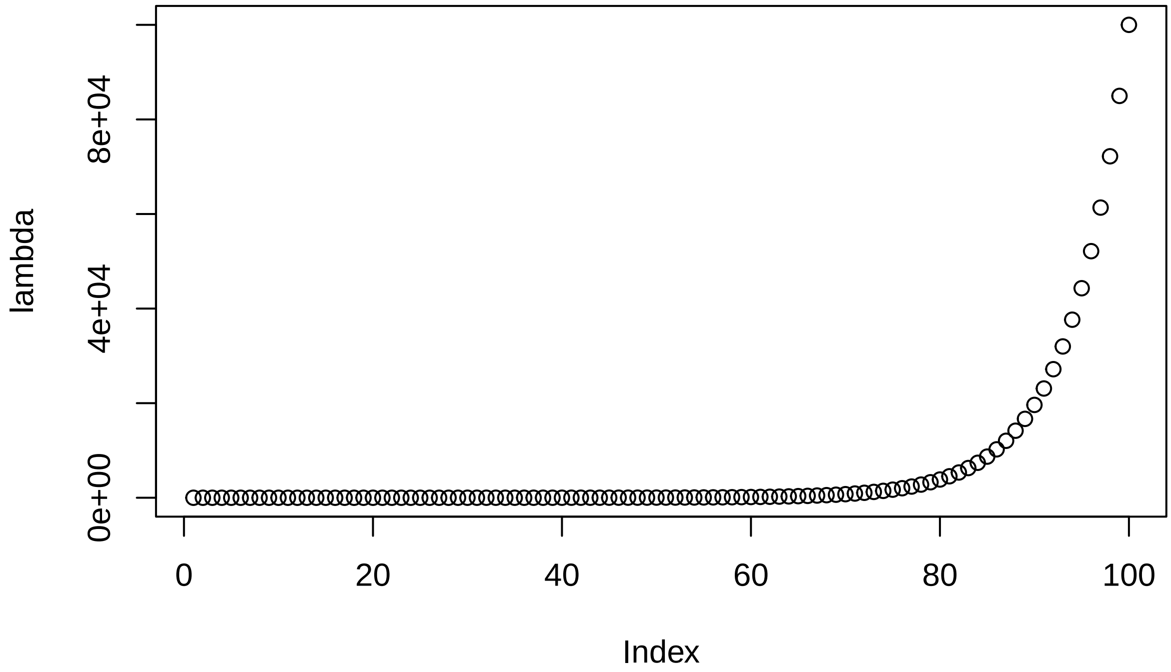
$$RSS = \sum_{i=1}^{n}(y_i - \beta_1 - \sum_{j=2}^{k}\beta_j x_i j)^2$$

Ridge is very similar to least squares, but we minimize a different quantity, including the shrinkage penalty term, which is the euclidean distance:

$$RSS = \sum_{i=1}^{n}(y_i - \beta_1 - \sum_{j=2}^{k}\beta_j x_i j)^2 + \lambda\sum_{j=2}^{k}\beta_j^2$$

In order to perform the regression estimation, we need to include the lambda. It is the tuning parameter that we need to choose. We have different solutions to choose the best tuning parameter, we define a grid of lambda values to include in the optimization function.

```
x <- model.matrix(insurance$charges~., insurance)[,-1]
 #creating the X matrix of regressors,
# excluding the dependent variable.
y <- insurance$charges #dependent variable.
lambda <- rev(10^seq(5, -2, length = 100)) #grid of lambda values
plot(lambda)
```



Now we define training and test set to perform our analysis.

```
set.seed(489)
train = sample(1:nrow(x), 0.7*nrow(x))
test = (-train)
ytest = y[test]
```

We are going to perform the difference between ridge and linear regression. We have splitted the data in training and test set (70%-30%), so we can compare the predictions of the two type of regression. We start from the linear model:

```
linearmodel <- lm(insurance$charges~., data = insurance, subset = train)
#linear regression
summary(linearmodel)
```

```
##
## Call:
## lm(formula = insurance$charges ~ ., data = insurance, subset = train)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -11550  -3146  -1111   1590  29794
```

```
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -12584.08    1194.29 -10.537  < 2e-16 ***
## age                  256.90      14.46  17.769  < 2e-16 ***
## sexmale              117.27     413.19   0.284    0.777
## bmi                  346.78      35.14   9.868  < 2e-16 ***
## children             686.56     175.18   3.919 9.54e-05 ***
## smokeryes          24021.15     512.19  46.899  < 2e-16 ***
## regionnorthwest      288.53     583.88   0.494    0.621
## regionsoutheast     -844.34     591.68  -1.427    0.154
## regionsouthwest     -661.60     596.87  -1.108    0.268
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 6264 on 927 degrees of freedom
## Multiple R-squared:  0.7467, Adjusted R-squared:  0.7445
## F-statistic: 341.6 on 8 and 927 DF,  p-value: < 2.2e-16
```

