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 contain information about characteristic of 62 movies 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. We have one qualitative variables "animated" and the others are continuos.

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	starpower	cmngsoon
191.67085	28.0	FALSE	19.83	10
631.06589	150.0	TRUE	32.69	59
54.01605	37.4	FALSE	15.69	24

```
     box                budget        animated    starpower          cmngsoon
 Min.   :  5.119   Min.   :  5.00   FALSE:56   Min.   : 0.00   Min.   :  2.00
 1st Qu.: 69.565   1st Qu.: 30.50   TRUE : 6   1st Qu.:12.16   1st Qu.: 19.25
 Median :169.309   Median : 37.40              Median :18.07   Median : 36.50
 Mean   :207.207   Mean   : 53.29              Mean   :18.03   Mean   : 78.21
 3rd Qu.:266.961   3rd Qu.: 60.00              3rd Qu.:24.09   3rd Qu.: 66.00
 Max.   :709.505   Max.   :200.00              Max.   :36.76   Max.   :594.00
```
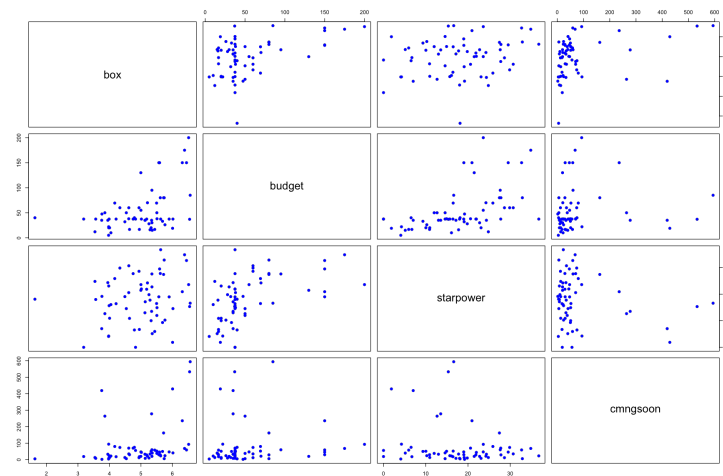

Histogram

Boxplot
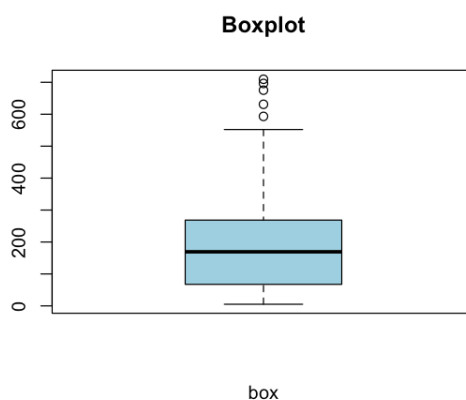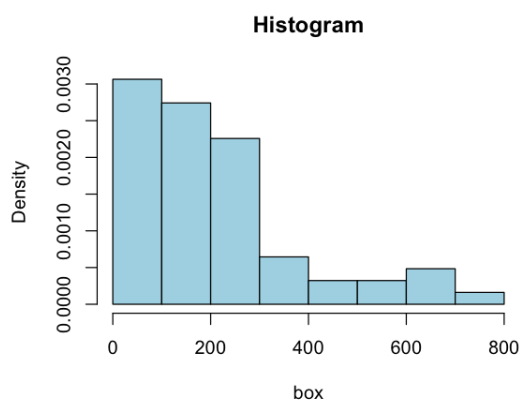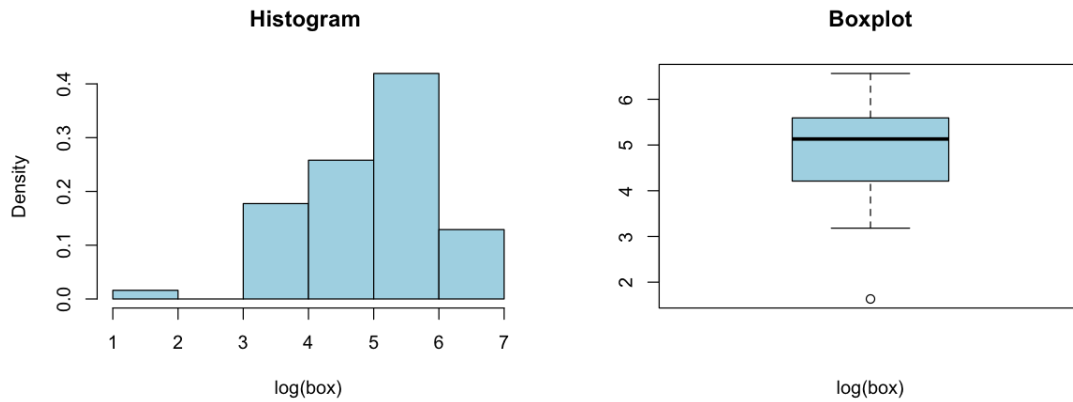
Let also check the relationship between box and the other covariates :

- between box and budget we see a kind of polynomial relationship
- between box and starpower not a particular one
- between box and cmngsoon we see a kind of linear relationship even if different points are spread around.

By the way it will be necessary to investigate more.

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 suggesting not interaction, 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 from left 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.

- For the third one from left we don't have a clear difference between the two medians suggesting not interactions, the width is different, the whiskers are different and with animated = FALSE cmngsoon is higher and we have and outliers for animated= FALSE

- For the fourth one from left we can perform the same consideration saying that budget is higher for animated=TRUE, the medians are different suggesting that there could be some interactions.The wihiskers are different between animated=TRUE and animated=FALSE and we don't have outliers.

