Estadística avanzada para ciencia de datos

Regresión - Proyecto

Jakub Maciążek

Project description

Risk prediction:

- Import in R the dataset riesgos.csv
- The insurance company wants to have a model to predict the medical expenses of the insured.

Perform an analysis of the dataset and develop one or more models that answer the following. Deliver a compressed file with: .RMarkDown with your analysis and the outputs, auxiliary files, images, etc.

- Analyze the structure of the dataset. What kind of data do we have? It raises the problems that we can have when working with this dataset.
- Statistically analyzes the attributes. Detect normality, biases, outliers, etc.
- Plot the expense attribute with a histogram. What knowledge do you extract from the visualized information?
- Obtain the correlation matrix between the attributes of the dataset. Which attributes seem to be more and less related? (cor).
- Visualize the relationships between the attributes - scatterplot (plot, pairs, pairs.panels).
- Set up a linear m1 regression model between expenses and another variable (the one you think best models the medical expenses of the insured).
- Try an m2 model using polynomial functions.
- Evaluates the efficiency of the models (summary). Extracts all the information about the validity of the two models created.
- Improve the model using generalized regression. Create an m3 model taking into account all the variables. Analyze which variables are significant. Look at the efficiency of the new model.
- Use anova to see which model of those created is more interesting.

Dataset analysis

Following section investigates dataset structure in order to detect problematic variables and outliers, as well as to find related variables and choose them for the model.

Variables description

> *ToDo: Analyze the structure of the dataset. What kind of data do we have? It raises the problems that we can have when working with this dataset.*

Dependent variable: gastos - main goal of the assignment is to develop a model to predict the medical expenses of the insured. Therefore *gastos* is the dependent variable. It takes form of real value numbers.

Independent variables:

- edad - age represented by integer number
- sexo - gender represented by labels [mujer, hombre]
- bmi - body mass index represented by real value number
- hijos - nr of children represented by integer number
- fumador - boolean value indicating whether insured person smokes
- region - label representing region of insured origin (4 different values)

```
riesgos <- read.csv("S:/0_Universidad_de_Malaga/MI_Ingenieria_y_ciencia_de_datos/Estatistica_avanzada_pa
head(riesgos)
##   edad   sexo    bmi hijos fumador    region    gastos
## 1   19  mujer 27.900     0      si Andalucía 16884.924
## 2   18 hombre 33.770     1      no    Murcia  1725.552
## 3   28 hombre 33.000     3      no    Murcia  4449.462
## 4   33 hombre 22.705     0      no    Madrid 21984.471
## 5   32 hombre 28.880     0      no    Madrid  3866.855
## 6   31  mujer 25.740     0      no    Murcia  3756.622
```

Linear regression assumptions regarding data

- Numerical inputs: regression requires independent variables to be numerical. For that reason *sexo*, *fumador* and *region* will be converted under the hood into boolean dummy variables, representing which category record applies to.
- Linearity: The relationship between X and the mean of Y is linear. - independent variables will be choosen based on visual analysis, that suggests linear correlation. Also later *residuals vs fitted values plot* will be examined. If the model does not meet the linear model assumption, we would expect to see residuals that are very large (big positive value or big negative value). To assess the assumption of linearity we want to ensure that the residuals are not too far away from 0 (standardized values less than -2 or greater than 2 are deemed problematic).
- Homoscedasticity: The variance of residual is the same for any value of X. - Will be checked through the same plot as in case of linearity, however in this case we would make sure that there is no pattern in the residuals and that they are equally spread around the y=0 line. Lack of the pattern can also be investigated in *scale-location plot*.
- Normality: For any fixed value of X, Y is normally distributed, however it is even more important that residuals are normally distributed. - This will be checked through *QQ-plot*. If residuals lie along 45 degree line, required normality of error distribution can be assumed.
- Independence: Observations are independent of each other. - This cannot be tested with diagnostic plots, only by examination of study design. In this case, there is no reason to suspect, that value of one person impacts insurance costs of another, therefore observations should be independent.

Attributes analysis

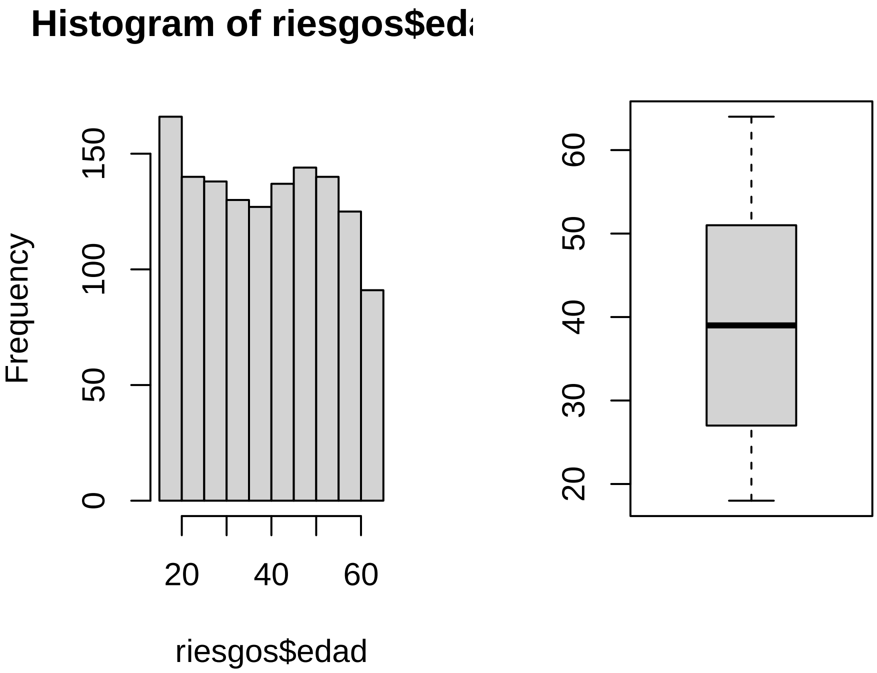
> *ToDo: Statistically analyzes the attributes. Detect normality, biases, outliers, etc.*

Possible outliers will be eliminated by examining histograms, boxplots and their cook's distance and influence. Also correlation between variables will be tested in order to eliminate possible multi-collinearity

Age [edad]

Age attribute does not follow normal distribution, in contrast, it is close to uniform distribution equally representing clients across all age groups. This fact, together with the lack of outliers, should mean that model won't be biased by the age of predicted client.

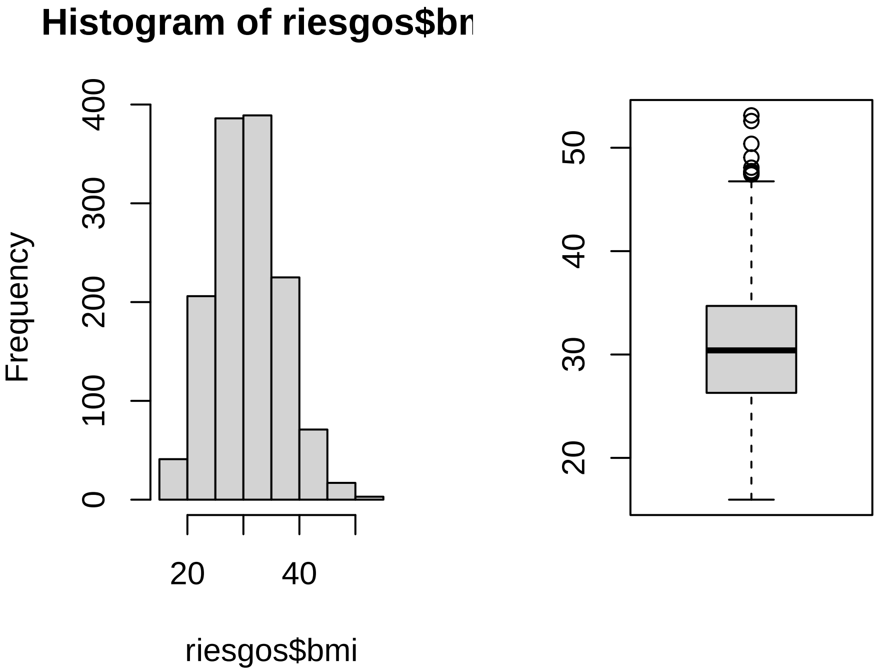
```
par(mfrow=c(1,2))
hist(riesgos$edad)
boxplot(riesgos$edad)
```



Body mass index [bmi]

BMI seems to follow normal distribution, but Shapiro-Wilk tests proves that to be incorrect (p-value lower than 0.05). Moreover, bar plot shows that data set contains many outliers with extremly high BMI values. They may have big influence on model reducing it quality, therefore they should be monitored and may require to be removed from dataset.

```
par(mfrow=c(1,2))
hist(riesgos$bmi)
boxplot(riesgos$bmi)
```