From the summary of the linear model we denote the significativity of the age, bmi and smoker variables, like seen previously in the graphical analysis of the variables. Now we perform the prediction and validate the model:

```
linear.pred <- predict(linearmodel, newdata = insurance[test,])#prediction of
#the linear model
MSEL <- mean((linear.pred-ytest)^2) #MSE
MSEL
```

```
## [1] 31722180
```

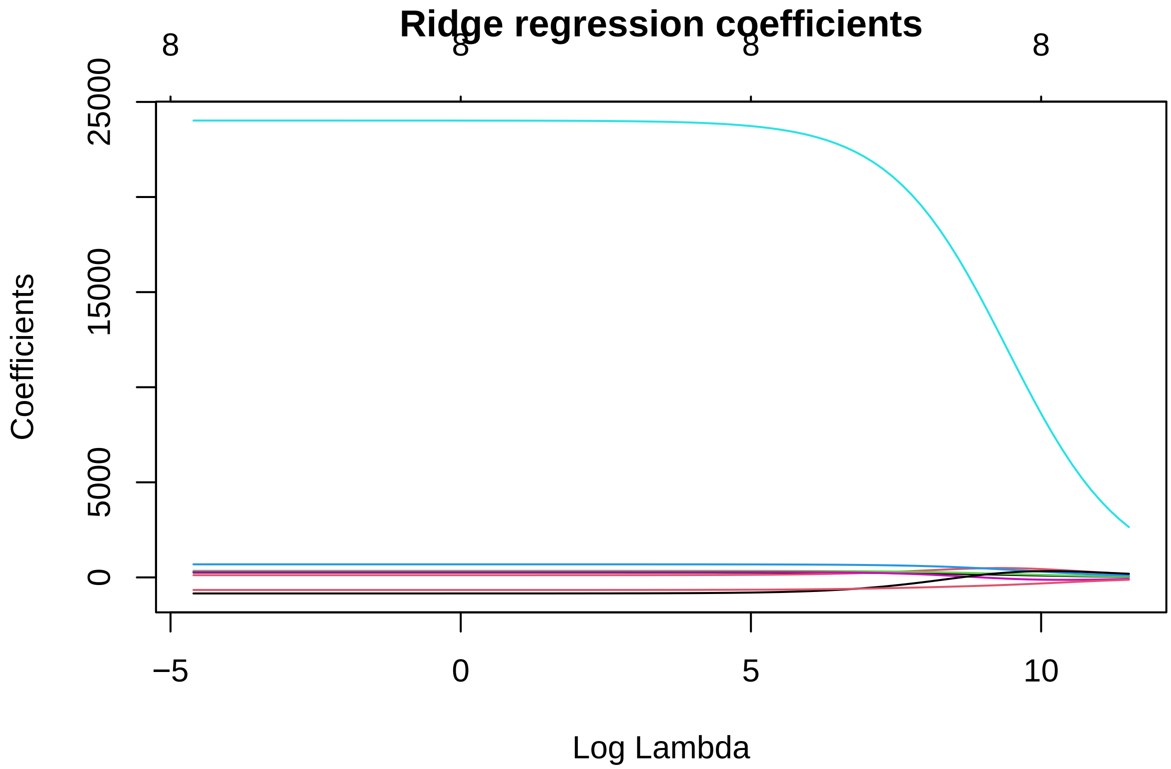
```
sst <- sum((ytest - mean(ytest))^2)
sse.l <- sum((linear.pred - ytest)^2)
rsq <- 1 - sse.l/sst
rsq #R-squared
```

```
## [1] 0.7558625
```

We have calculated the MSE and than the R-squared index, that indicates a good model prediction.