4   Point 1- Data Analysis

In order to check possible interactions we have to investigate more with regression.
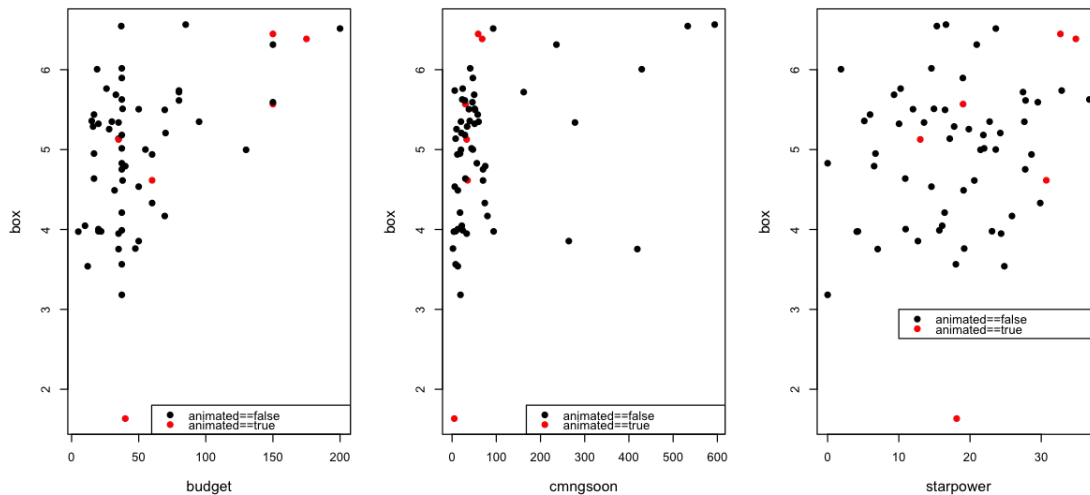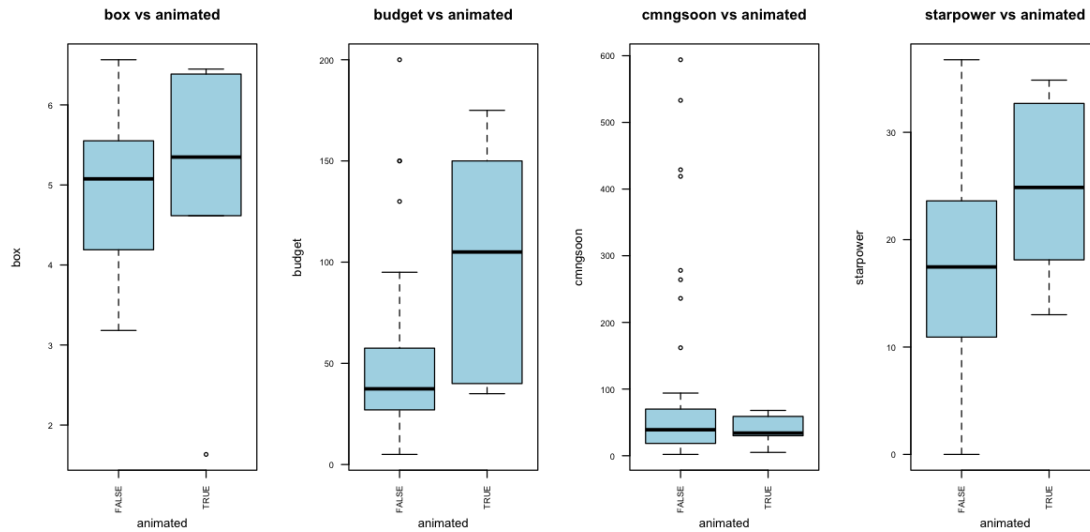
4.0.1 Multiple Linear Regression

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: $box = +4.421 + 0.008 budget -2.344 animated*$ $$ I_{animated=TRUE} $$ $+0.002 * cmngsoon + 0.052 * animated * cmngsoon$

is the following output. In the table below the 95% CI for coefficients is reported.

```
Call:
lm(formula = box ~ budget + cmngsoon + animated + cmngsoon:animated,
    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)          4.4211114  0.1652971  26.746  < 2e-16 ***
budget               0.0080583  0.0025631   3.144 0.002646 **
cmngsoon             0.0019240  0.0007752   2.482 0.016031 *
animatedTRUE        -2.3436926  0.6569115  -3.568 0.000738 ***
cmngsoon:animatedTRUE 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)	4.0901096969	4.752113005
budget	0.0029257614	0.013190791
cmngsoon	0.0003718007	0.003476288
animatedTRUE	-3.6591351480	-1.028250123
cmngsoon:animatedTRUE	0.0204690696	0.083487393

I tried also with polynomial terms for cmngsoon and budget but they are not significant as shown in the following output (I also tried removing as quadratic term for budget keeping it for cmngsoon and viceversa but nothing changes).

```
Call:
lm(formula = box ~ budget + cmngsoon + animated + I(budget^2) +
    I(cmngsoon^2) + cmngsoon:animated, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.68843 -0.57156  0.07404  0.61563  1.22495

