University of Padua - Department of Physics and Astronomy
Degree course: Physics of Data
Course: Data Mining
Year: 2022-23
Professor in charge: A. Guolo

Luca Menti - 2063594 - luca.menti@studenti.unipd.it

`Exam's date: 27/06/2023`

Data Mining Report - Exam

1 Aim of the Report

The goal of this report is to analyze a genuine dataset and identify the most suitable model that explains the data. To accomplish this objective, I plan to first study a subset of the dataset, followed by the entire set. Utilizing various techniques, I will then determine which approach yields the most precise results for my analysis.

2 Analysis Techniques

2.1 Point 1

2.1.1 Multilinear Regression

Multilinear Regression is a statistical technique used to analyze the relationship between two or more independent variables and a dependent variable. It is an extension of Simple Linear Regression, which deals with only one independent variable. Multilinear Regression involves creating a mathematical model that calculates the best-fit line for the data by minimizing the sum of the squared differences between the observed and predicted values. This method is often used to make predictions or forecasts of the dependent variable based on the values of the independent variables. It is widely used in fields such as economics, finance, social sciences, and others for developing models and understanding complex relationships between variables.

2.1.2 Polynomial Regression

Polynomial Regression is a statistical technique used to model the relationship between a dependent variable Y and one or more independent variables X. It involves fitting a polynomial equation (quadratic, cubic, or higher-order) to the data instead of a linear relationship seen in simple and multiple linear regression. This method allows for the modeling of non-linear relationships between variables and can result in a better fit to the data than linear regression.

2.1.3 Backward Selection

We start with all the variables in the model; we remove the variables with the largest p-value, one by one. Go on until the remaining covariates have a small $p - value$. The procedure can't be used when $p > n$.

2.1.4 Generalized additive models (GAM)

In Gam we have flexibility in predicting Y using p covariates. Furthermore non-linear functions of the covariates are allowed and the additivity of the components is maintained. We also have applicability outside linear models. The prons are:

- we can model non-linear relationships between Y and X_j;
- predictions can potentially be more accurate than those from a linear model;
- we can still examine the effect of each X_j on Y individually, while holding all of the other variables fixed, as in the linear model;
- the smoothness of the function f_j for the variable X_j can be summarized via degrees of freedom.

While the cons:

- the model is additive;
- in case of many variables, important interactions can be missed;
- interaction functions of the form $f_{jk}(X_j, X_k)$ can be fit with different smoothers (it can be complex).

2.2 Point 2

2.2.1 Ridge Regression

With high-dimensional data (p very large w.r.t n or even $p > n$) the maximum likelihood estimates can be difficult to calculate, it can be not unique, it can have a large associated standard error and in general there are problems of identifiability and efficiency. One possible solution are shrinkage methos and one of the is Ridge Regression. Shrinkage methods :

- are useful to regularize the estimation process;
- they shrink the estimates of the coefficients towards zero;
- in this way, the fitting of the model is improved the variability associated to the estimates is smaller.

In Ridge regression there is a small shrinkage penalty when the coefficients of model β_j are close to zero. In particular for $\lambda = 0$ the penalty term has no effect, ridge regression will produce the least squares estimates, while for $\lambda \longrightarrow +\infty$ the impact of the shrinkage penalty grows, ridge regression coefficient estimates will approach zero. The choice of λ is a crucial point in penalized regression. If λ is too small there is no substantial penalization and the estimate is close to least squares estimate. If λ is too large there is too much penalization and the estimates are shrunk to zero. A good value of λ is chosen typically using cross validation.

2.2.2 Lasso

Lasso is a recent alternative to ridge regression which does not select variables. There is a small shrinkage penalty if β_j are close to zero or even equal to zero. Furthermore, Lasso performs variable selection as it forces some of the coefficient estimates to be zero for sufficiently large λ and it is easier to interpret than ridge regression. By the way the disadvantage is that the estimates are not in closed form, neither are the associated variances.

2.2.3 Automatic selection-Stepwise selection

We Evaluate a set of models and construct a rank based on one criterion or more criteria. The pros are:

- Quick evaluation of a large number of models
- Useful to have an initial idea about the relationships among the variables

On the other hand, the cons are linked to open problems such as:

- the variability associated to the model choice is not accounted for estimates biased towards zero, small standard errors, t and F statistics are far from the classical distribution and so on.
- More importantly, it is a blind procedure: it does not allow to think about the choice of a model

Forward selection In Forward selection we :

- search for the best model with one covariate
- search for the best model with two covariates constructed upon the previous one
- and so on
- research on a subset of $1 + \frac{p(p+1)}{2}$ models

Backward selection

In Backward selection we :

- have alternative to forward selection
- start from the model with all the covariates
- eliminate the nonsignificant covariates one at a time
- research on a subset of $1 + \frac{p(p+1)}{2}$ models

Hybrid selection

In Hybrid selection we :

- add covariates as in forward selection
- but remove them when they do not improve on the model
- same spirit of best subsets selection

2.2.4 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique used in machine learning and data analysis. It is used to reduce the dimensionality of large datasets by transforming a large set of variables into a smaller one that still contains most of the information in the large set. PCA is used in exploratory data analysis and for making predictive models. It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. The first principal component can equivalently be defined as a direction that maximizes the variance of the projected data. PCA creates variables that are linear combinations of the original variables. The new variables have the advantage of being uncorrelated and capturing the maximum amount of variation in the original dataset. In conclusion, PCA is a linear dimensionality reduction technique that transforms a set of correlated variables into a smaller number of uncorrelated variables called principal components while retaining as much of the variation in the original dataset as possible.

3 Dataset

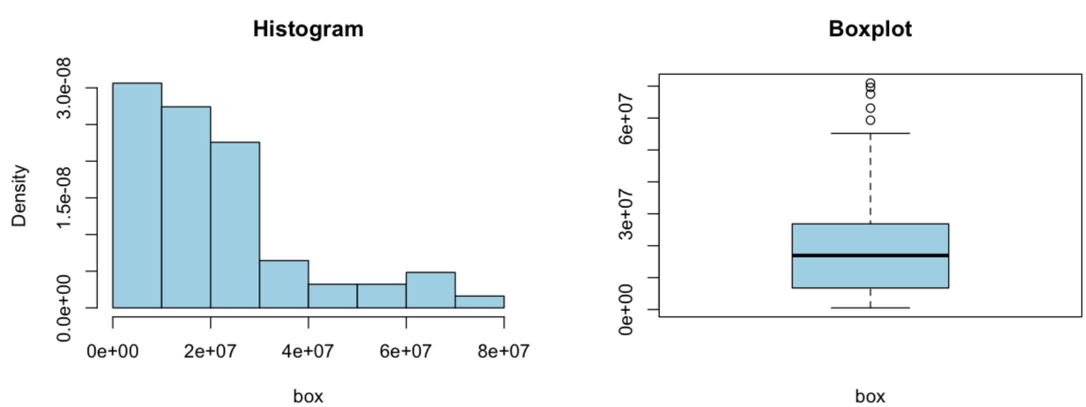
We see that our dataset is mad of $n = 62$ and $p = 13$. These are its main features of subset of plot (first 3 rows).

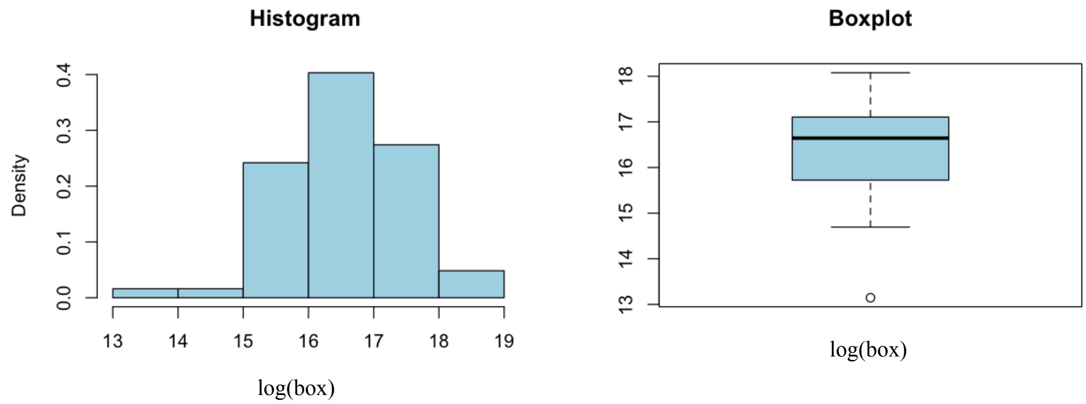
Let's also plot the histogram of box in order to check normality and the boxplot, otherwise let's apply a log transformation. It easy to see that is better consider a log-tranformation. So I will refer to box variable as log(box) in the following analysis.

box	budget	animated	cmngsoon
19167085	28.0	FALSE	10
63106589	150.0	TRUE	59
5401605	37.4	FALSE	24