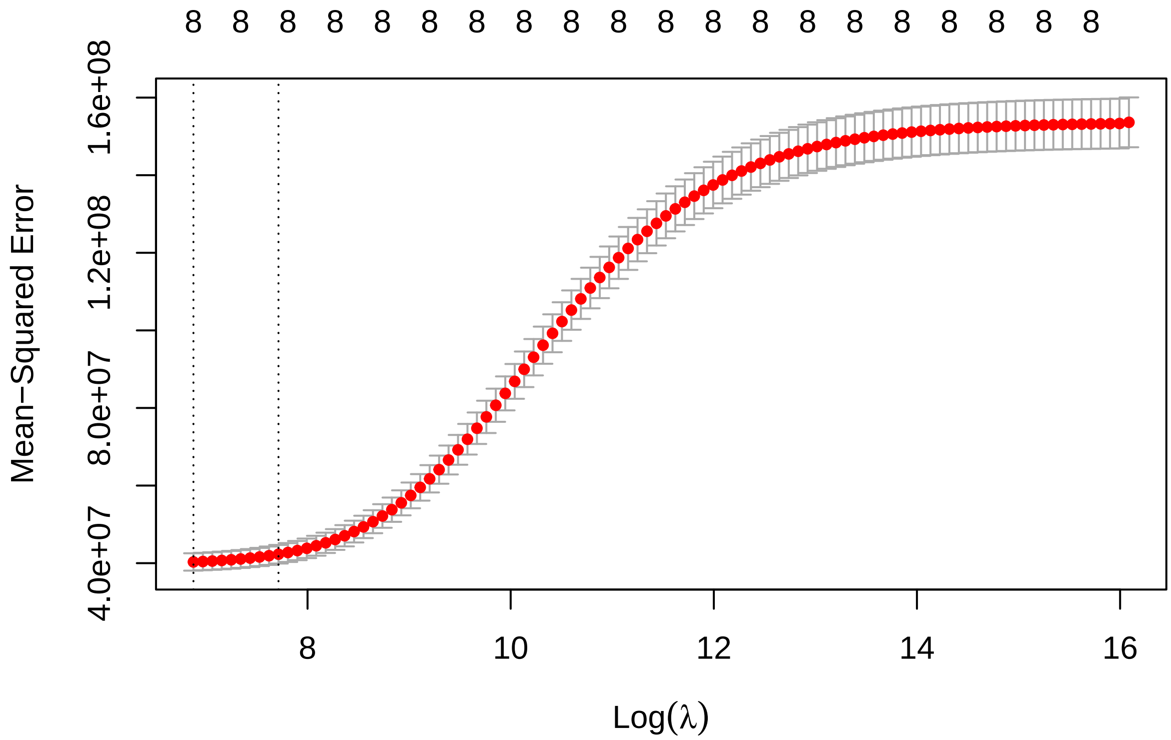
Now we perform the ridge regression to compare the results.

```
ridge.insur <- glmnet(x[train,], y[train], alpha = 0, lambda = lambda)
#ridge regression
plot(ridge.insur, xvar = "lambda", main= "Ridge regression coefficients")
```



The figure above represent the path of the lambda coefficients selected to perform the ridge regression. Now we perform the k-fold cross validation to find the best lambda value.

```
cv.out <- cv.glmnet(x[train,], y[train], alpha = 0) #k-fold cross validation
bestlamridge <- cv.out$lambda.min #extracting the optimal lambda value
plot(cv.out)
```



The figure shows the minimum value for the mean squared error as the best value of my lambda. The vertical line indicates us the optimal lambda.

```
ridge.pred <- predict(ridge.insur, s = bestlamridge, newx = x[test,])
MSER <- mean((ridge.pred-ytest)^2) #MSE

#SST and SSE
sst <- sum((ytest - mean(ytest))^2)
sse <- sum((ridge.pred - ytest)^2)

#R-Squared
rsqr <- 1 - sse/sst
rsqr
```

```
## [1] 0.7551682
```

Performing prediction, we calculated MSE and R-squared. Comparing the results of the two models, we argue that ridge regression improve slightly the performance of the estimation.

2.1.2 *Lasso*

Lasso is an acronym for *Least Absolute Selection and Shrinkage Operator*. It includes a different penalty term from ridge regression, overcoming its limits of variable selection.