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             4.500e+00  2.898e-01  15.527  < 2e-16 ***
budget                  7.299e-03  8.592e-03   0.849 0.399317
cmngsoon                9.640e-05  2.897e-03   0.033 0.973571
animatedTRUE           -2.384e+00  6.781e-01  -3.515 0.000888 ***
I(budget^2)             6.146e-06  4.642e-05   0.132 0.895146
I(cmngsoon^2)           3.641e-06  5.534e-06   0.658 0.513263
cmngsoon:animatedTRUE   5.242e-02  1.632e-02   3.213 0.002198 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7541 on 55 degrees of freedom
Multiple R-squared:  0.4234,        Adjusted R-squared:  0.3605
F-statistic: 6.731 on 6 and 55 DF,  p-value: 2.203e-05
```

From anova we see that we keep model without all the terms and interaction (the initial one).

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
51	29.82083	NA	NA	NA	NA
57	31.52650	-6	-1.705674	0.4861779	0.8155907

Now we can judge also our model considering the residuals. The graph of residuals indicates that the model does not have an pretty good fit. In fact, the first graph (scatter plot of the residuals) doesn't show a non- 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,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.
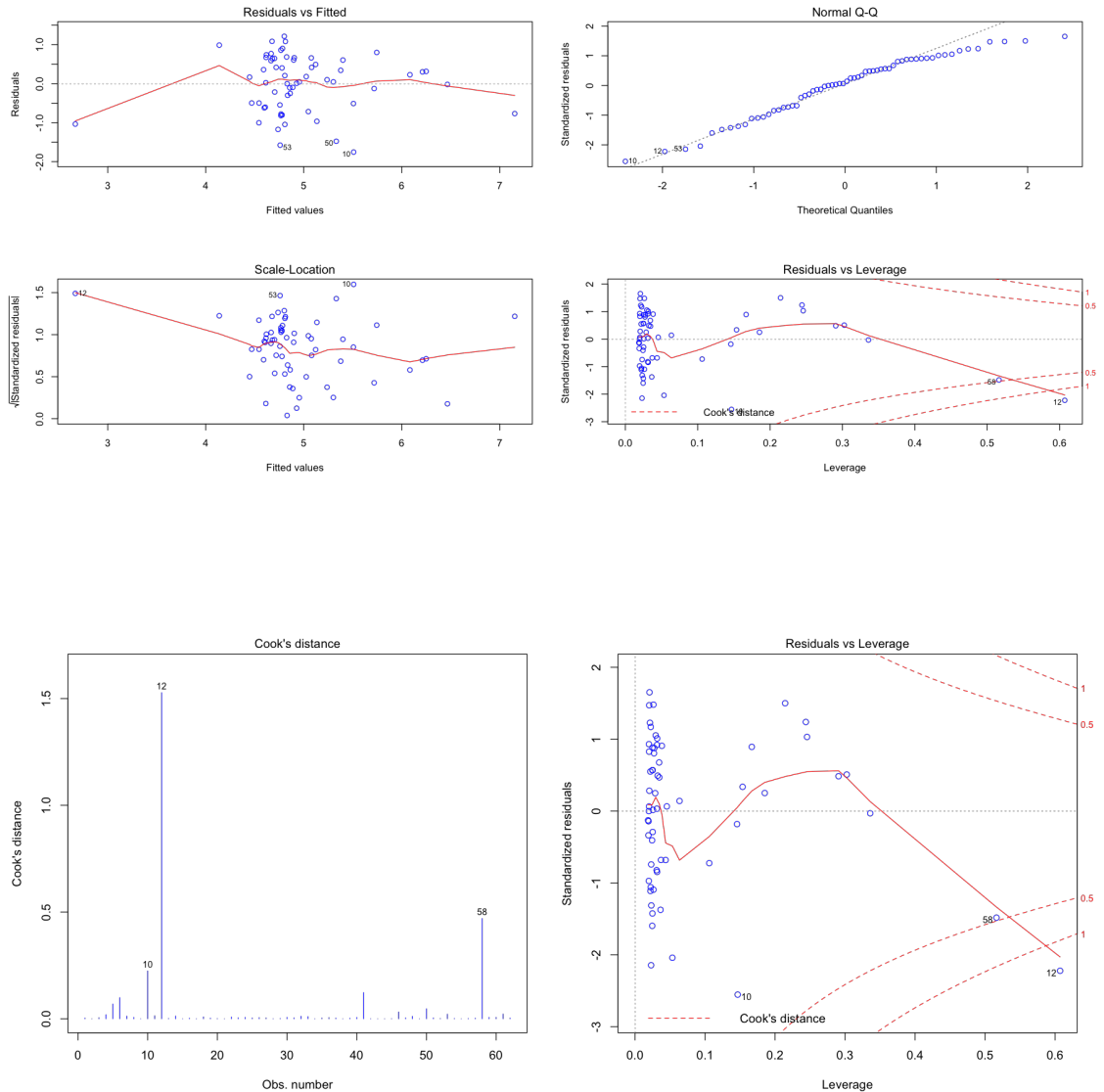
Removing the outliers leads the animated to lose significance in the model (also its interactions), so let's try another fit removing it and the interactions

```
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) 4.3772836  0.1535636  28.505  < 2e-16 ***
budget      0.0092566  0.0021263   4.353 5.52e-05 ***
cmngsoon    0.0018777  0.0007334   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)	4.0698927765	4.684674482
budget	0.0050003890	0.013512714
cmngsoon	0.0004096702	0.003345711