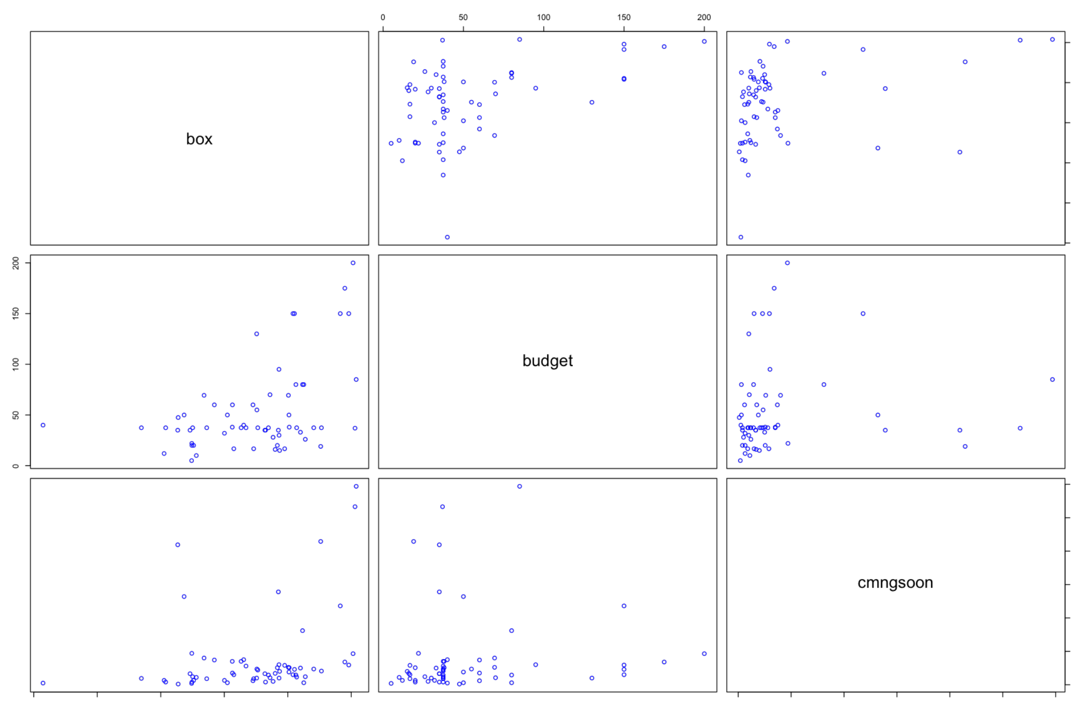
```
      box                budget         animated      cmngsoon
 Min.   :  511920   Min.   :  5.00   FALSE:56   Min.   :  2.00
 1st Qu.: 6956492   1st Qu.: 30.50   TRUE : 6   1st Qu.: 19.25
 Median :16930926   Median : 37.40              Median : 36.50
 Mean   :20720651   Mean   : 53.29              Mean   : 78.21
 3rd Qu.:26696144   3rd Qu.: 60.00              3rd Qu.: 66.00
 Max.   :70950500   Max.   :200.00              Max.   :594.00
```

0





Let also check the relationship between Y and the covariates X. From the first plot we can see a kind of linear relationship (even if it is not so clear and evident) between box and budget.

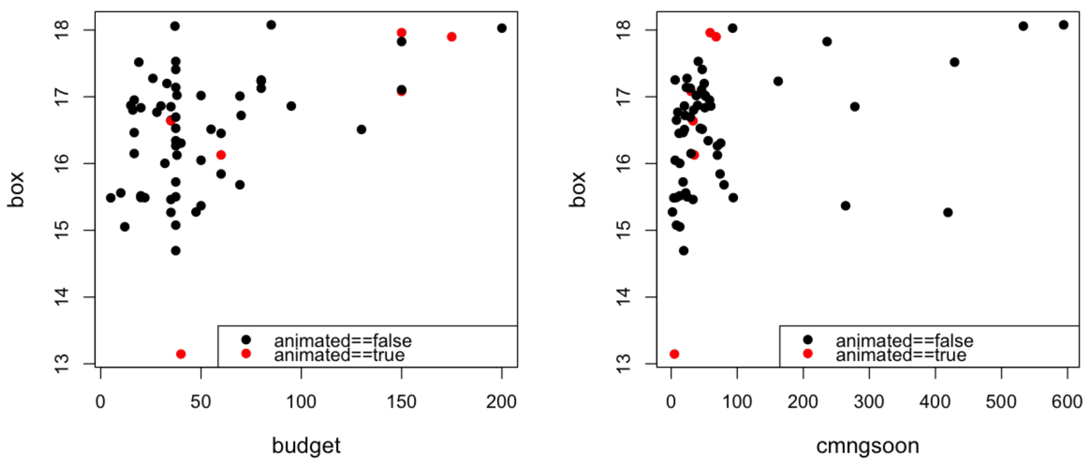


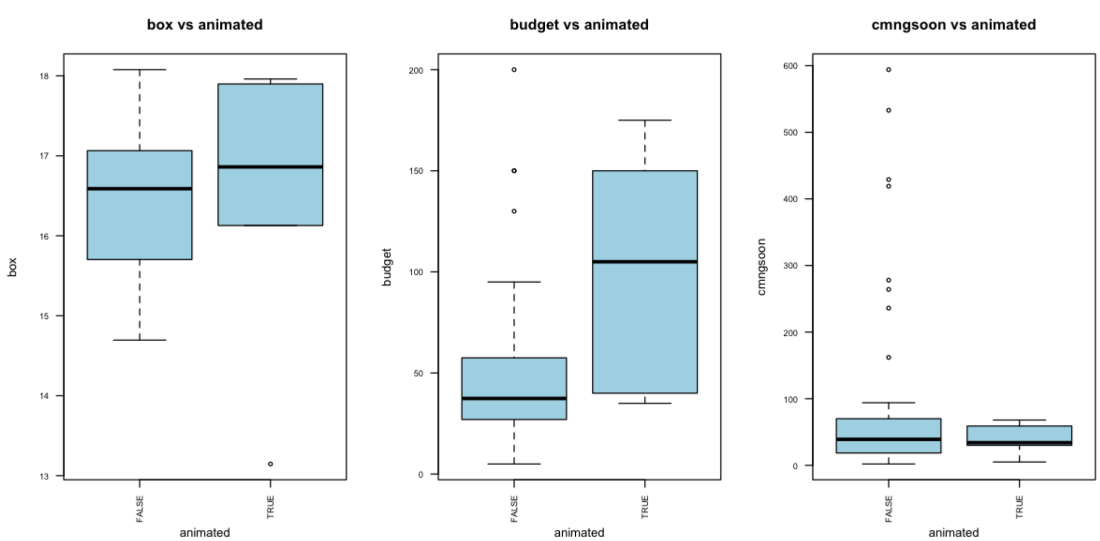
Regardign the interections between variables we see from the following plots that there is some overlapping (in the three scatter plots) so maybe there will not be interaction but we have to check it. As concern the boxplots:

- in the first one from left we don't have a clear difference between the two medians, the width is different, the whiskers are different and with animated = true box is higher and we have an outlier for TRUE.

- For the second one we can perform the same consideration saying that budget is higher for animated, the medians are different suggesting that there could be some interactions.The width is different between animated=TRUE and animated=FALSE and for FALSE we have outliers.
- we don't have a clear difference between the two medians, the width is different, the whiskers are different and with animated = FALSE cmngsoon is higher and we have and outliers for animated= FALSE

In order to check possible interactions we have to investigate more with regression.





4 Data Analysis - Point 1

4.1 Multiple Linear Regression

After this preliminary analysis we can apply a linear regression. Let's start with a model with all variables and interactions and then perform model selection base ond $P - value$.

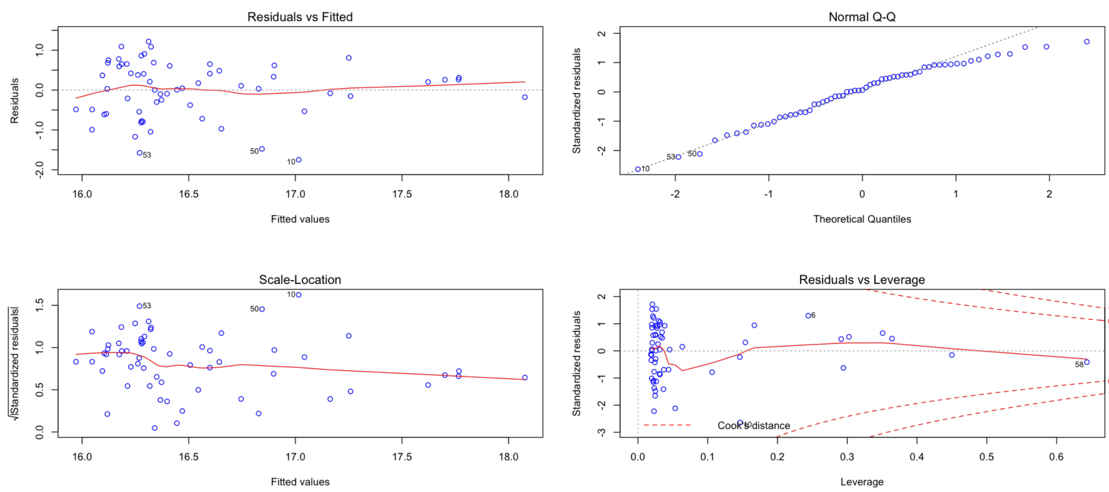
So I started with a model including also the interactions between the covariates then the final model I have obtained, that is $box = +15.9 + 0.008 * budget - 2.3 * animatedTRUE + 0.002 * cmngsoon + 0.05 * animated * cmngsoon$ is the following output. In the table below the 95% CI for coefficients is reported. I tried also with polynomial terms for cmngsoon and budget but they are not significant.

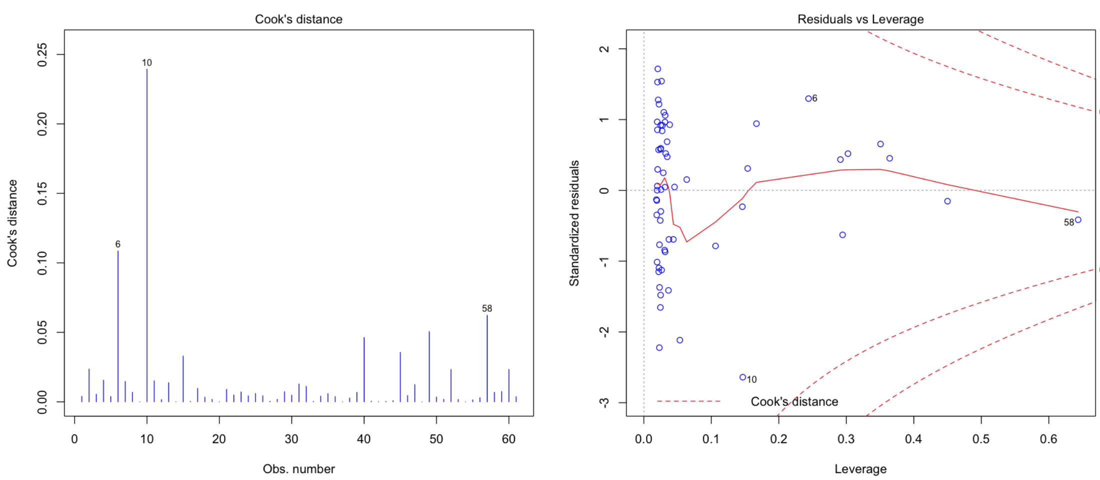
```
Call:
lm(formula = box ~ budget + animated + cmngsoon + animated:cmngsoon,
    data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.75479 -0.53717  0.07359  0.60779  1.21553

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           15.9340368  0.1652971  96.396  < 2e-16 ***
budget                 0.0080583  0.0025631   3.144 0.002646 **
animatedTRUE          -2.3436926  0.6569115  -3.568 0.000738 ***
cmngsoon               0.0019240  0.0007752   2.482 0.016031 *
animatedTRUE:cmngsoon  0.0519782  0.0157352   3.303 0.001654 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7437 on 57 degrees of freedom
Multiple R-squared:  0.4188,       Adjusted R-squared:  0.378
F-statistic: 10.27 on 4 and 57 DF,  p-value: 2.487e-06
```