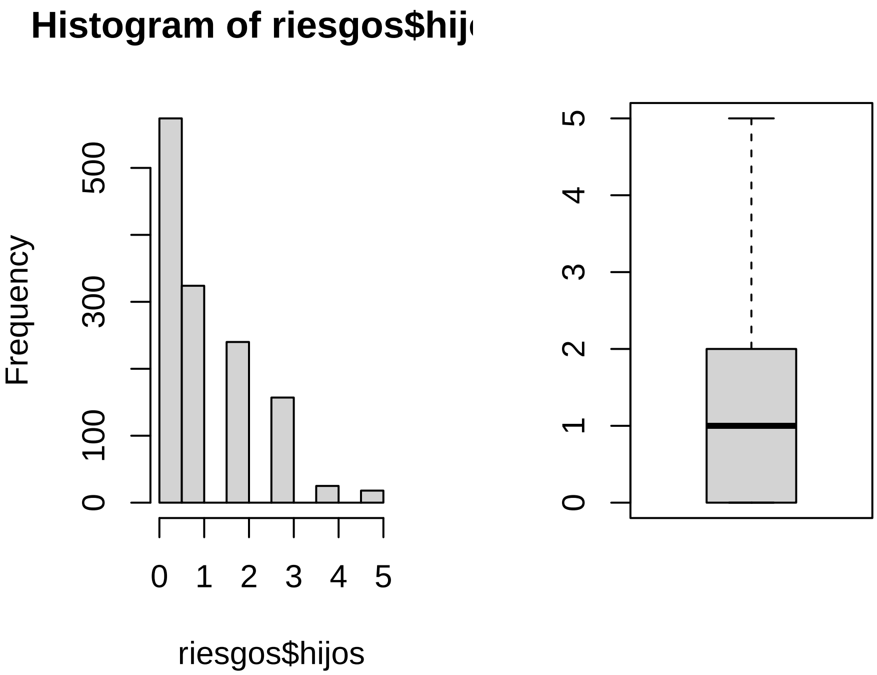


```
shapiro.test(riesgos$bmi)
##
##  Shapiro-Wilk normality test
##
## data:  riesgos$bmi
## W = 0.99389, p-value = 2.605e-05
```

Number of children [hijos]

Nr of childern does not follow normal distribution either. It is strongly positively skewed. Therefore, if this variable is choosen as a predictor, it may also effect distribution of residuals. If this is true, modifications to the variable or its omitance might be necessary.

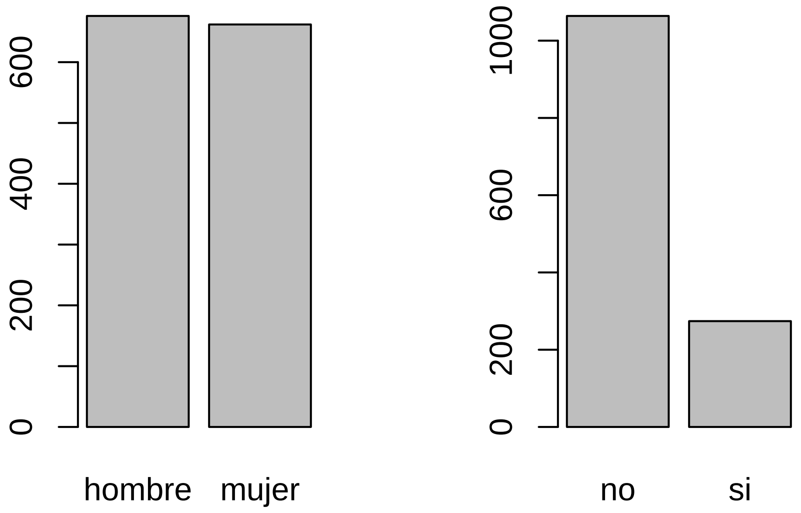
```
par(mfrow=c(1,2))
hist(riesgos$hijos)
boxplot(riesgos$hijos)
```



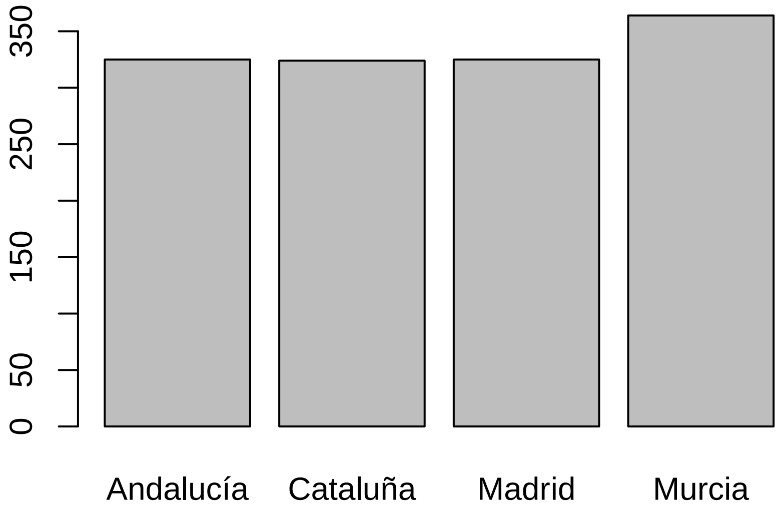
Non-numerical attributes analysis

Both gender and regions are close to uniform distribution, which means model should not be biased in their case. For smoking indication, number of smoking persons is about only the fifth of all insured, but its number of samples should be enough to avoid biases.

```
par(mfrow=c(1,2))
barplot(table(riesgos$sexo))
barplot(table(riesgos$fumador))
```



```
par(mfrow=c(1,1))
barplot(table(riesgos$region))
```

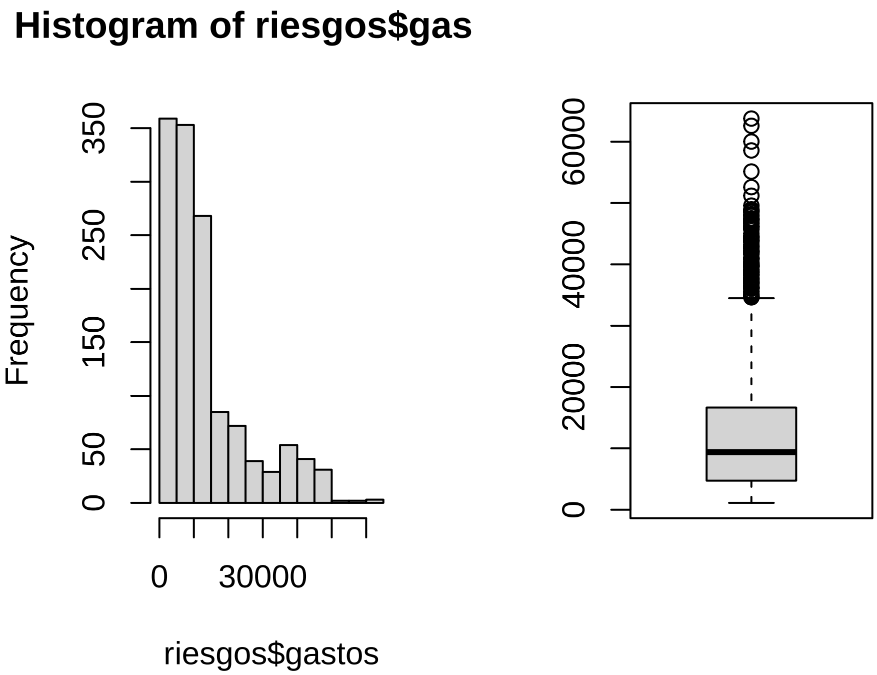


Dependent variable analysis

> *ToDo: Plot the expense attribute with a histogram. What knowledge do you extract from the visualized information?*

Expenses variable does not follow normal distribution and is highly positively skewed. Moreover, it has a large number of outliers, observations with expense value higher than 50 thousand. They might highly influence the model and worsen the prediction and may required to be removed.