We obtain a model $box = 4.377 + 0.009*budget + 0.002*cmngsoon$. We see that the residuals in this case are goos and doesn't not present a deterministic path. In addition, the mean of the residuals does appear to be almost 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 fact, does not deviate from the theoretical quantiles of a standard normal (except for one of the tails).
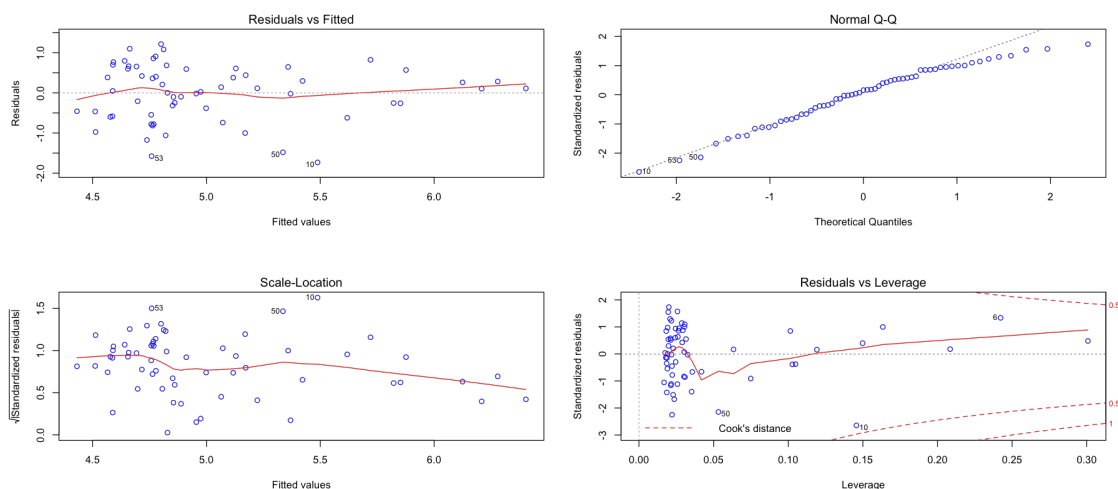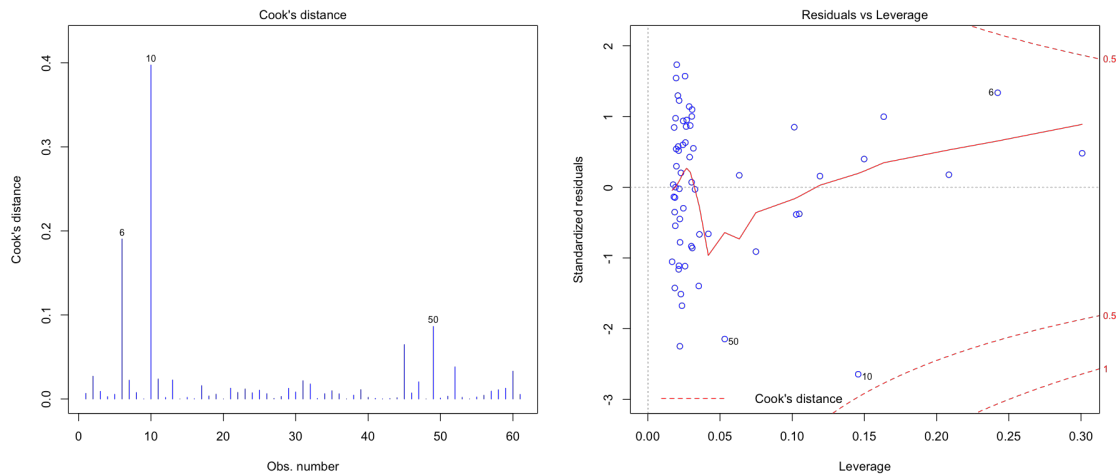
Let's compare the model obtained with the outlier with the one without the outlier using anova (on the same dataset without the outlier). We see the model without animated (due to elimination of outlier value) is preferable. So our final model is : $box = 4.377 + 0.009 * budget + 0.003 * cmngsoon$

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

4.1 Gam

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). 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.

```
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
```

5 Conclusion Point 1

We have that: - boxs depend positevly from budget - boxs depend positevly from cmngsoon

The final model for the subset is: $box = 4.377 + 0.009 * budget + 0.003 * cmngsoon$.

6 Point 2- Data Analysis

6.0.1 RIDGE

Let's consider all the dataset. In the following we plot the informations about covariates and first three rows of the dataset.

```
      box              mprating      budget         starpower        sequel
 Min.   :  5.119   1: 2     Min.   :  5.00  Min.   : 0.00   FALSE:53
 1st Qu.: 69.565   2:15     1st Qu.: 30.50  1st Qu.:12.16   TRUE : 9
 Median :169.309   3:28     Median : 37.40  Median :18.07
 Mean   :207.207   4:17     Mean   : 53.29  Mean   :18.03
 3rd Qu.:266.961            3rd Qu.: 60.00  3rd Qu.:24.09
 Max.   :709.505            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
```

box	mprating	budget	starpower	sequel	action	comedy	animated	horror	addict	cmngsoon	fa...
191.67085	4	28.0	19.83	FALSE	FALSE	TRUE	FALSE	FALSE	7860.5	10	14
631.06589	2	150.0	32.69	TRUE	FALSE	FALSE	TRUE	FALSE	5737.0	59	46
54.01605	4	37.4	15.69	FALSE	FALSE	TRUE	FALSE	FALSE	850.0	24	19

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.
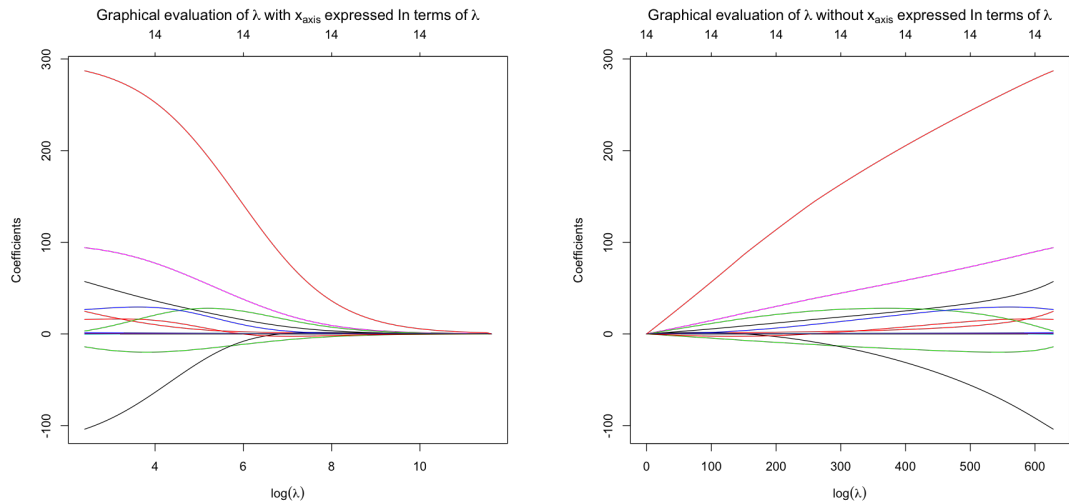
Now let's look for the best λ using cross validation. The plot below shows the values of *cvm* for each $log(\lambda)$ together with the associated confidence interval. The two dashed lines are the values of minimun $log(\lambda)$ and $log(\lambda)$ 1σ far from the minimum.
So the best λ from cross validation is: 71
And the MSE is: 16003
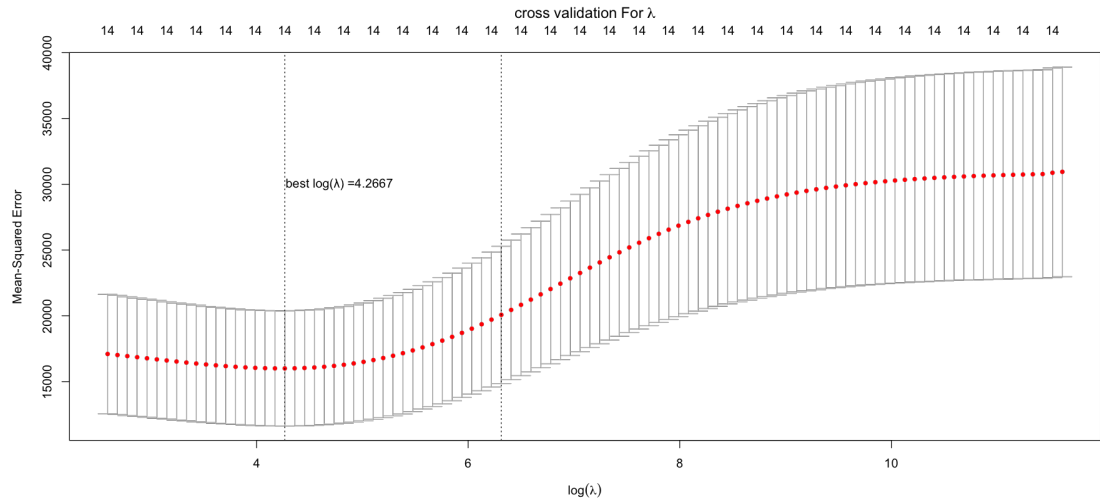
13