	2.5 %	97.5 %
(Intercept)	15.6030351619	16.265038470
budget	0.0029257614	0.013190791
animatedTRUE	-3.6591351480	-1.028250123
cmngsoon	0.0003718007	0.003476288
animatedTRUE:cmngsoon	0.0204690696	0.083487393

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
55	31.37372	NA	NA	NA	NA
57	31.52650	-2	-0.1527797	0.1339159	0.8749479

From anova we see that we keep model without all the terms and interaction (the initial one). Now we can judge also our model considering the residuals. The graph of residuals indicates that the model does not have an acceptable fit. In fact, the first graph (scatter plot of the residuals) shows a deterministic pattern . In addition, the mean of the residuals does not appear to be 0 and the variance of the residuals does not appear to be constant, as it should be based on the assumptions that the regression model places on the ε errors. Furthermore, the normality of the residuals is not satisfied as highlighted in the second graph: the empirical quantiles, in fact, do deviate from the theoretical quantiles of a standard normal (except for one of the tails). To complete the analysis of the residuals, outliers appear to be present as shown from Cook's distance>1. So let's remove it and see if something change.





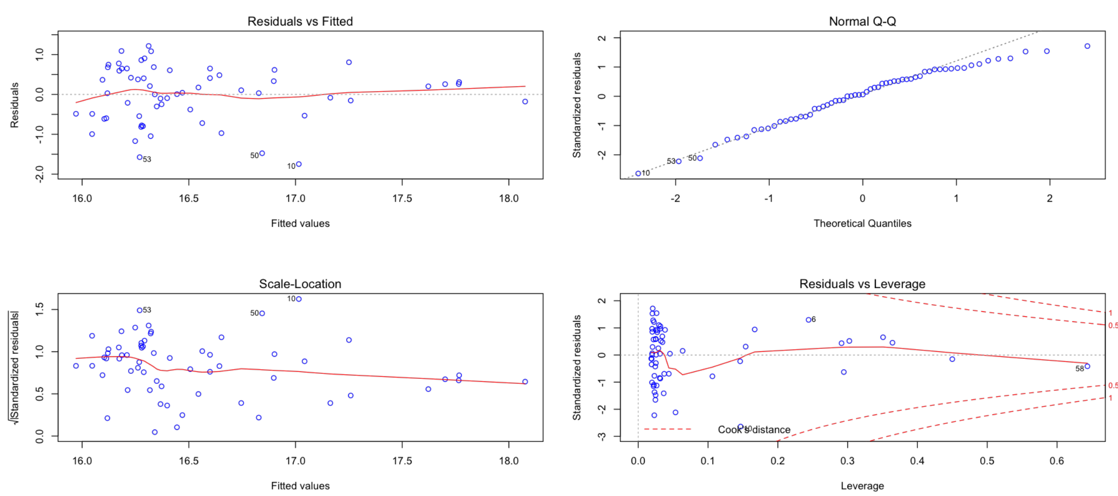
Removing the outliers leads the animated to lose significance in the model, so let's try another fit removing it and the interactions

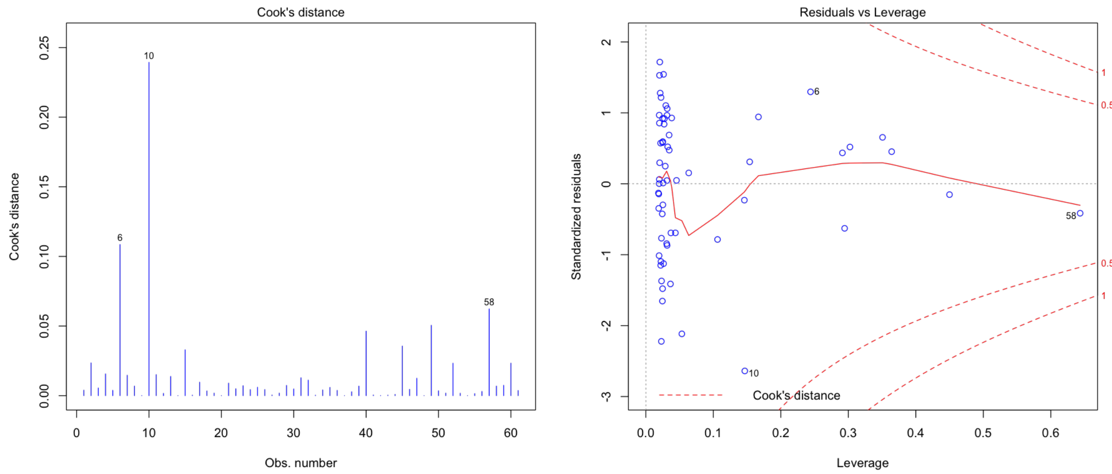
```
Call:
```

```
lm(formula = box ~ budget + cmngsoon, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7335 -0.4642  0.1049  0.5941  1.2165

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.589e+01  1.536e-01 103.476  < 2e-16 ***
budget      9.257e-03  2.126e-03   4.353 5.52e-05 ***
cmngsoon    1.878e-03  7.334e-04   2.560   0.0131 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7093 on 58 degrees of freedom
Multiple R-squared:  0.3204,	Adjusted R-squared:  0.297
F-statistic: 13.67 on 2 and 58 DF,  p-value: 1.364e-05
```

	2.5 %	97.5 %
(Intercept)	1.558282e+01	16.197599947
budget	5.000389e-03	0.013512714
cmngsoon	4.096702e-04	0.003345711

We obtain a model $box = 15.9 + 0.009 * budget + 0.003 * cmngsoon$. We see that the residuals in this case are goods and don't present a deterministic path.





Let's compare the model obtained with the outlier with the one without the outlier using anova. We see the model without animated (due to elimination of outlier value) is preferable. So our final model is : $box = 15.9 + 0.009 * budget + 0.003 * cmngsoon$

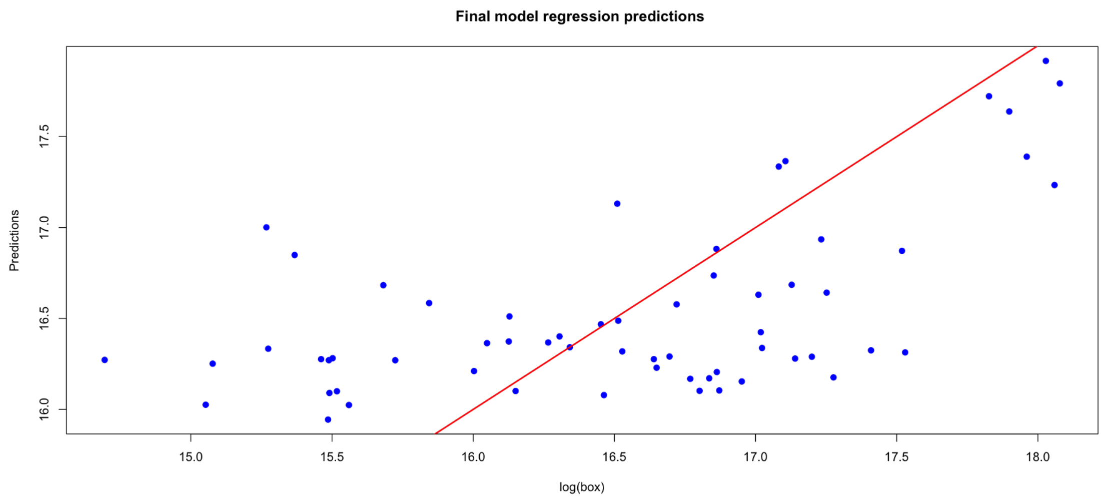
Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
56	28.79292	NA	NA	NA	NA
58	29.18307	-2	-0.3901517	0.3794074	0.6860122

Let's also try with smooth splines to see if we obtaine better model. In order to find the degrees of freedom I have used the cross validation (9 for budget 4 for cmngsoon). I report just the gam output for cmngsoon since it is the most interested(for budget splines is not useful based on p-value)

```
Warning message in model.matrix.default(mt, mf, contrasts):
‘‘non-list contrasts argument ignored’’


Call: gam(formula = box ~ budget + s(cmngsoon, 4), data = mydata)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.53341 -0.46280  0.07629  0.50166  1.09710

(Dispersion Parameter for gaussian family taken to be 0.4628)

    Null Deviance: 42.9432 on 60 degrees of freedom
Residual Deviance: 25.453 on 55 degrees of freedom
AIC: 133.794

Number of Local Scoring Iterations: 2

Anova for Parametric Effects
               Df  Sum Sq Mean Sq F value    Pr(>F)
budget          1  9.0506  9.0506 19.5569 4.652e-05 ***
```

```
s(cmngsoon, 4)  1  3.3608  3.3608  7.2622  0.009322 **
Residuals      55 25.4530  0.4628
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova for Nonparametric Effects
               Npar Df Npar F   Pr(F)
(Intercept)
budget
s(cmngsoon, 4)       3 2.6867 0.05532 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the output,based on $P - value$ we see that we don't need smoothing splines (even if we are in a borderline for cmngsoon since we have $p - value$=0.055 (greater than 0.05 by the way)). In the following plot we there are the predictions based on this model.



5 Conclusion Point 1

We have that:

- boxs depend positevly from budget

- boxs depend positevly from cmngsoon

The final model for the subset is: $box = 15.9 + 0.009 * budget + 0.003 * cmngsoon$

6 Data Analysis - Point 2

6.0.1 RIDGE

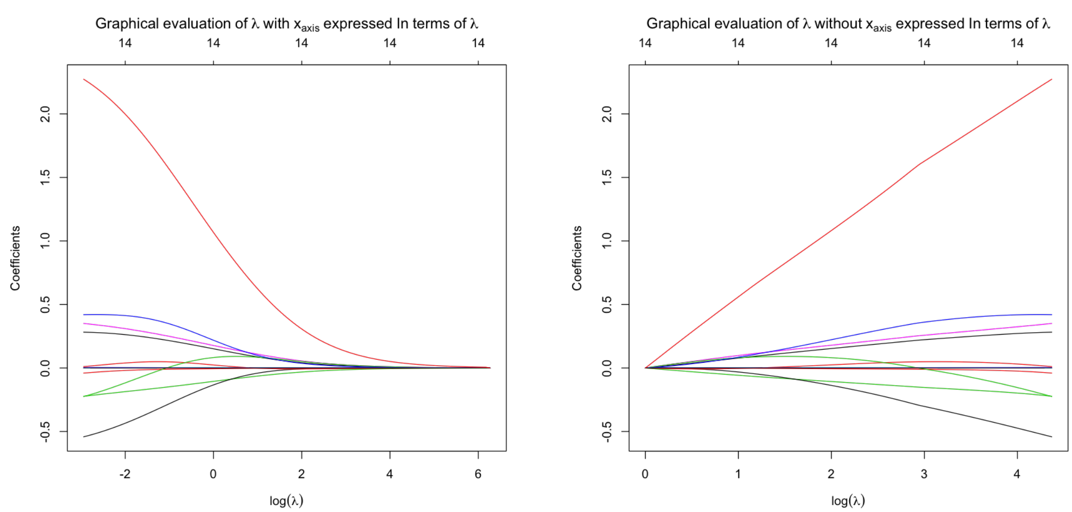
Let's consider all the dataset.