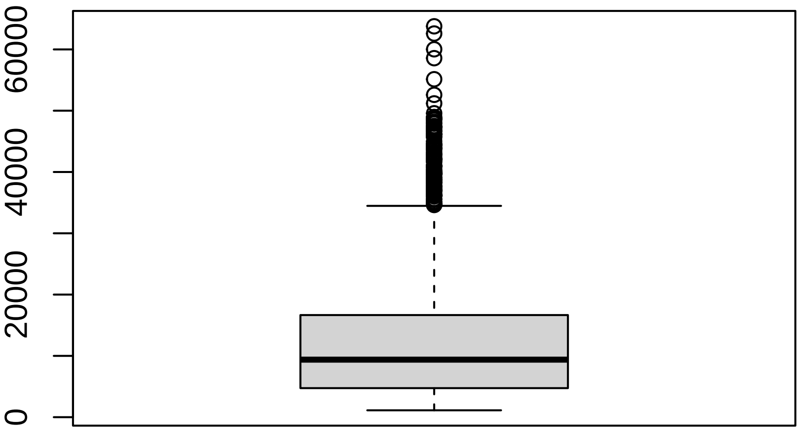
```
par(mfrow=c(1,2))
hist(riesgos$gastos)
boxplot(riesgos$gastos)
```



Outliers detection and removal

From histogram analysis, outliers can be defined as observations with expense value over 50 thousand. At the same time, according to boxplot, outliers start for expense values over ~34 thousand.

```
Summary<-boxplot(riesgos$gastos)$stats
```



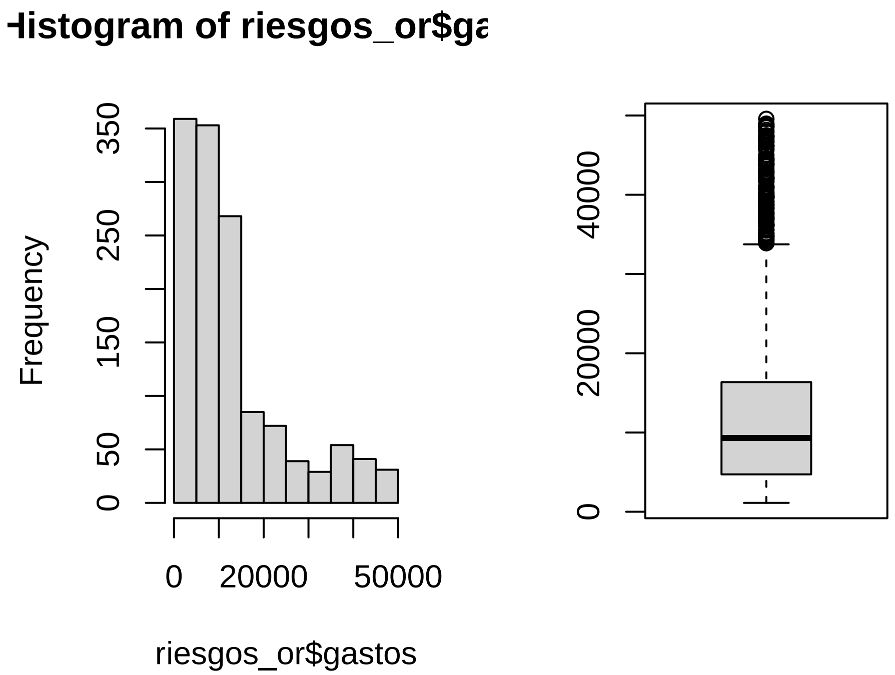
```
rownames(Summary)<-c("Min","First Quartile","Median","Third Quartile","Maximum")
Summary
##                       [,1]
## Min               1121.874
## First Quartile    4738.268
## Median            9382.033
## Third Quartile   16657.717
## Maximum          34472.841
```

As the values between 25 thousand and 35 thousand, seem to occure with similar frequency to those between 35 and 50 thousand, it seems reasonable to either keep or remove both of them. Therefore threshold for outliers was set to be 50 thousand.

```
riesgos_or <- filter(riesgos,gastos<=50000)
head(riesgos_or)
##   edad   sexo    bmi hijos fumador    region    gastos
## 1   19  mujer 27.900     0      si Andalucía 16884.924
## 2   18 hombre 33.770     1      no    Murcia  1725.552
## 3   28 hombre 33.000     3      no    Murcia  4449.462
## 4   33 hombre 22.705     0      no    Madrid 21984.471
## 5   32 hombre 28.880     0      no    Madrid  3866.855
## 6   31  mujer 25.740     0      no    Murcia  3756.622
```

After filtering, dataset still contains outliers according to boxplot. However, setting threshold for 35 thousands present similar outcome, and would require threshold at level of 25 thousand, to significantly reduce number of such defined outliers. That would be a big amount of observaions, therefore that was not done for now.

```
par(mfrow=c(1,2))
hist(riesgos_or$gastos)
boxplot(riesgos_or$gastos)
```



Correlation matrix, and attributes relationship analysis

> *ToDo: Obtain the correlation matrix between the attributes of the dataset. Which attributes seem to be more and less related? (cor)*

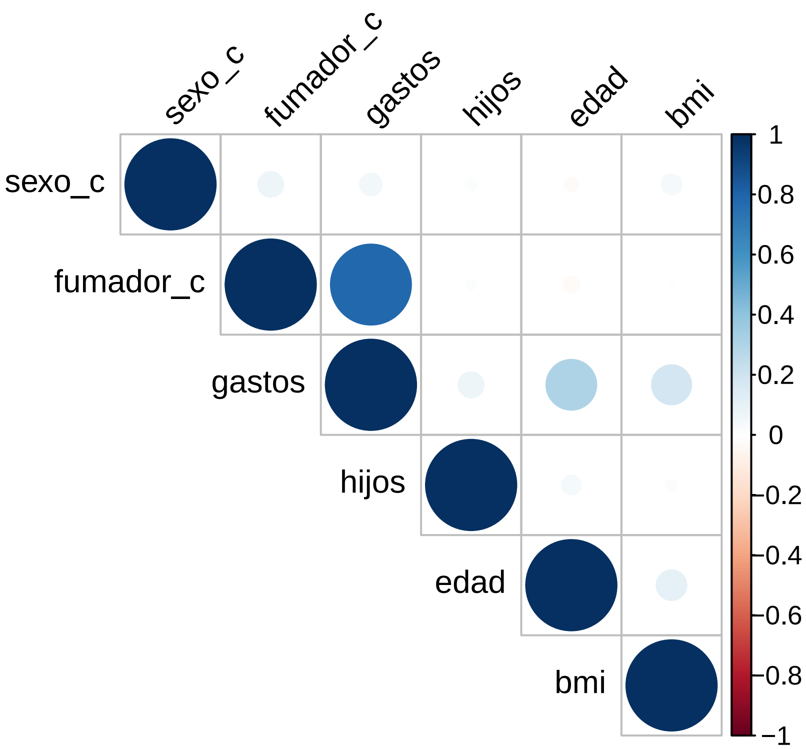
Correlation matrix can only be calculated for numeric values, therefore *sexo* and *fumador* were converted into binary values for that reason. For another analysis also regions were converted, showing only small correlation to bmi value.

From graph below it can be seen that biggest influence on expenses has variable *fumador*, and has high correlation value of ~0.8. Expenses are also correlated to *edad* by ~0.30, and *bmi* by ~0.20.

Other correlations between dependent and independent variables are very small. Moreover, correlations between dependent variables are also almost nonexistent, meaning there is no problem with multi-collinearity of variables, which means picking one variable for model does not influence decision on also including another one.