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.

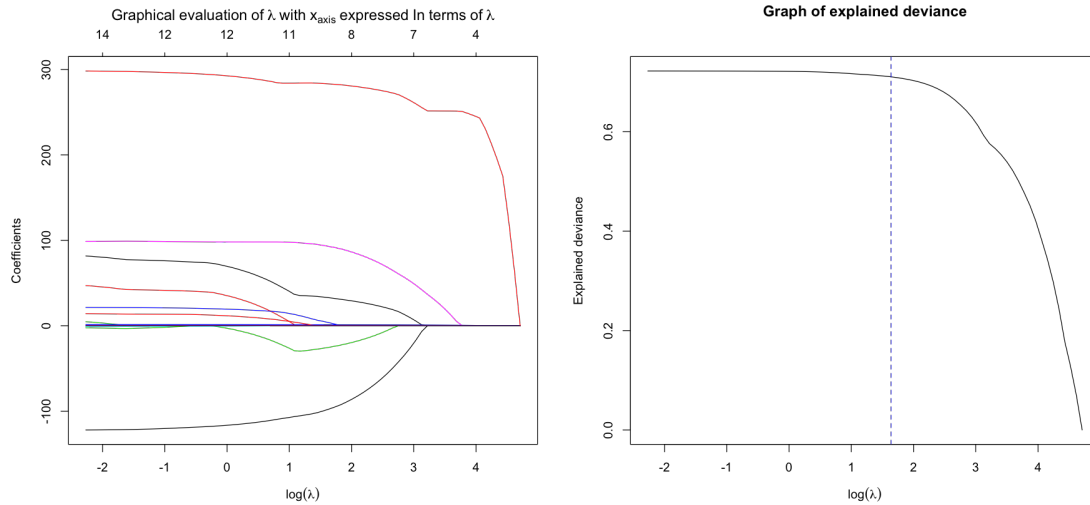The maximum explained deviance is obtained for the minimum (best) λ and it is equal to:5
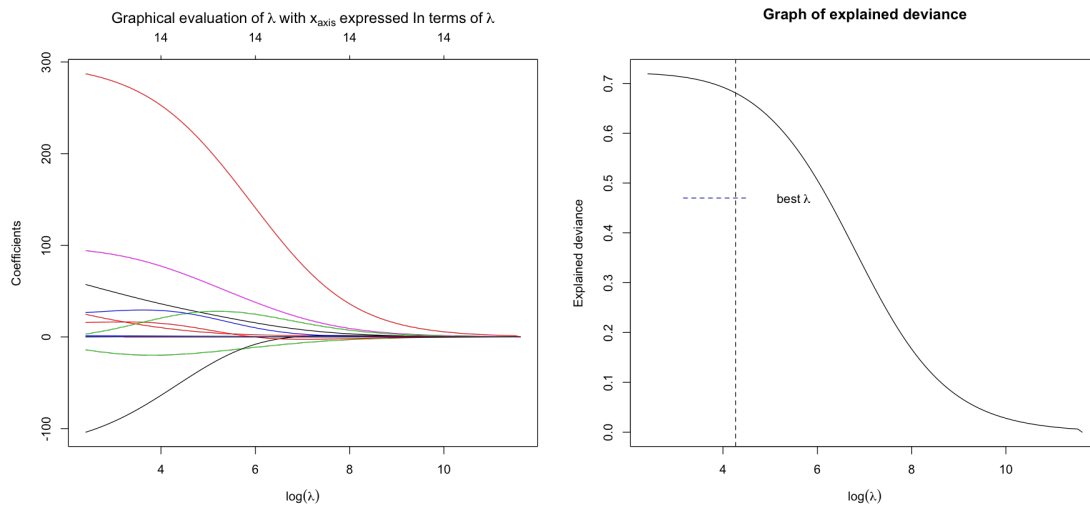
6.0.2 LASSO