```
      box              mprating     budget          starpower        sequel
 Min.    :  511920   1: 2      Min.    :  5.00   Min.    : 0.00   FALSE:53
 1st Qu.: 6956492   2:15      1st Qu.: 30.50   1st Qu.:12.16   TRUE : 9
 Median :16930926   3:28      Median : 37.40   Median :18.07
 Mean   :20720651   4:17      Mean   : 53.29   Mean   :18.03
 3rd Qu.:26696144             3rd Qu.: 60.00   3rd Qu.:24.09
 Max.   :70950500             Max.   :200.00   Max.   :36.76
   action       comedy      animated      horror          addict          cmngsoon
 FALSE:48    FALSE:42    FALSE:56    FALSE:56    Min.    :  568   Min.    :  2.00
 TRUE :14    TRUE :20    TRUE : 6    TRUE : 6    1st Qu.: 1671   1st Qu.: 19.25
                                                 Median : 3480   Median : 36.50
                                                 Mean   : 5934   Mean   : 78.21
                                                 3rd Qu.: 7836   3rd Qu.: 66.00
                                                 Max.   :45866   Max.   :594.00
    fandango          cntwait
 Min.    :  35.0   Min.    :0.1500
 1st Qu.: 254.8   1st Qu.:0.3600
 Median : 430.5   Median :0.4850
 Mean   : 522.3   Mean   :0.4824
 3rd Qu.: 663.5   3rd Qu.:0.5875
 Max.   :1778.0   Max.   :0.7900
```

0

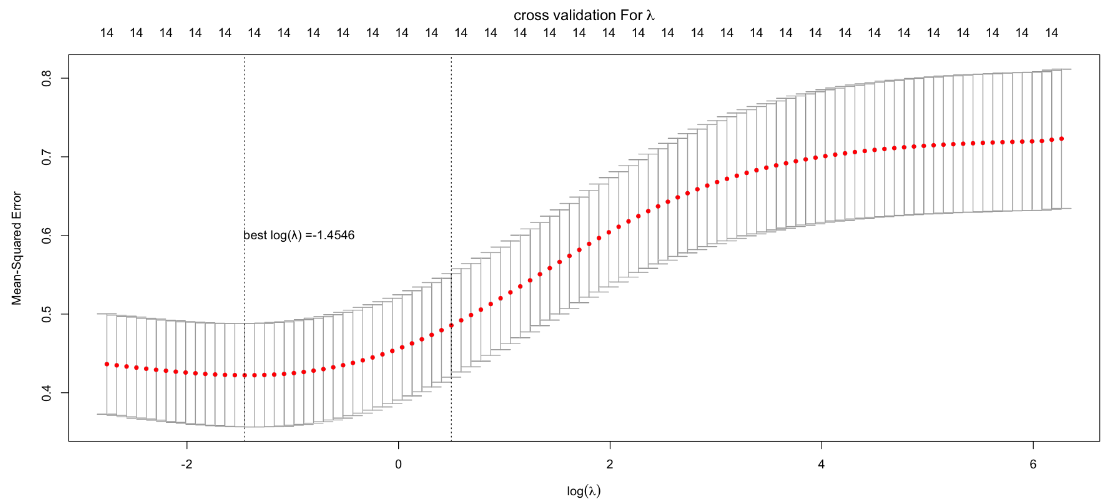
box	mprating	budget	starpower	sequel	action	comedy	animated	horror	addict	cmngsoon	fandango	cntwait
16.76871	4	28.0	19.83	FALSE	FALSE	TRUE	FALSE	FALSE	7860.5	10	144	0.49
17.96034	2	150.0	32.69	TRUE	FALSE	FALSE	TRUE	FALSE	5737.0	59	468	0.79
15.50221	4	37.4	15.69	FALSE	FALSE	TRUE	FALSE	FALSE	850.0	24	198	0.36

We see that 14 over the graph indicate the number of covariates entering the model as λ varies: 14 is repeated, as ridge regression is not a selection method.



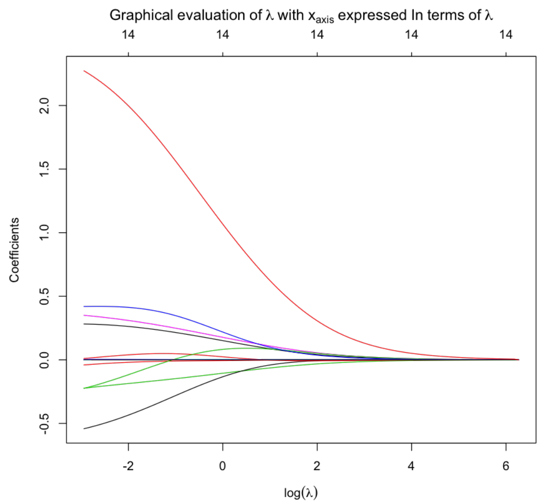
Now let's look for the best λ using cross validation. The plot below shows the values of *cvm* for each $log(\lambda)$ together with the associated confidence interval. The two dashed lines are the values of

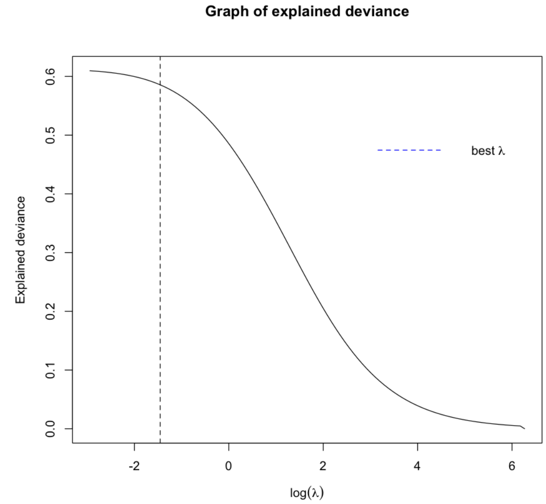
minimun $log(\lambda)$ and $log(\lambda)$ 1σ far from the minimum.
So the best λ from cross validation is: 0.23
And the MSE is: 0.42



Now we can Re-estimate the model using the best λ. Below we seen the coefficients of the model, graphical representation of the coefficients for the best λ and model deviance.
The maximum explained deviance is obtained for the minimum (best) λ and it is equal to:0.61

```
Call:  glmnet(x = X, y = y, alpha = 0, lambda = best.lambda)