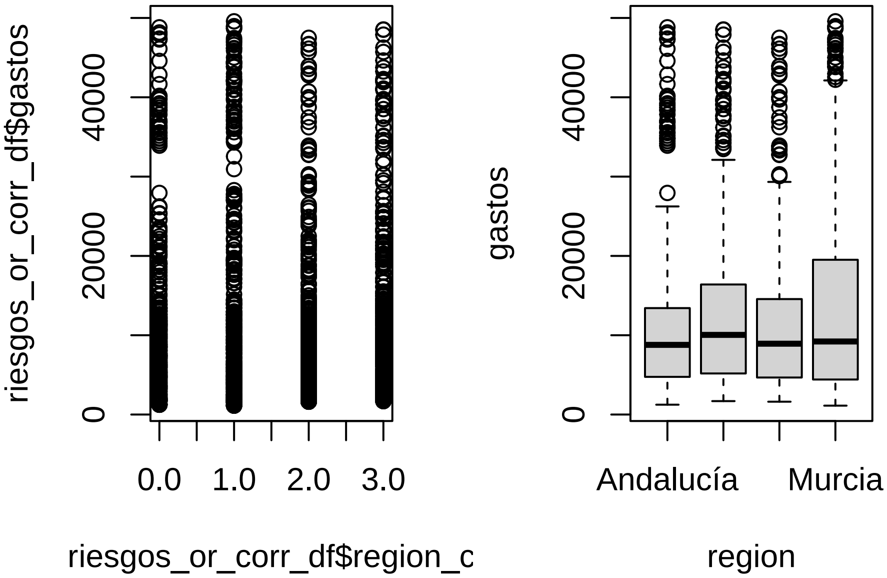
```
riesgos_or_corr_df <- riesgos_or
riesgos_or_corr_df$sexo_c <- c(mujer=0, hombre=1)[riesgos_or_corr_df$sexo]
riesgos_or_corr_df$fumador_c <- c(no=0, si=1)[riesgos_or_corr_df$fumador]

#Cataluna label and region conversion
cataluna<-unique(riesgos_or_corr_df$region)[4]
riesgos_or_corr_df$region_c <- c(Andalucía=0, Murcia=1, Madrid=2, cataluna=3)[riesgos_or_corr_df$region]
```

```
riesgos_or_corr_df[is.na(riesgos_or_corr_df)] <- 3

head(riesgos_or_corr_df)
##   edad   sexo    bmi hijos fumador    region    gastos sexo_c fumador_c
## 1   19  mujer 27.900     0      si Andalucía 16884.924      0         1
## 2   18 hombre 33.770     1      no    Murcia  1725.552      1         0
## 3   28 hombre 33.000     3      no    Murcia  4449.462      1         0
## 4   33 hombre 22.705     0      no    Madrid 21984.471      1         0
## 5   32 hombre 28.880     0      no    Madrid  3866.855      1         0
## 6   31  mujer 25.740     0      no    Murcia  3756.622      0         0
##   region_c
## 1        0
## 2        1
## 3        1
## 4        2
## 5        2
## 6        1

#corr = cor(riesgos_or_corr_df[ , c("edad", "bmi", "hijos", "sexo_c", "fumador_c", "region_c", "gastos"
corr = cor(riesgos_or_corr_df[ , c("edad", "bmi", "hijos", "sexo_c", "fumador_c", "gastos")] , method =
corr
##                  edad          bmi      hijos      sexo_c    fumador_c
## edad       1.00000000  0.108331456 0.04319684 -0.02242451 -0.028377266
## bmi        0.10833146  1.000000000 0.01485951  0.04891478 -0.006159009
## hijos      0.04319684  0.014859511 1.00000000  0.01604194  0.011084313
## sexo_c    -0.02242451  0.048914781 0.01604194  1.00000000  0.075786462
## fumador_c -0.02837727 -0.006159009 0.01108431  0.07578646  1.000000000
## gastos     0.30456096  0.187061332 0.07699300  0.05751925  0.785616809
##               gastos
## edad      0.30456096
## bmi       0.18706133
## hijos     0.07699300
## sexo_c    0.05751925
## fumador_c 0.78561681
## gastos    1.00000000

corrplot(corr, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
```



To further investigate correlation between region and expenses below graphs were generated. They also confirm that there is barely no correlation, as all regions have similar minimal and median expense value. They only differ in maximum values, but not significantly.

```
par(mfrow=c(1,2))
plot(riesgos_or_corr_df$region_c, riesgos_or_corr_df$gastos)
boxplot(gastos~region, data=riesgos_or_corr_df)
```

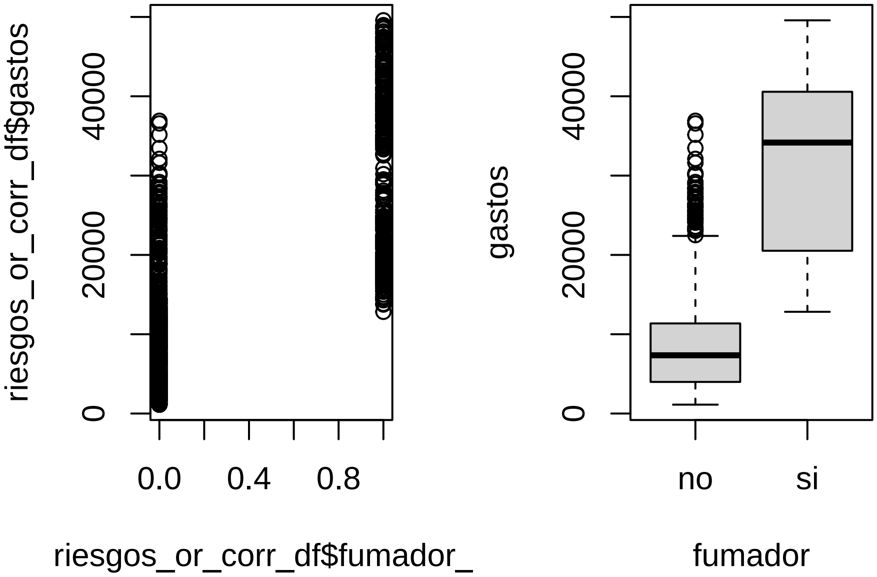


Variables relationship analysis

> *ToDo: Visualize the relationships between the attributes - scatterplot (plot, pairs, pairs.panels).*

First and second plot visualize high influence of smoking on insured expenses.

```
par(mfrow=c(1,2))
plot(riesgos_or_corr_df$fumador_c, riesgos_or_corr_df$gastos)
boxplot(gastos~fumador, data=riesgos_or_corr_df)
```



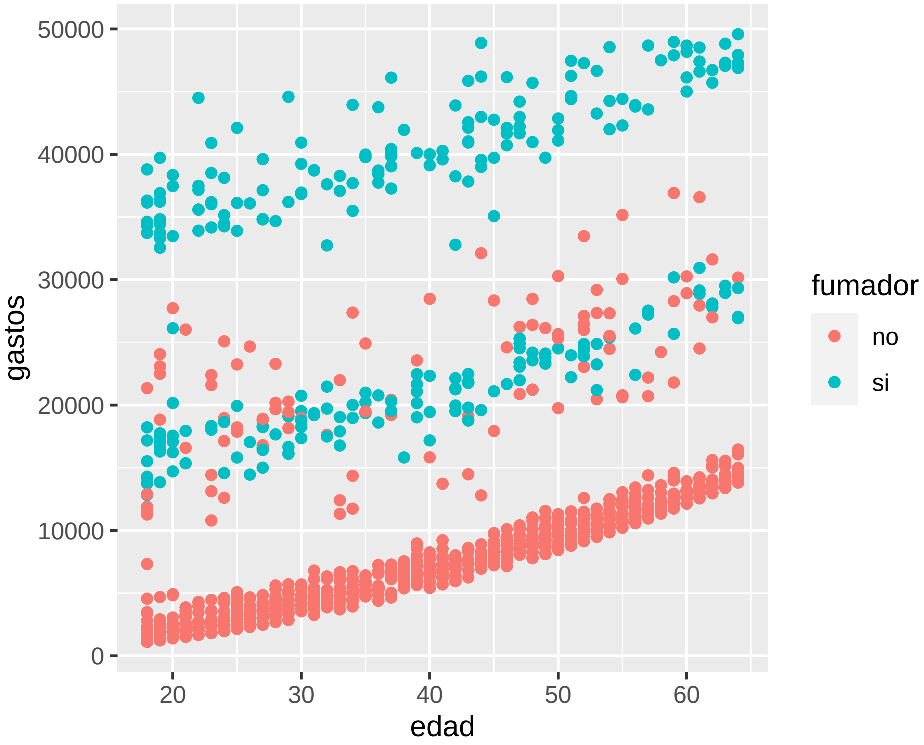
Third graph shows, that expenses value rises with almost linear relation to the insured age. Interestingly, also outliers form two linear lines.

This may mean that bottom line corresponds to non-smoking people, while upper or middle line corresponds to smoking people. It would mean smoking people would follow collinear line, but moved by difference in means visible on previous plots.

Question is, whether there is a 3rd variable, similarly to smoking, significantly increases costs both for smoking and non smoking people. It would move some of non-smokers to middle smoking line, and create upper line for smokers. That value may be discovered in generalized regression in later section.

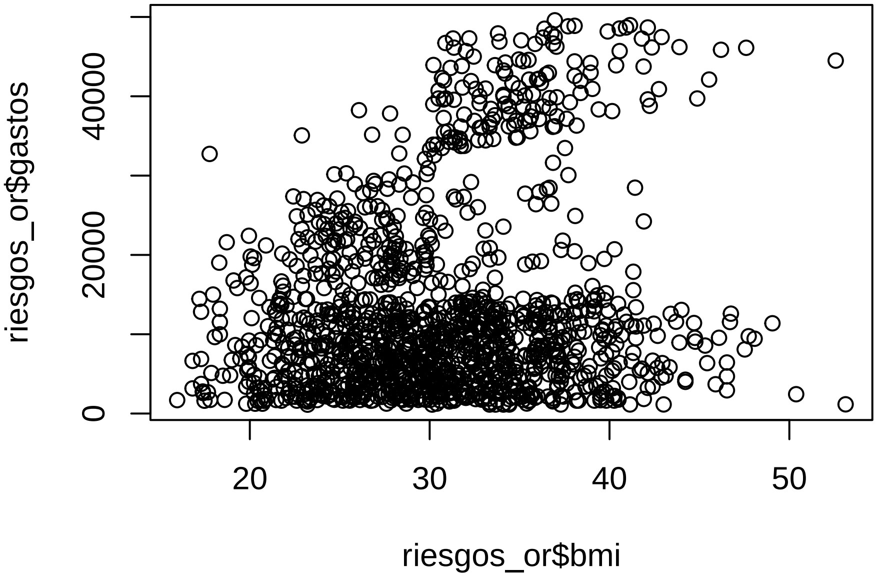
Outliers analysis: It is important to note, that 2nd and 3rd line most likely correspond to 2 spikes in the tail of *gastos* histogram. Moreover, it seems that it was a good decision not to remove outliers according to boxplot, but rather to histogram.

```
ggplot(data = riesgos_or, aes(x=edad, y=gastos)) + geom_point(aes(colour=fumador))
```



Even though *bmi* presented small correlation value, no clear pattern is visible.

```
plot(riesgos_or$bmi, riesgos_or$gastos)
```



Model developement

Based on previous analysis 3 models were developed.