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.
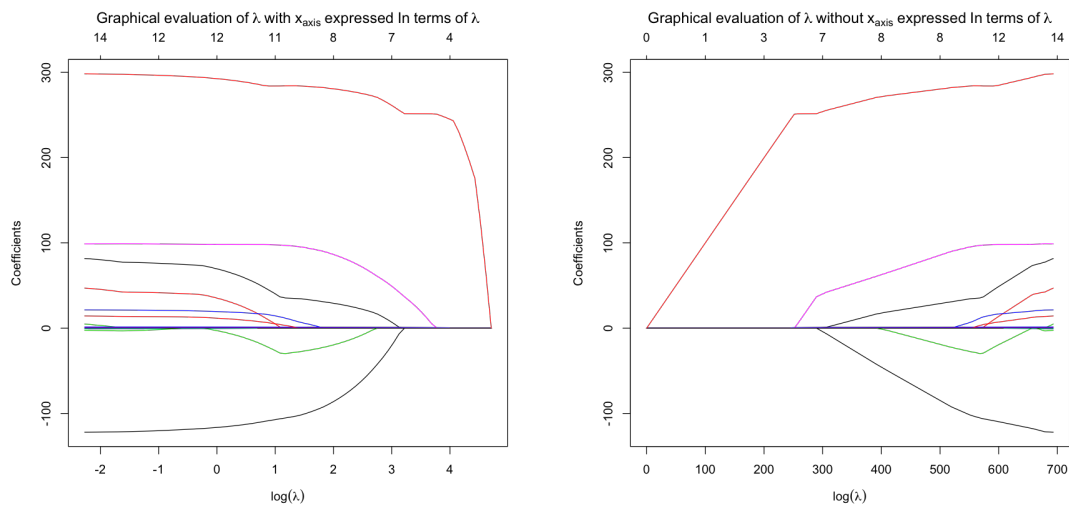
Now let's look for the best λ using cross validation. So the best λ from cross validation is: 5
And the MSE is: 16201
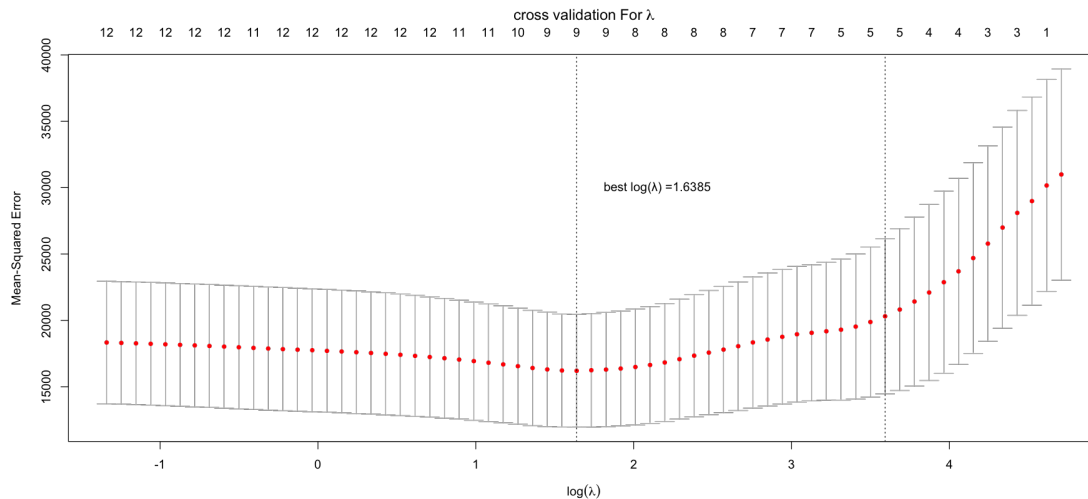
On the basis of MSE , the model fitted with lasso has got the same MSE by the way The resulting model with lasso is simpler. Now we can Re-estimate the model using the best λ. Below we seen the coefficients of the model, graphical representation of the coefficients for the best λ and model deviance.
The maximum explained deviance is obtained for the minimum (best) λ and it is equal to: 0.72
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: comedytrue, animatedtrue , fandango, and starpower. Also mprating3 is set to 0. So the total number of coefficients different from 0 is 10.