     Df   %Dev Lambda
[1,] 14 0.5857 0.2335

15 x 1 sparse Matrix of class "dgCMatrix"
                        s0
(Intercept)   1.518598e+01
mprating2     2.395706e-01
mprating3    -1.278229e-02
mprating4    -1.657099e-01
budget        3.691353e-03
starpower     2.619087e-03
sequelTRUE    2.789123e-01
actionTRUE   -3.535089e-01
comedyTRUE    4.861000e-02
animatedTRUE -4.695996e-02
horrorTRUE    3.868300e-01
addict        2.164091e-05
cmngsoon      4.786077e-04
fandango      1.094037e-04
cntwait       1.774320e+00
```

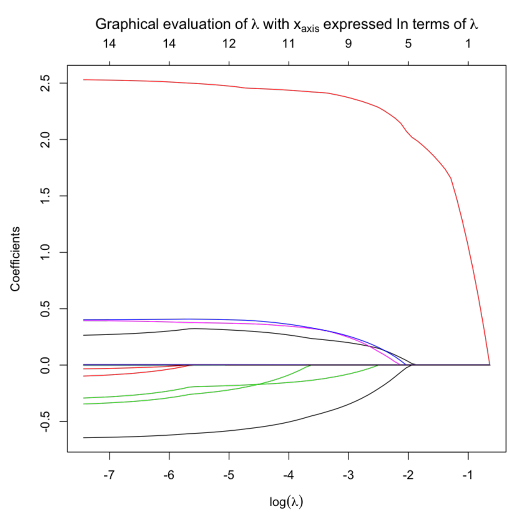


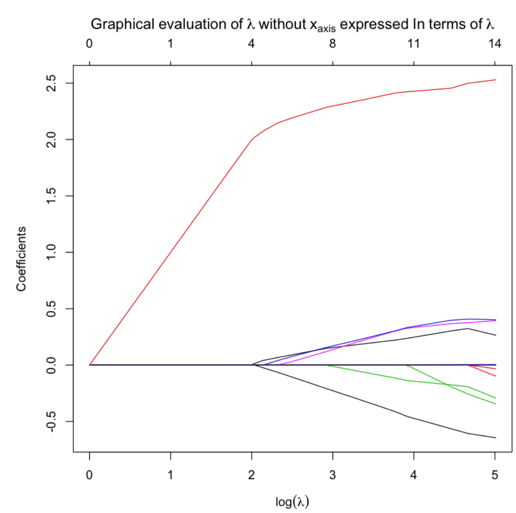


6.1 LASSO

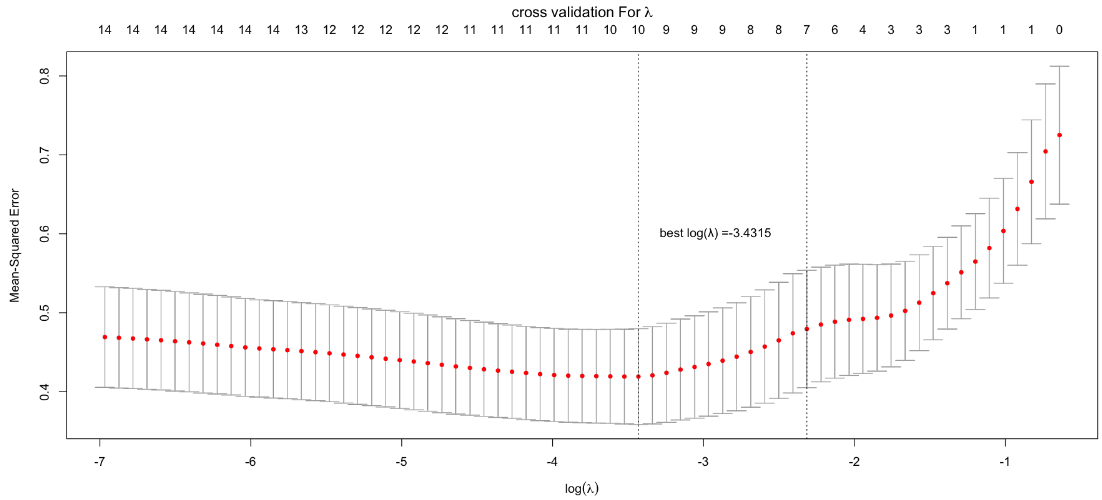
Let's perform the analysis usign lasso.

Above we can see the graphical evaluation of the coefficients associated to the covariates. We see that 14 over the graph indicate the number of covariates entering the model as λ varies:14 is not repeated, as lasso regression is a selection method.





Now let's look for the best λ using cross validation. So the best λ from cross validation is: 0.03
And the MSE is: 0.42



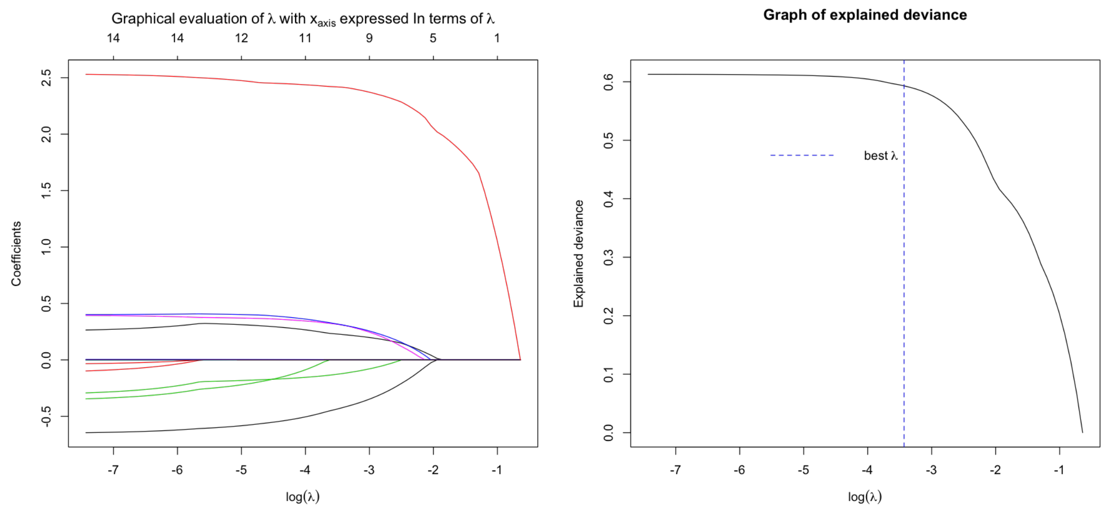
On the basis of MSE , the model fitted with lasso has got the same MSE by the way The resulting model with lasso is simpler. Now we can Re-estimate the model using the best λ. Below we seen the coefficients of the model, graphical representation of the coefficients for the best λ and model deviance.

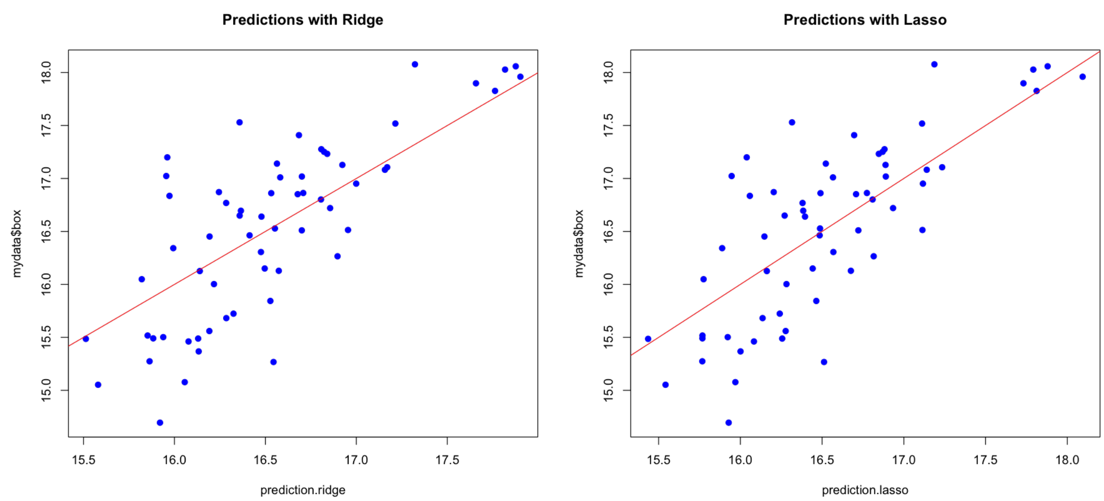
The maximum explained deviance is obtained for the minimum (best) λ and it is equal to: 0.61

Furthermore from the new coefficients we can see that some of the coefficients are zero, so the lasso performed a model selection. In particular thenot coefficients equal to 0 are= comedy, animated and starpower. Also mprating3 is set to 0.

```
Call:  glmnet(x = X, y = y, alpha = 1, lambda = best.lambda)

     Df  %Dev  Lambda
[1,] 10 0.593 0.03234

15 x 1 sparse Matrix of class "dgCMatrix"
                        s0
(Intercept)   1.503304e+01
mprating2     2.252352e-01
mprating3     .
mprating4    -1.216300e-01
budget        3.095863e-03
starpower     .
sequelTRUE    3.091528e-01
actionTRUE   -4.247813e-01
comedyTRUE    .
animatedTRUE  .
horrorTRUE    3.110032e-01
addict        2.485345e-05
cmngsoon      1.653581e-05
fandango      2.052026e-05
```