Simple linear regression model developement

> *ToDo: Set up a linear m1 regression model between expenses and another variable (the one you think best models the medical expenses of the insured).*

Analysis in previous section indicated, that smoking is the most strongly correlated to expenses, therefore it was choosen for simple linear regression.

```
m1 <- lm(gastos ~ fumador, data = riesgos_or)
```

Polynomial regression model developement

> *ToDo: Try an m2 model using polynomial functions.*

Age was second most correlated variable and showed linear tendency at the same time, therefore it was choosen to be a polynomial predictor.

```
m2 <- lm(gastos ~ fumador  + I(edad^2), data = riesgos_or)
```

Evaluation of models M1 & M2

> *ToDo: Evaluates the efficiency of the models (summary). Extracts all the information about the validity of the two models created.*

M1 model evaluation

Starting with the basics, *F-statistic* is very high, and corresponding *p-value* essentially equal 0, meaning selected predictor is significantly related to outcome. In coefficients table, it can be seen that in fact *p-value* for *fumadorsi* is essentially 0, meaning it is a significant variable.

Therefore, first received model has a following equation:

$gastos = 8434.3 + 22943.6 * fumadorsi$

From the model we can interpret, that expected insurance costs for non smoking person are about ~8,434. If insured person smokes, costs rise by 22,943.6, making expected insurance costs about ~31,400.

According to *Adj. R^2* model explains ~62% of variance when predicting value of the insurance expenses.

From residuals summary it can be seen that mean residual value is far from 0. Moreover, even though 1st and 3rd quantile are at similar distance from median suggesting symmetry, shapiro-wilk test confirms that residuals does not follow normal distribution.

```
summary(m1)
##
## Call:
## lm(formula = gastos ~ fumador, data = riesgos_or)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18548.4  -5045.4   -988.4   3693.4  28476.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8434.3      222.0   37.99   <2e-16 ***
## fumadorsi    22943.6      495.7   46.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7241 on 1329 degrees of freedom
## Multiple R-squared:  0.6172, Adjusted R-squared:  0.6169
## F-statistic:  2143 on 1 and 1329 DF,  p-value: < 2.2e-16
shapiro.test(m1$residuals)
##
##  Shapiro-Wilk normality test
##
## data:  m1$residuals
## W = 0.96467, p-value < 2.2e-16
```

As residuals distribution significantly differs from normal distribution, variables significance is not confirmed. Though that does not affect model predictions, it suggests that choosen predictor may not be significant and best choice.

M2 model evaluation

For second model *Pr(>|t|)* in coefficients table for all variables proofs that they are significant.

Adj. R^2 for 2nd model indicates that it explains about ~73% of predicted expenses value variability.

```
summary(m2)
##
## Call:
## lm(formula = gastos ~ fumador + I(edad^2), data = riesgos_or)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14788.8  -2031.0  -1393.9   -227.7  23923.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.427e+03  3.215e+02   7.549 8.14e-14 ***
## fumadorsi   2.322e+04  4.194e+02  55.374  < 2e-16 ***
## I(edad^2)   3.434e+00  1.492e-01  23.017  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6125 on 1328 degrees of freedom
## Multiple R-squared:  0.7264, Adjusted R-squared:  0.7259
## F-statistic:  1763 on 2 and 1328 DF,  p-value: < 2.2e-16
```

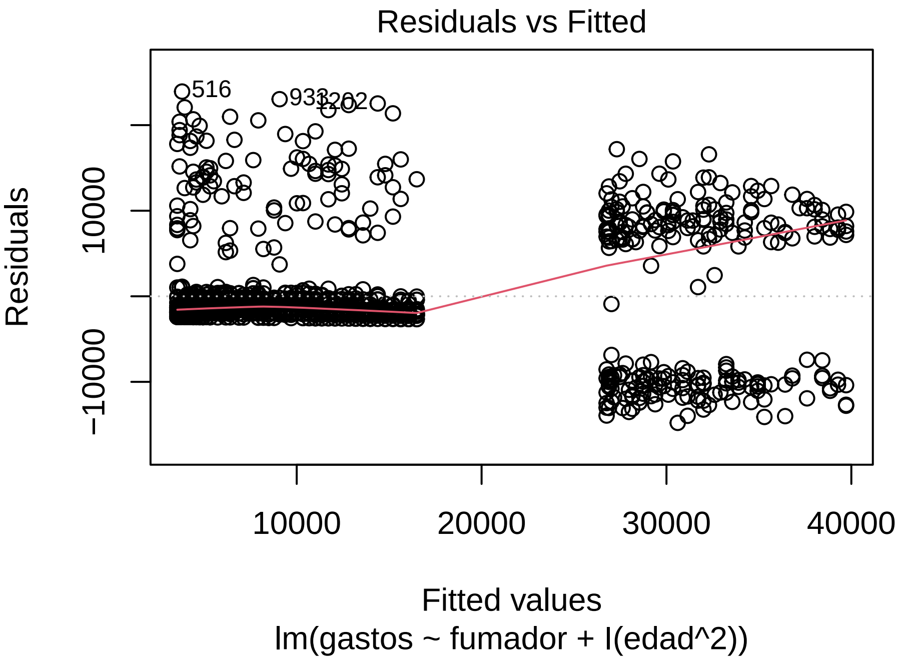
Recived residuals summary, together with Shapiro-Wilk test confirm that they again does not follow normal distribution.

```
shapiro.test(m2$residuals)
##
##  Shapiro-Wilk normality test
##
## data:  m2$residuals
## W = 0.81085, p-value < 2.2e-16
```

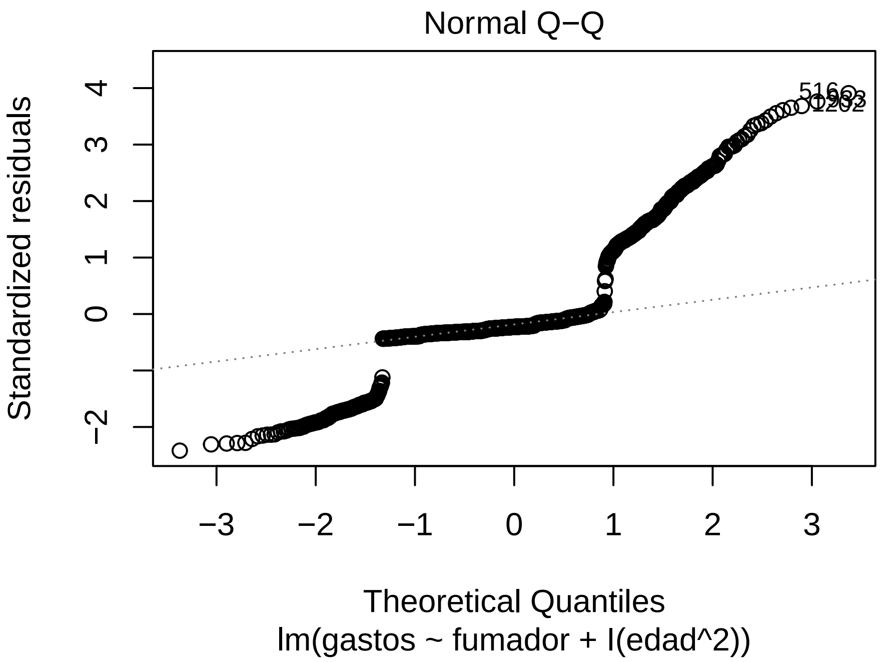
On 1st graph it can be seen, that for smaller values predictions are overestimated. This can be explaind by 1st line of smokers and 2nd mixed line overestimating costs of non-smokers. For higher expenses (mostly smokers) residuals starts to increase, meaning much more of the values are overestimated. This also suggests that model does not meet linearity condition, as well as Homoscedasticity, as residuals rise for higher expense values.