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

     Df   %Dev Lambda
[1,]  9 0.7104  5.147

15 x 1 sparse Matrix of class "dgCMatrix"
                    s0
(Intercept)  -46.209355470
mprating2     32.812449674
mprating3       .
mprating4    -25.156457342
budget         1.163192402
starpower       .
sequelTRUE    92.792137770
actionTRUE   -97.469669173
comedyTRUE      .
animatedTRUE    .
horrorTRUE     3.672794196
addict         0.006753868
cmngsoon       0.298359645
fandango        .
cntwait      282.999782870
```

Graphical evaluation of λ with x_axis expressed In terms of λ

Graph of explained deviance

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

- MSE for lasso is:16201
- MSE for linear model is:16332

There is a difference. Since there is substantial variable selection, lasso is more interesting and we keep it.
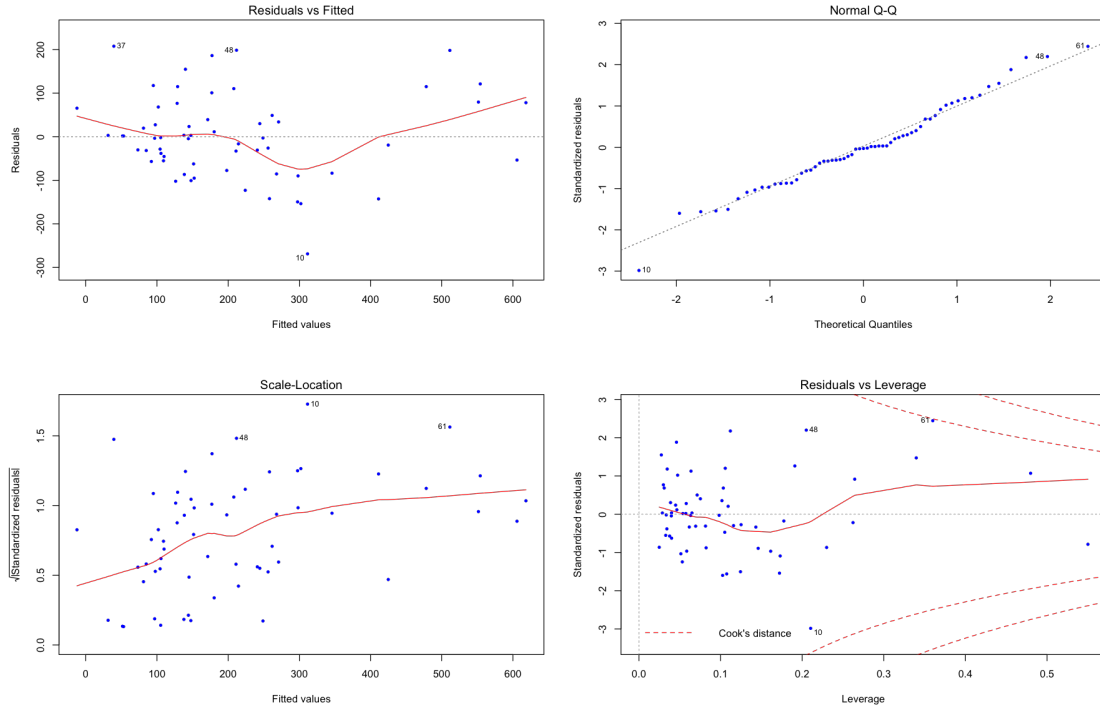Let's also plot the prediction for both methods.