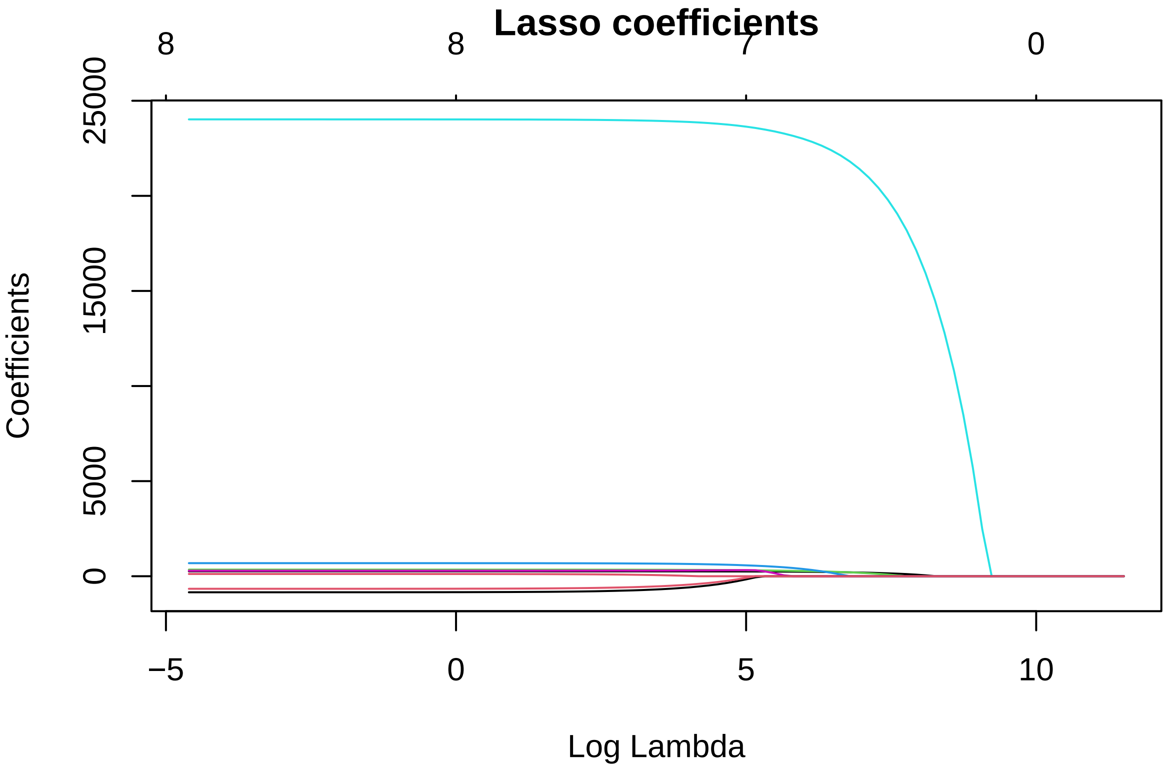
$$RSS = \sum_{i=1}^{n}(y_i - \beta_1 - \sum_{j=2}^{k}\beta_j x_i j)^2 + \lambda\sum_{j=2}^{k}|\beta|$$

It estimates beta hat that minimize the quantity:

$$RSS + \lambda\sum_{j=2}^{k}|\beta|$$

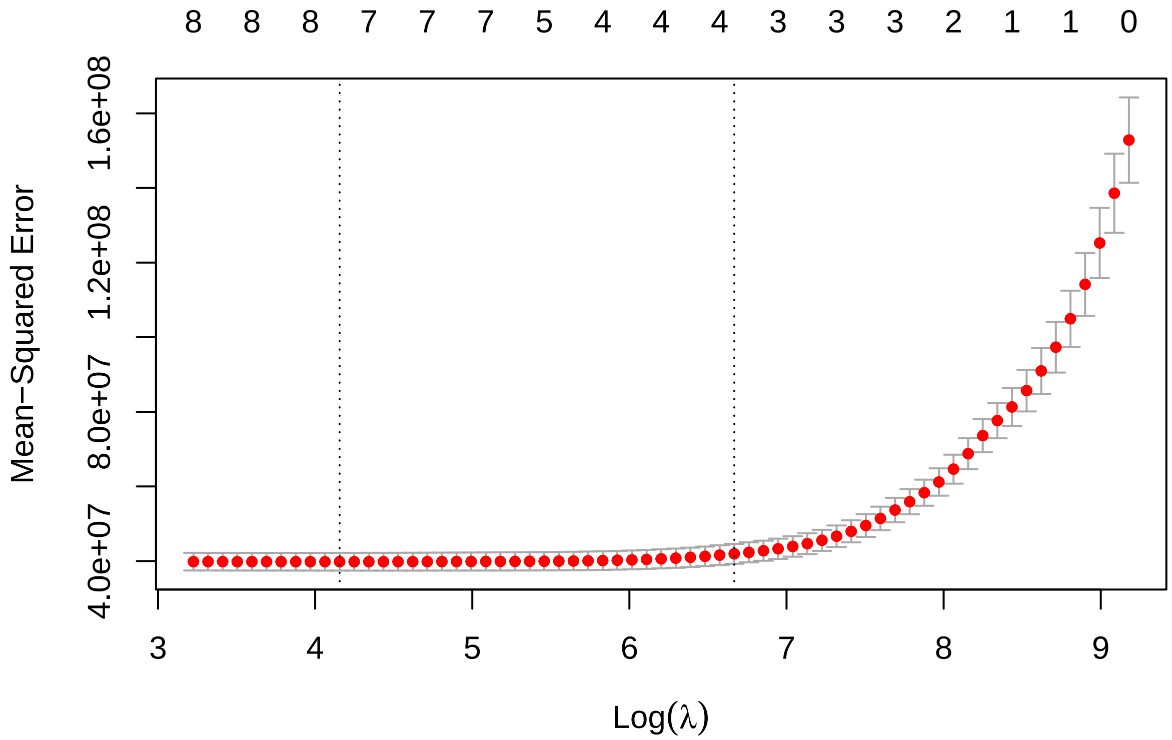
The difference between Lasso and Ridge is that Lasso use absolute value norm instead the euclidean norm.

```
lasso.insur<- glmnet(x[train,], y[train], alpha = 1, lambda = lambda) #lasso
plot(lasso.insur, xvar = "lambda", main= "Lasso coefficients")
```