From 2nd QQ-plot, it can be seen that only part of residuals follow normal distribution.

```
plot(m2, which=1)
```



```
plot(m2, which=2)
```



After examining Residual standard error, it can be said that 2nd model has average error rate of ~47%.

```
rse2 <- sigma(m2)/mean(riesgos_or$gastos)
rse2
## [1] 0.4697993
```

Generalized regression model developement

> *ToDo: Improve the model using generalized regression. Create an m3 model taking into account all the variables. Analyze which variables are significant. Look at the efficiency of the new model.*

To determine which variables are significant when constructing generalized regression model, at first all variables were included. Based on the summary and *Pr(>|t|)* value in coefficients table, it can be seen that gender is insignificant, and region is significant only in case of Cataluña. For that reason next version of the model contains all variables except gender. In next section it will be investigated whether it was correct to keep region.

```
m3 <- lm(gastos ~ . , data = riesgos_or)
summary(m3)
##
## Call:
## lm(formula = gastos ~ ., data = riesgos_or)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max 
## -10888  -2744  -1041   1226  24337 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    -12382.862    998.416 -12.403  < 2e-16 ***
## edad              255.167     11.475  22.237  < 2e-16 ***
## sexomujer         108.692    321.284   0.338 0.735188    
## bmi               319.013     27.650  11.538  < 2e-16 ***
## hijos             513.419    132.904   3.863 0.000117 ***
## fumadorsi       23238.447    401.867  57.826  < 2e-16 ***
## regionCataluna    985.430    461.038   2.137 0.032747 *  
## regionMadrid      548.694    460.661   1.191 0.233828    
## regionMurcia       -9.182    454.313  -0.020 0.983878    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5834 on 1322 degrees of freedom
## Multiple R-squared:  0.7529, Adjusted R-squared:  0.7514 
## F-statistic: 503.4 on 8 and 1322 DF,  p-value: < 2.2e-16
```

Model m3 has value of $Adj.\ R\hat{}2$ of about ~76%. This means that it explains $3/4th$ of variability in predicted expenses.

Median value of residuals again is not close to 0, and shapiro-wilk test confirms that distribution of residuals is significantly different from normal.

```
m3 <- lm(gastos ~ I(edad^2) + bmi  + hijos + fumador + region , data = riesgos_or)
m3_sr <- lm(gastos ~ I(edad^2) + bmi  + hijos + fumador, data = riesgos_or)
summary(m3)
## 
## Call:
## lm(formula = gastos ~ I(edad^2) + bmi + hijos + fumador + region, 
##     data = riesgos_or)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -11215  -2774   -969   1089  24226 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    -7956.2594   919.0718  -8.657  < 2e-16 ***
## I(edad^2)          3.2437     0.1421  22.829  < 2e-16 ***
## bmi              314.4985    27.4269  11.467  < 2e-16 ***
## hijos            649.9717   131.8073   4.931 9.21e-07 ***
## fumadorsi      23233.0751   397.8155  58.402  < 2e-16 ***
## regionCataluna   979.2039   457.6665   2.140   0.0326 *  
## regionMadrid     530.7352   457.3050   1.161   0.2460    
## regionMurcia      -8.6363   450.9937  -0.019   0.9847    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5791 on 1323 degrees of freedom
## Multiple R-squared:  0.7563, Adjusted R-squared:  0.755 
## F-statistic: 586.5 on 7 and 1323 DF,  p-value: < 2.2e-16
shapiro.test(m3$residuals)
##
##  Shapiro-Wilk normality test
##
## data:  m3$residuals
## W = 0.89507, p-value < 2.2e-16
```

Error rate of the model is about ~44%.

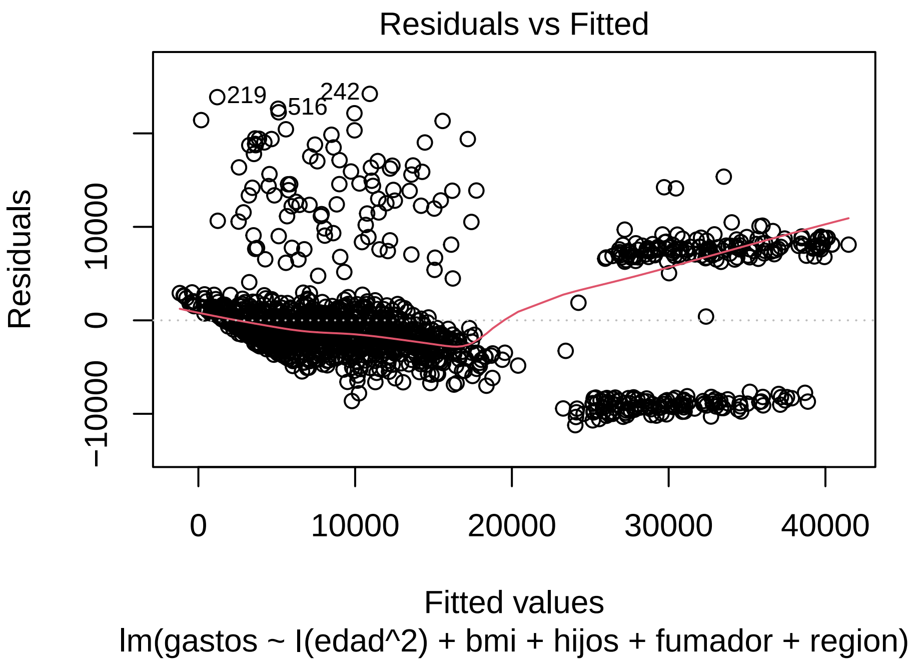
```
rse3 <- sigma(m3)/mean(riesgos_or$gastos)
rse3
## [1] 0.4442061
```

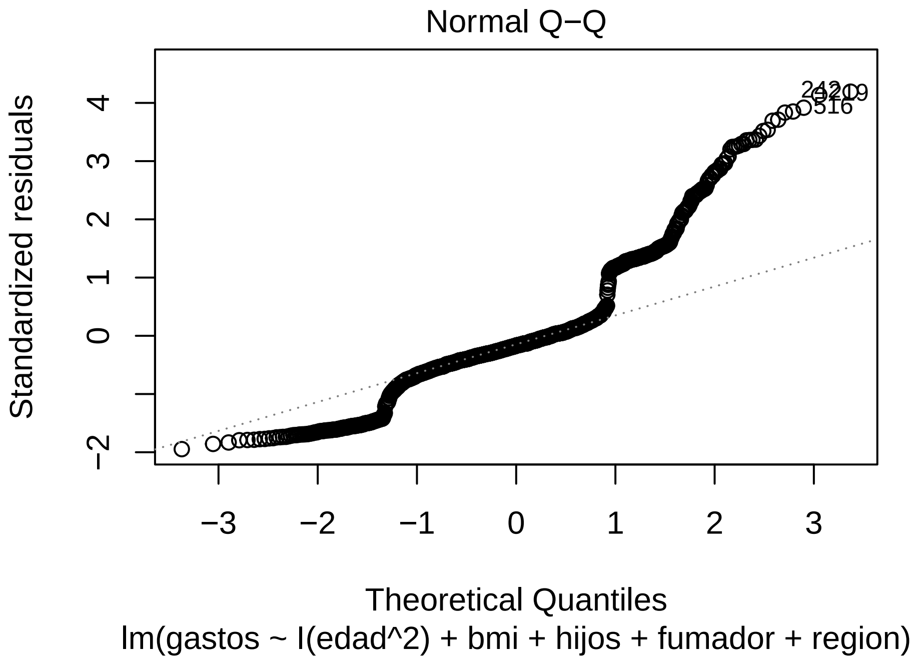
First and third plot show that for higher expenses residuals grow, meaning model does not satisfy homoscedasticity assumption.

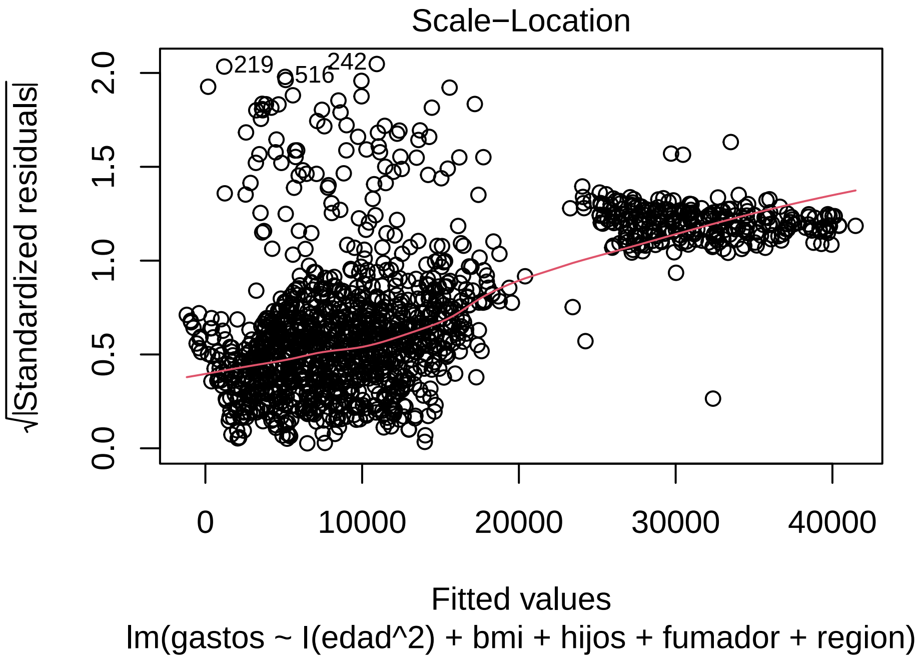
QQ-plot confirms that residuals does not follow normal distribution.