```
cntwait        2.416193e+00
```





Compare the results with those from the linear model. We have that :

-MSE for lasso is:0.42

- MSE for linear model is:0.42

No Net difference. Since there is substantial variable selection, lasso is more interesting and we keep it.

6.2 AUTOMATIC SELECTION

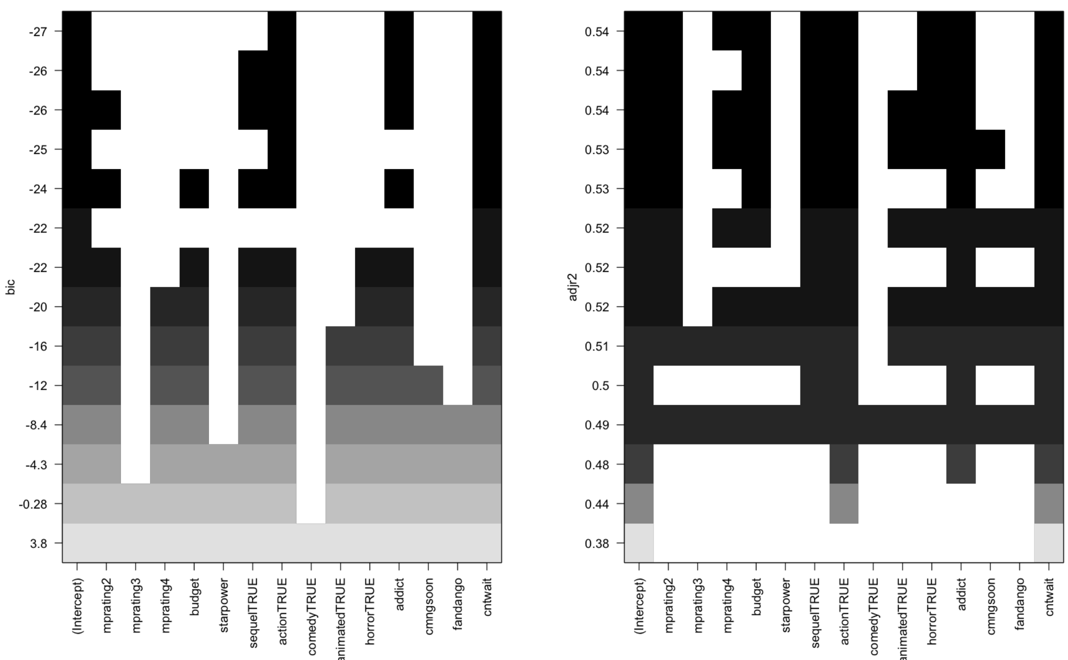
6.2.1 FORWARD SELECTION

Let's perform forward selection.

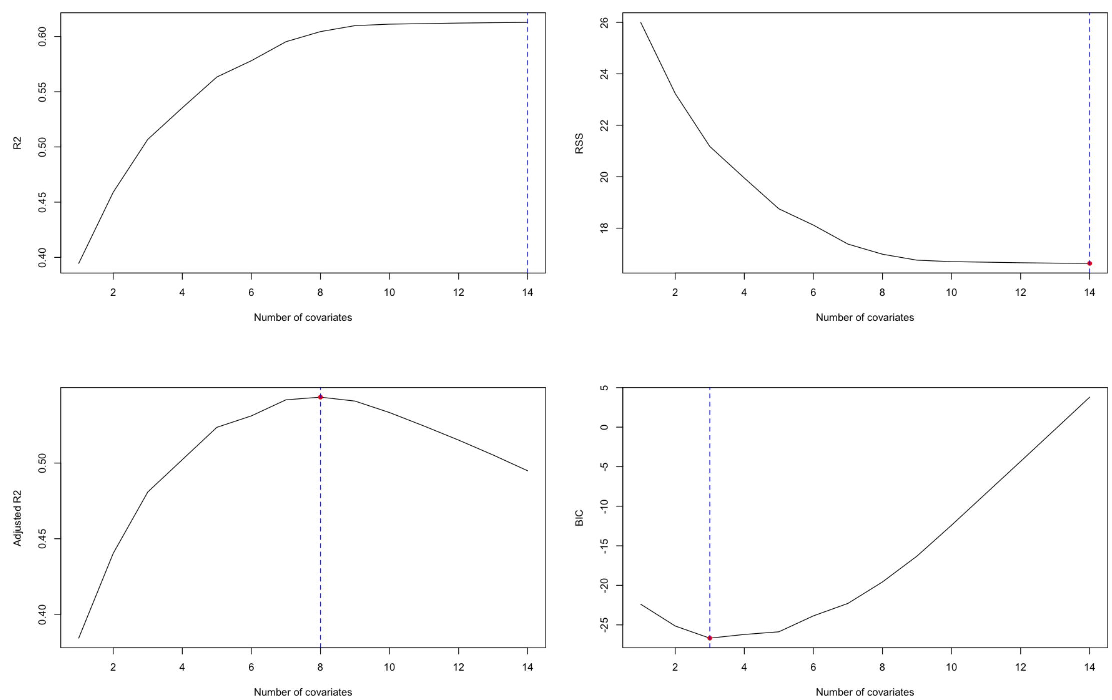
```
Subset selection object
Call: regsubsets.formula(box ~ ., data = mydata, nvmax = 17, method = "forward")
14 Variables  (and intercept)
             Forced in Forced out
mprating2        FALSE      FALSE
mprating3        FALSE      FALSE
mprating4        FALSE      FALSE
budget           FALSE      FALSE
starpower        FALSE      FALSE
sequelTRUE       FALSE      FALSE
actionTRUE       FALSE      FALSE
comedyTRUE       FALSE      FALSE
animatedTRUE     FALSE      FALSE
horrorTRUE       FALSE      FALSE
addict           FALSE      FALSE
cmngsoon         FALSE      FALSE
fandango         FALSE      FALSE
cntwait          FALSE      FALSE
1 subsets of each size up to 14
Selection Algorithm: forward
          mprating2 mprating3 mprating4 budget starpower sequelTRUE actionTRUE
1  ( 1 )  " "       " "       " "       " "    " "       " "        " "
2  ( 1 )  " "       " "       " "       " "    " "       " "        "*"
3  ( 1 )  " "       " "       " "       " "    " "       " "        "*"
4  ( 1 )  " "       " "       " "       " "    " "       "*"        "*"
5  ( 1 )  "*"       " "       " "       " "    " "       "*"        "*"
6  ( 1 )  "*"       " "       " "       "*"    " "       "*"        "*"
7  ( 1 )  "*"       " "       " "       "*"    " "       "*"        "*"
8  ( 1 )  "*"       " "       "*"       "*"    " "       "*"        "*"
9  ( 1 )  "*"       " "       "*"       "*"    " "       "*"        "*"
10  ( 1 ) "*"       " "       "*"       "*"    " "       "*"        "*"
11  ( 1 ) "*"       " "       "*"       "*"    " "       "*"        "*"
12  ( 1 ) "*"       " "       "*"       "*"    "*"       "*"        "*"
13  ( 1 ) "*"       "*"       "*"       "*"    "*"       "*"        "*"
14  ( 1 ) "*"       "*"       "*"       "*"    "*"       "*"        "*"
          comedyTRUE animatedTRUE horrorTRUE addict cmngsoon fandango cntwait
1  ( 1 )  " "        " "          " "        " "    " "      " "      "*"
2  ( 1 )  " "        " "          " "        " "    " "      " "      "*"
3  ( 1 )  " "        " "          " "        "*"    " "      " "      "*"
4  ( 1 )  " "        " "          " "        "*"    " "      " "      "*"
5  ( 1 )  " "        " "          " "        "*"    " "      " "      "*"
6  ( 1 )  " "        " "          " "        "*"    " "      " "      "*"
7  ( 1 )  " "        " "          "*"        "*"    " "      " "      "*"
```

```
8  ( 1 )  " "         " "          "*"         "*"    " "     " "     "*"
9  ( 1 )  " "         "*"          "*"         "*"    " "     " "     "*"
10  ( 1 ) " "         "*"          "*"         "*"    "*"     " "     "*"
11  ( 1 ) " "         "*"          "*"         "*"    "*"     "*"     "*"
12  ( 1 ) " "         "*"          "*"         "*"    "*"     "*"     "*"
13  ( 1 ) " "         "*"          "*"         "*"    "*"     "*"     "*"
14  ( 1 ) "*"         "*"          "*"         "*"    "*"     "*"     "*"
```

- the model with the smallest RSS is the model with 14 covariates
- usign BIC instead the best model includes 3 covariates

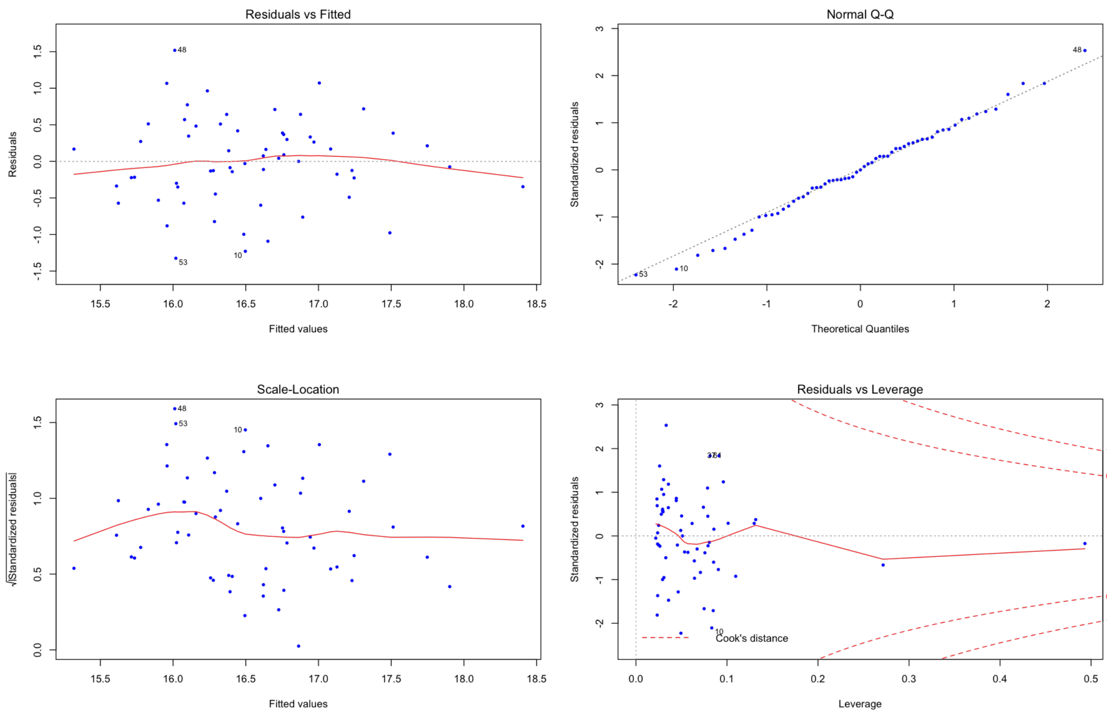


we see as computed before that the best model basing on BIC is 3



base on BIC we keep the model with the lowest BIC so with a number of covariates equal to : 3

```
Call:
lm(formula = box ~ action + addict + cntwait, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.32437 -0.34675 -0.00036  0.38525  1.51818