Now we perform k-fold cross-validation to find the best lambda.

```
cv.lasso <- cv.glmnet(x[train,], y[train], alpha = 1) #cross-validation for lasso
plot(cv.lasso)
```



```
#extracting best lambda
bestlamlasso <- cv.lasso$lambda.min
```

We have to make predictions with the model.

```
lasso.pred <- predict(lasso.insur, s = bestlamlasso, newx = x[test,])#prediction
MSE.L <- mean((lasso.pred-ytest)^2) #MSE
```

We calculated the MSE, smaller than the MSE of the previous models. Now we compute the R-squared index.

```
#SSE for lasso prediction
ssel <- sum((lasso.pred - ytest)^2)

#R-Squared for lasso
rsql <- 1 - ssel/sst
rsql
```

```
## [1] 0.7573986
```

Now we compare the results of the three models.

```r
mse <- c(MSEL, MSER, MSE.L)
r2 <- c(rsq, rsqr, rsql)
results <- data.frame(mse, r2, row.names = c("linear", "ridge", "lasso"))
kable(results,caption = "MSE and R-squared", digits = 4, "pipe")
```

Table 2: MSE and R-squared

	mse	r2
linear	31722180	0.7559
ridge	31812402	0.7552
lasso	31522593	0.7574

As we can see from the results, MSE is lower for lasso model, also the R^2 index is higher in the same model. In addiction, we can say that the other two model are also good for the prediction, because their results aren't so different from the lasso.

2.1.3 *Prediction with Lasso*

Now we can see the first 20 predictions made by lasso model.

```r
new.insurance <- model.matrix(insurance$charges~., insurance)[,-1]
charges <- predict(lasso.insur, s = bestlamlasso, newx = new.insurance)
#applying our model
data.predictions <- cbind(insurance[,-7], charges)
colnames(data.predictions)[7] <-"Predicted charges"
kable(data.predictions[1:20,],caption = "Predicted charges", digits = 4, "pipe")
```

Table 3: Predicted charges

age	sex	bmi	children	smoker	region	Predicted charges
19	female	27.900	0	yes	southwest	25536.574
18	male	33.770	1	no	southeast	3873.906
28	male	33.000	3	no	southeast	7422.977
33	male	22.705	0	no	northwest	4206.001
32	male	28.880	0	no	northwest	6009.483
31	female	25.740	0	no	southeast	3850.949
46	female	33.440	1	no	southeast	10850.085
37	female	27.740	3	no	northwest	8803.376
37	male	29.830	2	no	northeast	8557.782
60	female	25.840	0	no	northwest	12083.062
25	male	26.220	0	no	northeast	3043.539
62	female	26.290	0	yes	southeast	35746.828
23	male	34.400	0	no	southwest	4852.968
56	female	39.820	0	no	southeast	14869.942
27	male	42.130	0	yes	southeast	32164.197
19	male	24.600	1	no	southwest	1213.110
52	female	30.780	1	no	northeast	12032.085
23	male	23.845	0	no	northeast	1746.158
56	male	40.300	0	no	southwest	15172.941
30	male	35.300	0	yes	southwest	30789.125

For example, the first observation, is a 19 years old female, that is a smoker and has a bmi of 27.9, a little bit higher than the bmi that indicates the ideal weight. The insurance charges of 25536 dollars are a realistic estimation for that policyholder.

CONCLUSION

We have computed three regression models to predict the insurance costs. From the results we can say that lasso regression, with the absolute value norm as penalty term give us the better prediction, with a little greater R-squared index and a smaller MSE. Ridge and linear occupies the second position, while linear model has anyway lower value of MSE, ridghe has a higher value of Rˆsquared.