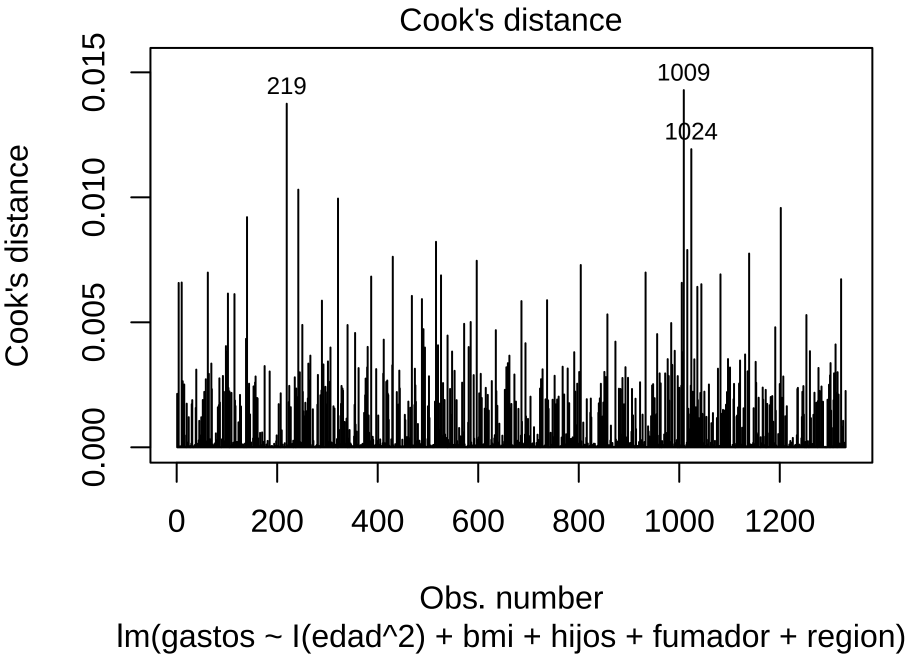
Last plot shows existance of outliers, however most significant 3 does not differ much from many others.

```
plot(m3, which=1:4)
```




Normal Q−Q
Standardized residuals
−2
0
1
2
3
4
−3
−2
−1
0
1
2
3
242
219
516
Theoretical Quantiles
lm(gastos ~ I(edad^2) + bmi + hijos + fumador + region)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
219
516
242
0
10000
20000
30000
40000
Fitted values
lm(gastos ~ I(edad^2) + bmi + hijos + fumador + region)