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.476e+01  2.747e-01  53.732  < 2e-16 ***
actionTRUE  -5.834e-01  2.043e-01  -2.855  0.00599 **
addict       2.644e-05  1.125e-05   2.351  0.02220 *
cntwait      3.592e+00  6.080e-01   5.908 2.03e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6095 on 57 degrees of freedom
Multiple R-squared:  0.5068,	Adjusted R-squared:  0.4809
F-statistic: 19.53 on 3 and 57 DF,  p-value: 7.855e-09
```

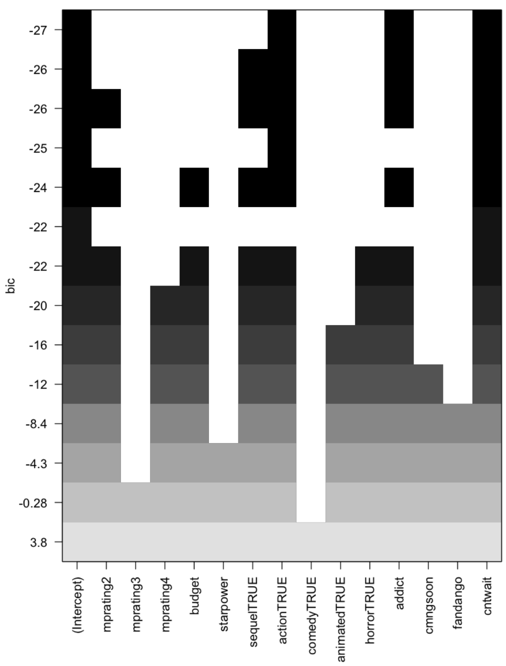


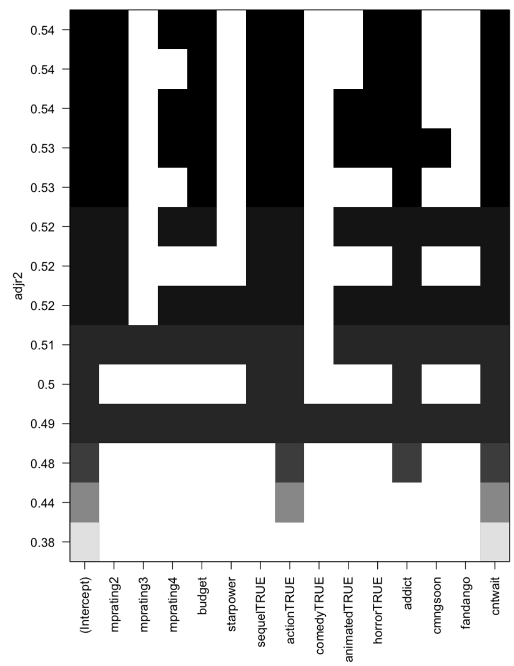
Now we can judge also our model considering the residuals. The graph of residuals indicates that the model does have a good fit. In fact, the first graph (scatter plot of the residuals) doesn't show a deterministic pattern . In addition, the mean of the residuals does appear to be 0 and the variance of the residuals does appear to be constant, as it should be based on the assumptions that the regression model places on the ε errors. Furthermore, the normality of the residuals is satisfied as highlighted in the second graph: the empirical quantiles in the tails , in fact, don't deviate from the theoretical quantiles of a standard normal. To complete the analysis of the residuals, no outliers appear to be present: although R highlights observations, these do not represent outlier observations since Cook's distance is not large.

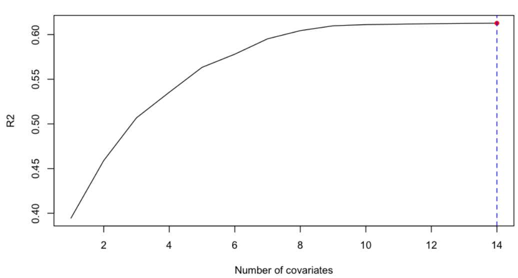
We also have that the MSE is:0.59

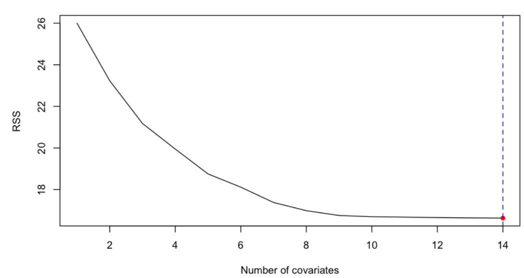
6.2.2 BACKWORD SELECTION

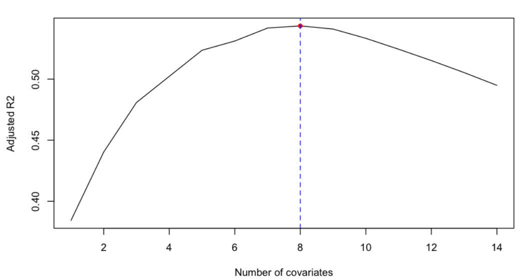
Let's perform backword selection.

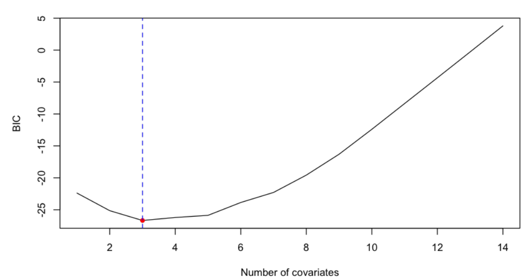

bic
-27
-26
-26
-25
-24
-22
-22
-20
-16
-12
-8.4
-4.3
-0.28
3.8
(Intercept)
mprating2
mprating3
mprating4
budget
starpower
sequelTRUE
actionTRUE
comedyTRUE
animatedTRUE
horrorTRUE
addict
cmngsoon
fandango
cntwait


adjr2
0.54
0.54
0.54
0.53
0.53
0.52
0.52
0.52
0.51
0.5
0.49
0.48
0.44
0.38
(Intercept)
mprating2
mprating3
mprating4
budget
starpower
sequelTRUE
actionTRUE
comedyTRUE
animatedTRUE
horrorTRUE
addict
cmngsoon
fandango
cntwait


R2
Number of covariates


RSS
Number of covariates

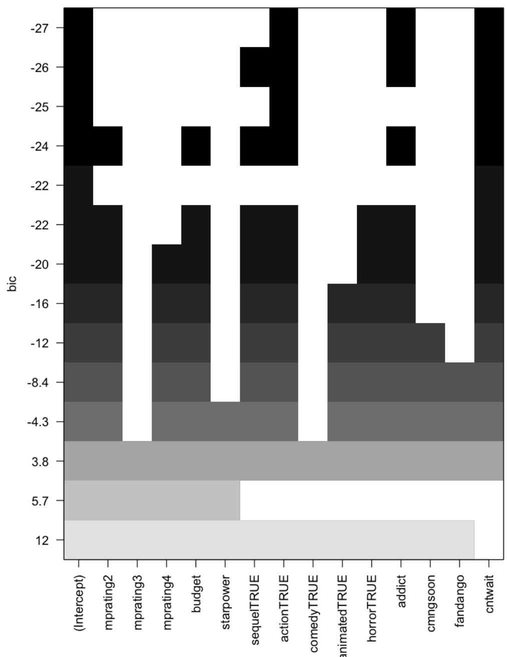

Adjusted R2
Number of covariates

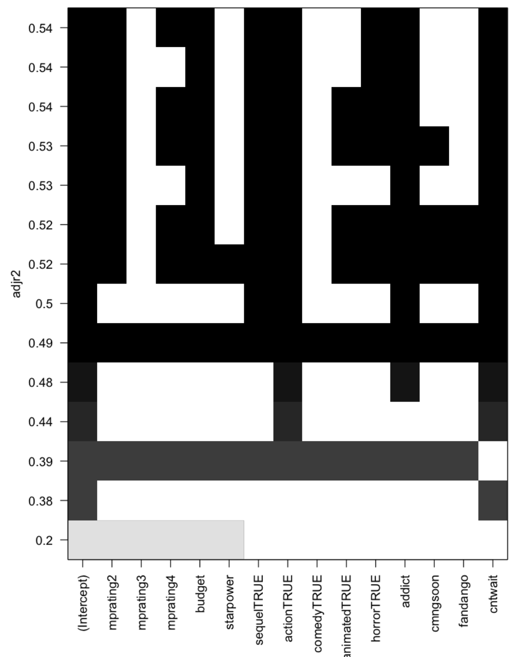

BIC
Number of covariates

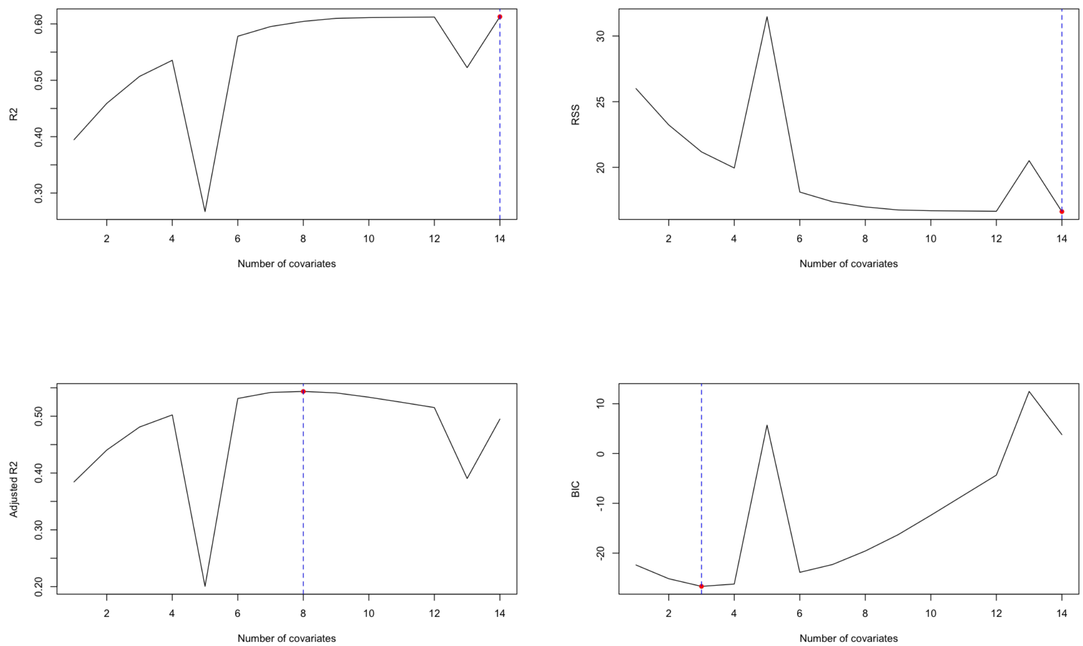
As you can see, forward and backward give us the same amount of covariates based on BIC. In fact in this case we have 3 covariates and the MSE is again 0.59 .

6.2.3 MIXED SELECTION

Let's perform mixed selection.







As you can see, forward ,backward and mixed selection give us the same amount of covariates based on BIC. In fact in this case we have 3 covariates and the MSE is again 0.59 .

6.2.4 PRINCIPAL COMPONENT ANALYSIS

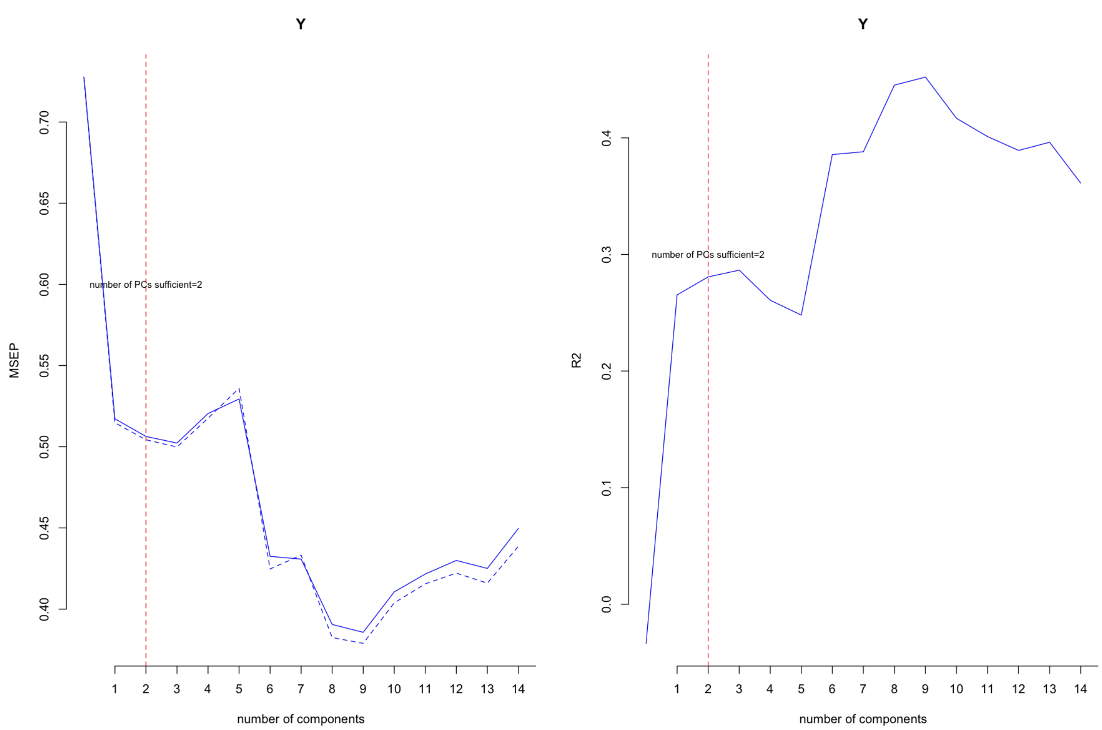
Let's consider Principal component analysis in order to see if it is useful. I set the seed at 222.