Models comparison using ANOVA

> *ToDo: Use anova to see which model of those created is more interesting.*

Anova test confirms that model m2 is better than m1, as well as that model m3 is better than m2. However, in model m3, despite caltalan region being significant enough to justify its addition to the model, model without it (m3_sr) is better than model that includes it (m3).

```
anova(m1, m2)
## Analysis of Variance Table
##
## Model 1: gastos ~ fumador
## Model 2: gastos ~ fumador + I(edad^2)
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   1329 6.9687e+10
## 2   1328 4.9815e+10  1 1.9872e+10 529.76 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
anova(m2, m3)
## Analysis of Variance Table
##
## Model 1: gastos ~ fumador + I(edad^2)
## Model 2: gastos ~ I(edad^2) + bmi + hijos + fumador + region
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   1328 4.9815e+10
## 2   1323 4.4368e+10  5 5447412982 32.487 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m2, m3_sr)
## Analysis of Variance Table
##
## Model 1: gastos ~ fumador + I(edad^2)
## Model 2: gastos ~ I(edad^2) + bmi + hijos + fumador
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   1328 4.9815e+10
## 2   1326 4.4581e+10  2 5234220130 77.842 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
anova(m3_sr, m3)
## Analysis of Variance Table
##
## Model 1: gastos ~ I(edad^2) + bmi + hijos + fumador
## Model 2: gastos ~ I(edad^2) + bmi + hijos + fumador + region
##   Res.Df        RSS Df Sum of Sq      F  Pr(>F)
## 1   1326 4.4581e+10
## 2   1323 4.4368e+10  3 213192852 2.1191 0.09601 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary

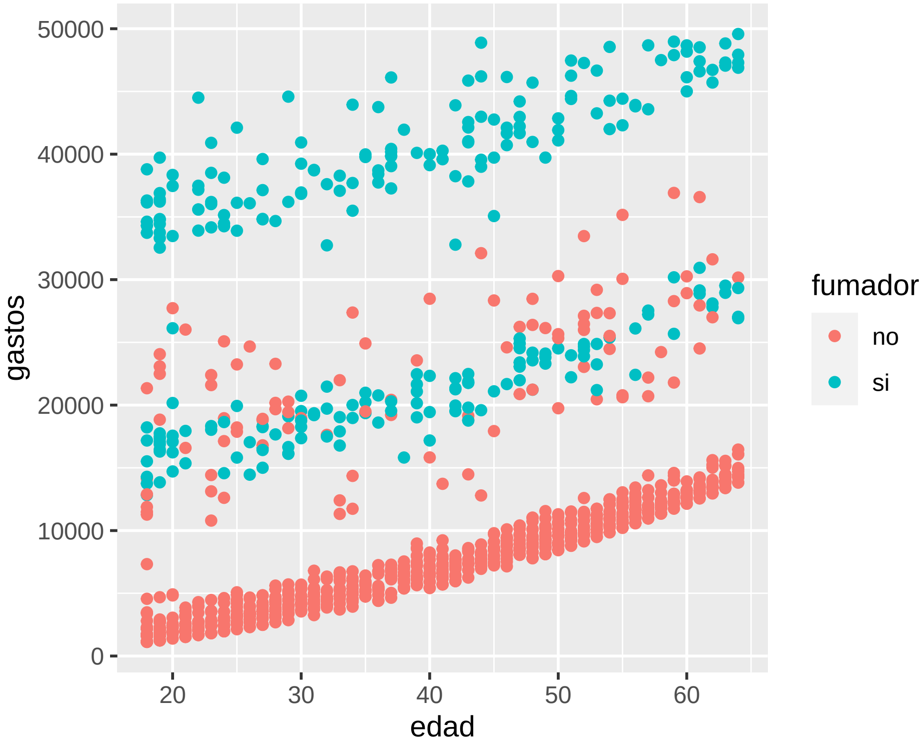
From all of the prepared models, m3_sr is the best one. However, it still explains only 75% of variance in predicted value and has error rate of about 44%. Moreover, it's residuals does not follow normal distribution and model does not satisfy homoscedasticity assumption.

To construct a better model, possibly removal of more outliers is necessary. However, it was tested that removal of samples with BMI values over 45, and expense values over 25,000 does not improve the model. Only radical filtering for expenses of values below 15,000, which removes smokers and most non-smoker outliers, resulted in model of 90% Adj. R^2. However, that removes important population from the model.

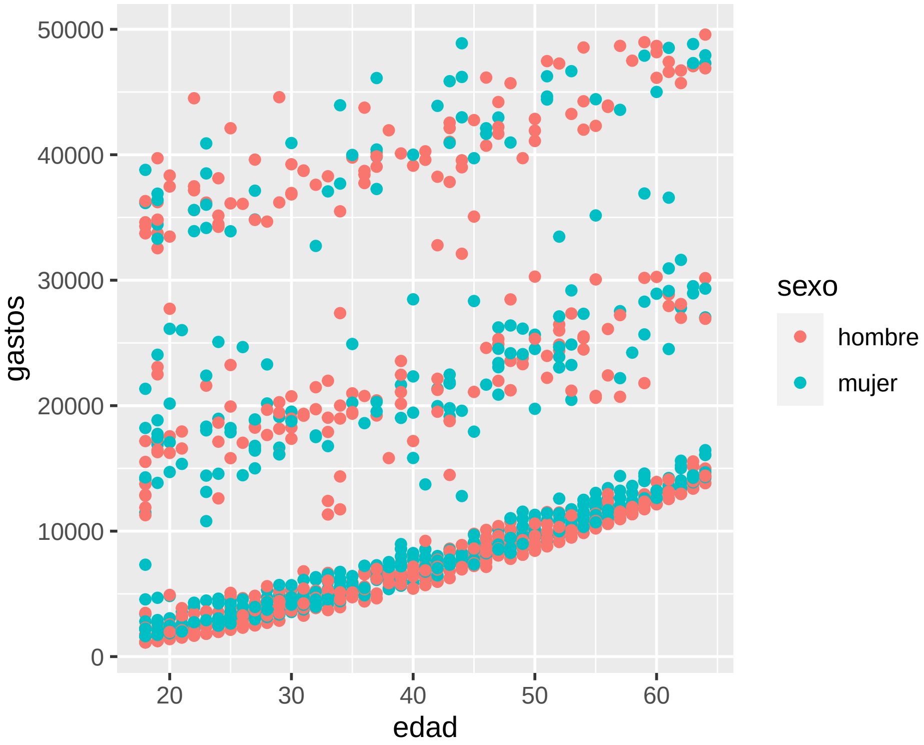
Therefore, to achieve better model, it may be necessary to construct a separate model for non-smokers only. To include smokers, possibly more information is required.

As can be seen on graphs below, there is clear linear relationship for between age and costs both for smokers and non-smokers. If the middle line didn't existed, is should have been easy to create a model based solely on age and smoking predictors. However, based on the analysis of all variables, there is no clear pattern explaining exsistance of mixed line. It may be, that those are outliers, for both smokers and non smokers. However, similar distances between lines, and clear pattern, suggest that there is rather some factor, that was not discovered.

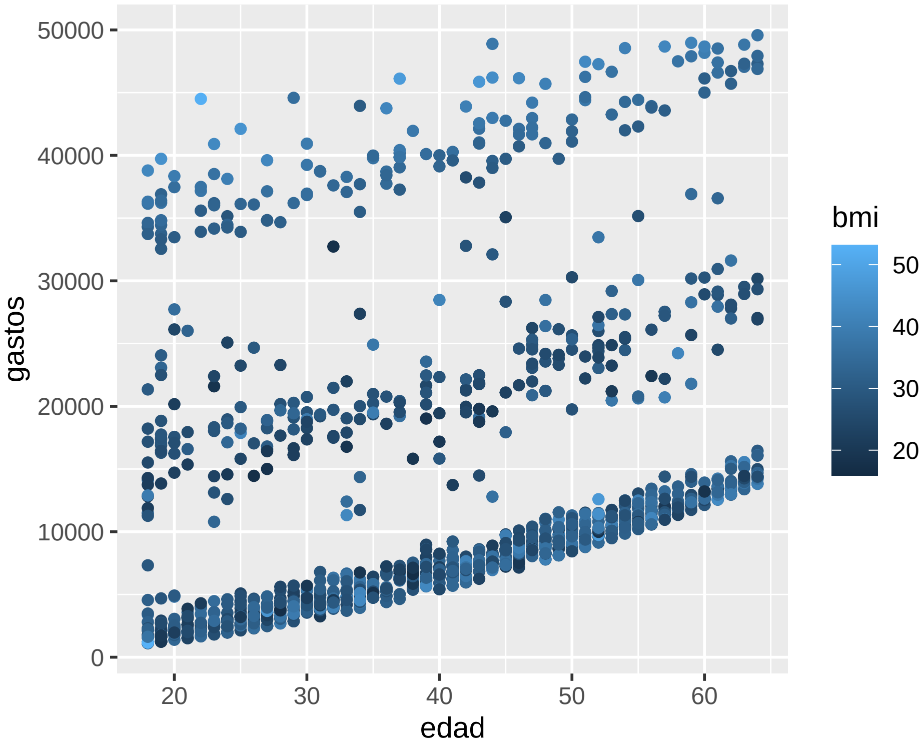
```
ggplot(data = riesgos_or, aes(x=edad, y=gastos)) + geom_point(aes(colour=fumador))
```



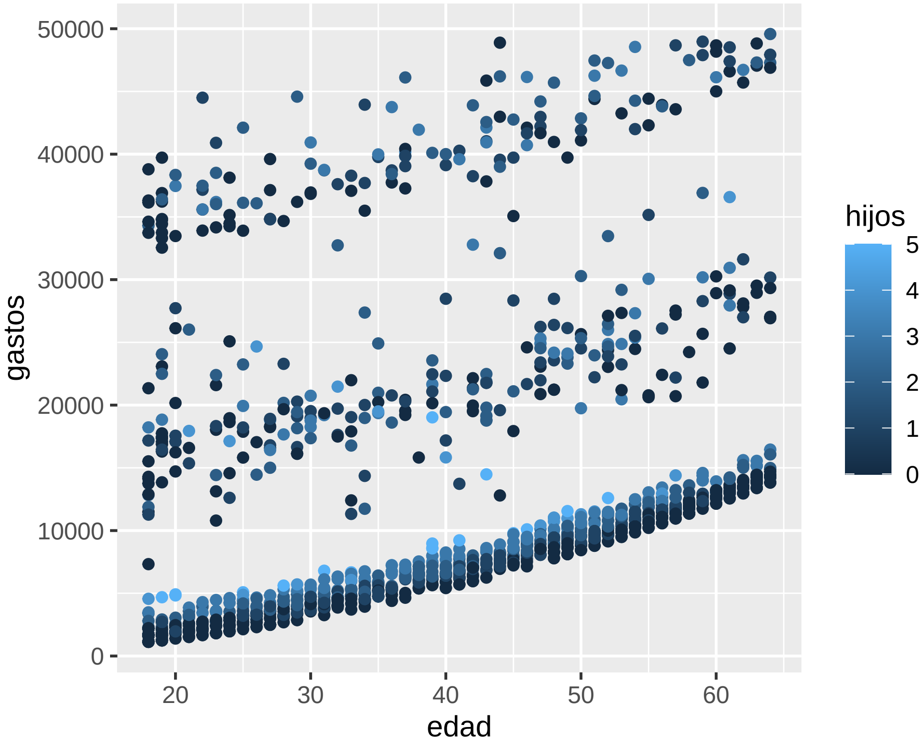
```
ggplot(data = riesgos_or, aes(x=edad, y=gastos)) + geom_point(aes(colour=sexo))
```



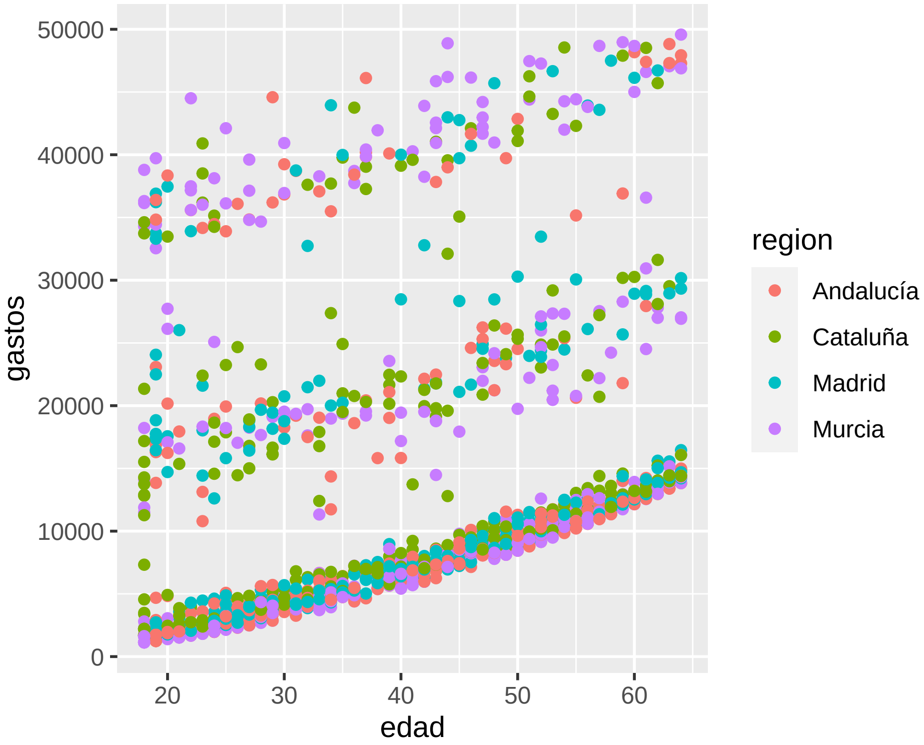
```
ggplot(data = riesgos_or, aes(x=edad, y=gastos)) + geom_point(aes(colour=bmi))
```



```
ggplot(data = riesgos_or, aes(x=edad, y=gastos)) + geom_point(aes(colour=hijos))
```



```
ggplot(data = riesgos_or, aes(x=edad, y=gastos)) + geom_point(aes(colour=region))
```


gastos
edad
region
● Andalucía
● Cataluña
● Madrid
● Murcia