```
Attaching package: 'pls'

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

    loadings


Data:    X dimension: 61 14
         Y dimension: 61 1
Fit method: svdpc
Number of components considered: 14

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV           0.853   0.7192   0.7116   0.7087   0.7214   0.7276   0.6576
adjCV        0.853   0.7175   0.7101   0.7070   0.7194   0.7321   0.6517
```

```
       7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
CV      0.6563   0.6249   0.6211    0.6408    0.6493    0.6557     0.652
adjCV   0.6582   0.6184   0.6155    0.6355    0.6446    0.6498     0.645
       14 comps
CV       0.6706
adjCV    0.6624

TRAINING: % variance explained
     1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X      24.46    42.85    55.35    64.45    71.92    78.49    83.74    88.35
box    30.99    33.54    35.48    37.01    37.01    50.74    52.57    57.45
     9 comps  10 comps  11 comps  12 comps  13 comps  14 comps
X      91.61     94.56     96.66     98.42     99.71    100.00
box    58.06     58.16     58.16     60.21     61.21     61.28
```

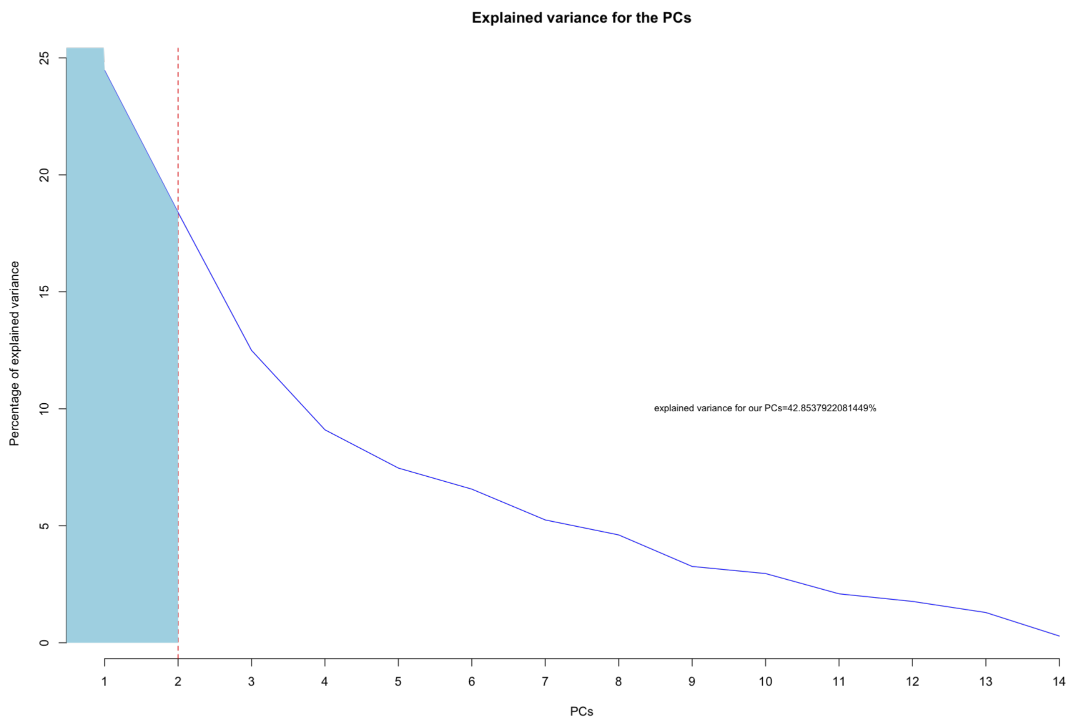
The output provides the result of the cross validation in terms of square root of the MSE for each number of PCs. -Choose the optimum through a graphical inspection of the results considering MSE and R^2. We see that the number of PCs needed is: 14 While the best number of components we can use for the analysis based on R comand SelectNcompo is : 2 We also have that the value of MSE is reported below.

```
       (Intercept)  2 comps
CV          0.7276   0.5064
adjCV       0.7276   0.5043
```

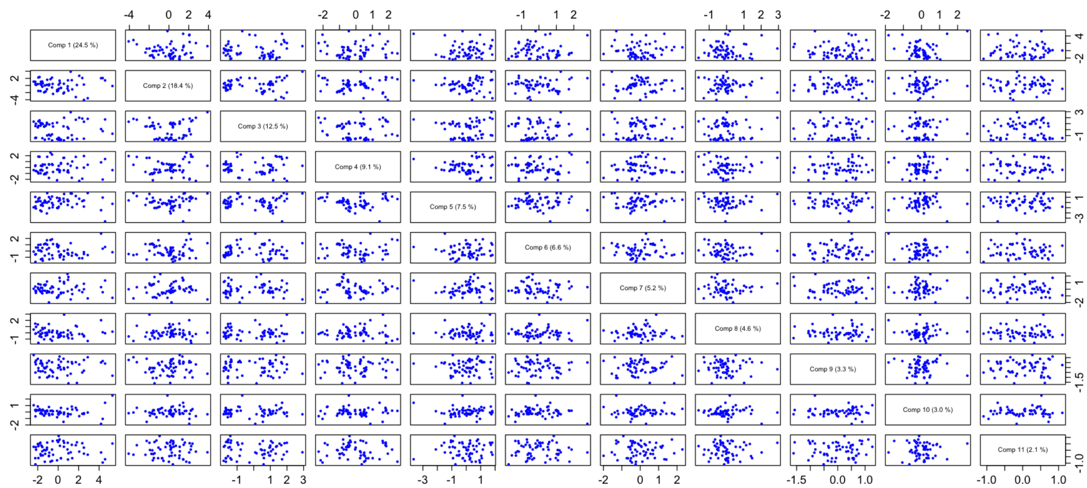


Let's look how much variance is explained by the 14 components in the plot below. While te explained variance for our 2 PCs obtained before is 43% that is a bit low. So we will also consider 11 components that are the same number obtained by lasso which lead to 94% of explained deviance.

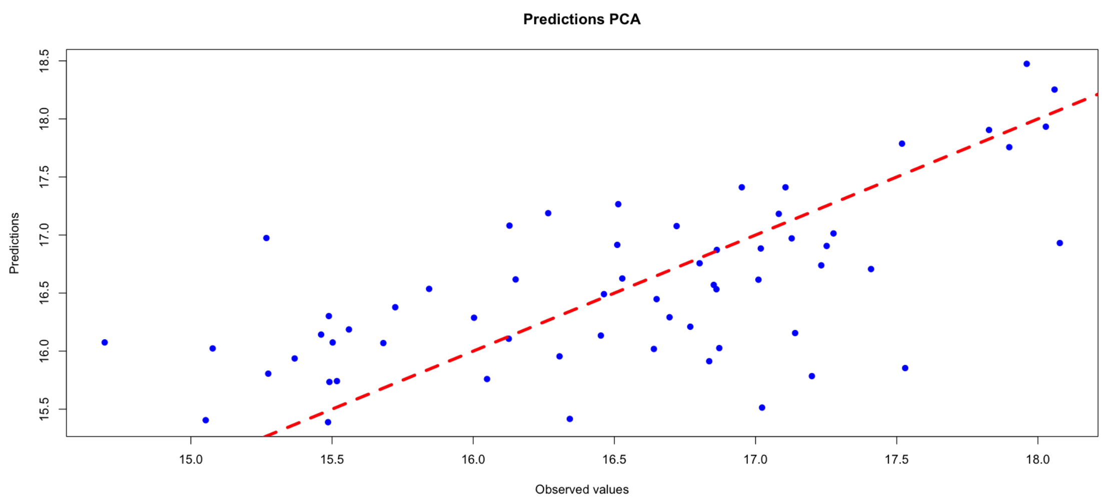
Comp 1 24.4608910912727 **Comp 2** 18.3929011168722 **Comp 3** 12.4991911845226 **Comp 4** 9.10185199488781 **Comp 5** 7.46658095613463 **Comp 6** 6.56616471024323 **Comp 7** 5.24916881017593 **Comp 8** 4.60878978003381 **Comp 9** 3.26151855999341 **Comp 10** 2.95755897338834 **Comp 11** 2.09127110876997 **Comp 12** 1.76656610355031 **Comp 13** 1.29086136351849 **Comp 14** 0.286684246636644



Let's now plot Plot the regression coefficients associated to the models with increasing PCs, from 1 to 11 We see that we have our 11 models. We look for the picks. As picks are higher as our model is better. The model with 11 comps give us the largest amount of informations (it is the better one).



Finally, evaluate the predictions from the model. Values around the bisector does suggest a good behavior of the model.



Finally the MSE of PCA is 0.42.

6.3 Conclusion Point 2

From The analysis of al dataset we have that :

- MSE for lasso is:0.42

-MSE for linear model is:0.42

- MSE for automatic selection :0.59

- MSE for PCA is:0.42

Since PCA is more useful for clustering, the best approach seems to be lasso due to the fact that it has got a less MSE and perform a variable selection. By the way it is important to consider that the MSE for automatic selection is obtained considering BIC criteria and just 3 variables. Maybe with more variables and different criteria the MSE coulde decrease.