Predictions with Ridge

Predictions with Lasso
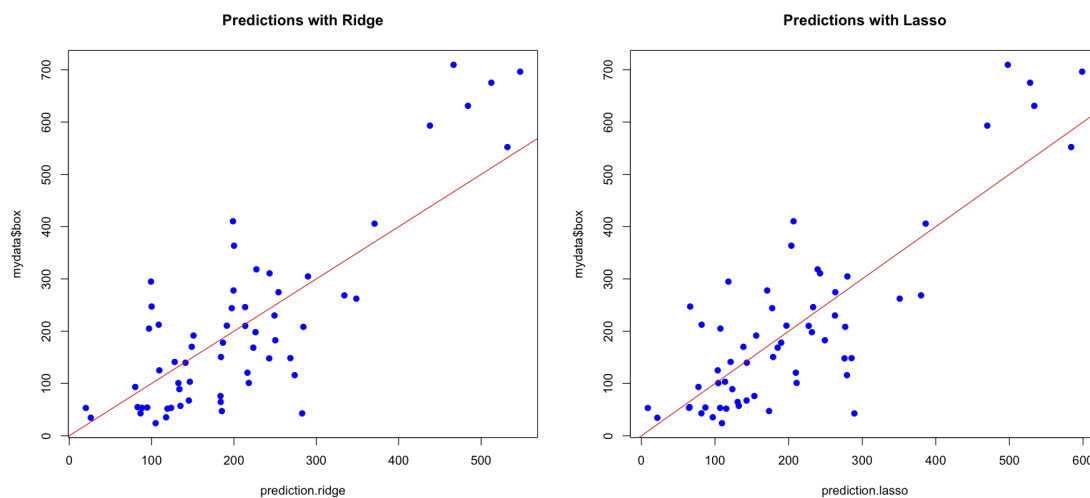
6.0.3 AUTOMATIC SELECTION

FORWARD SELECTION   Let's try automatic selection

From the following plot we see as computed before that the best model basing on BIC is 6 variables.
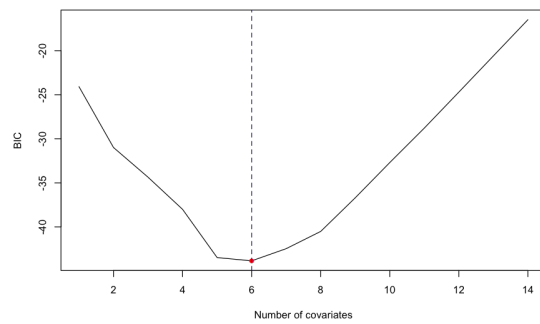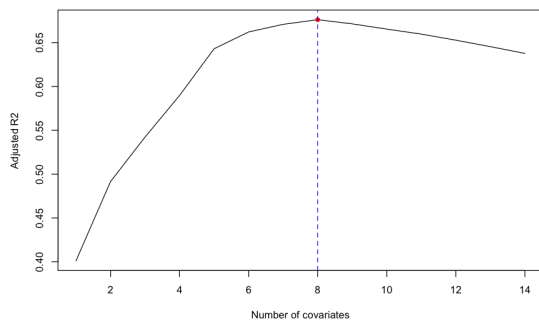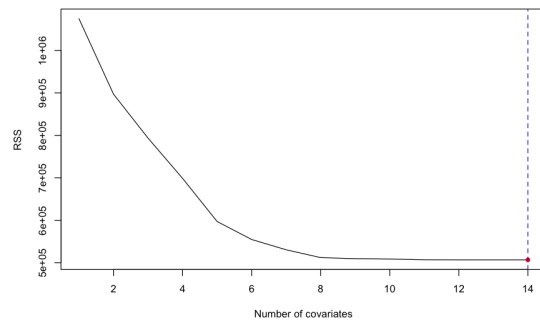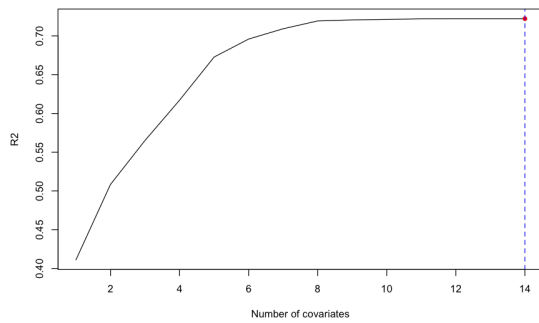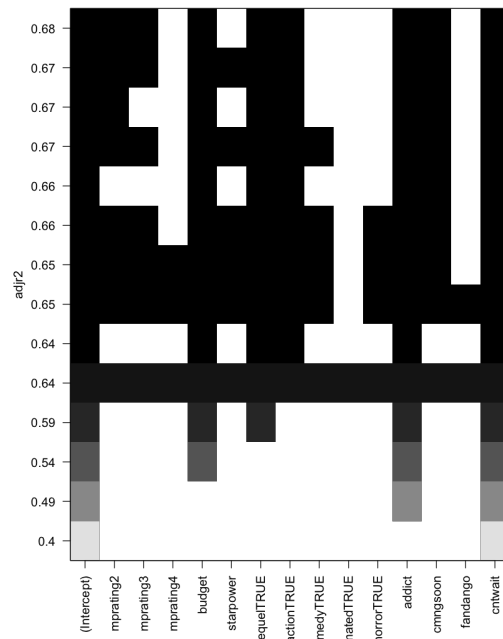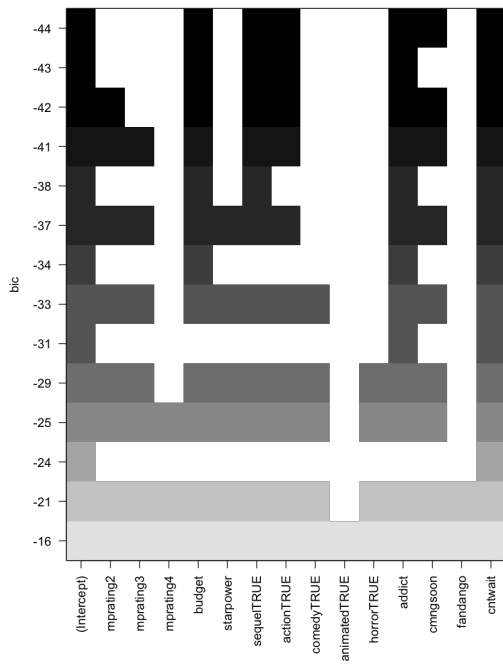
base on BIC we keep the model with the lowest BIC so with a number of variablesequal to : 6.

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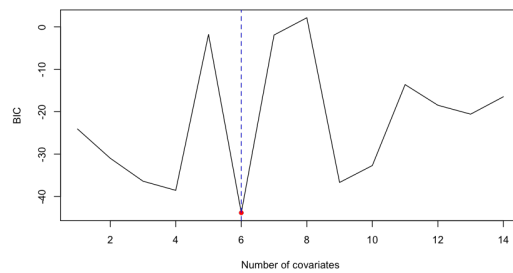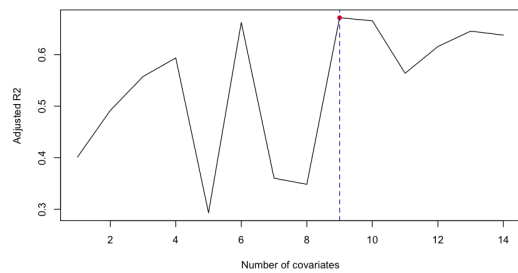
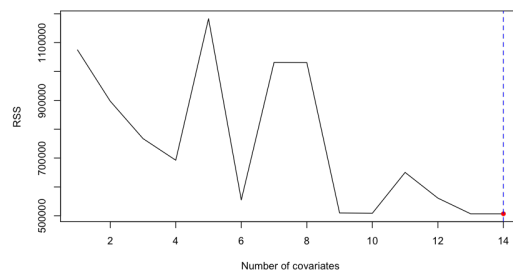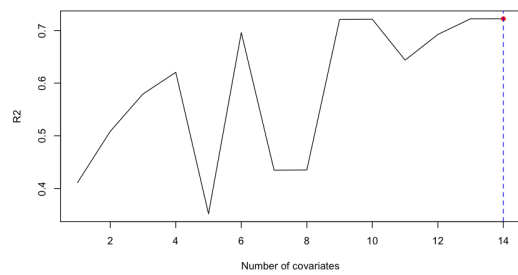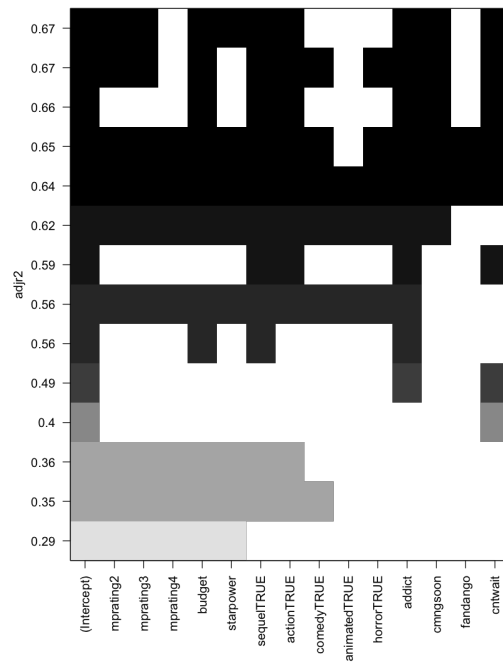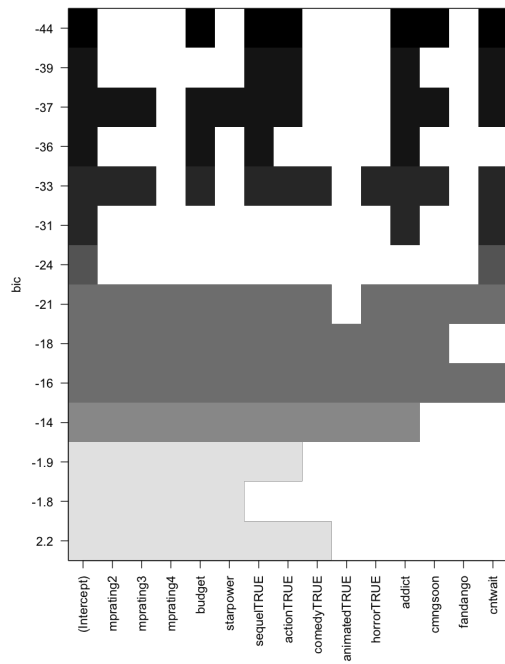### 6.0.4   BACKWORD SELECTION

Let's perform backword selection.
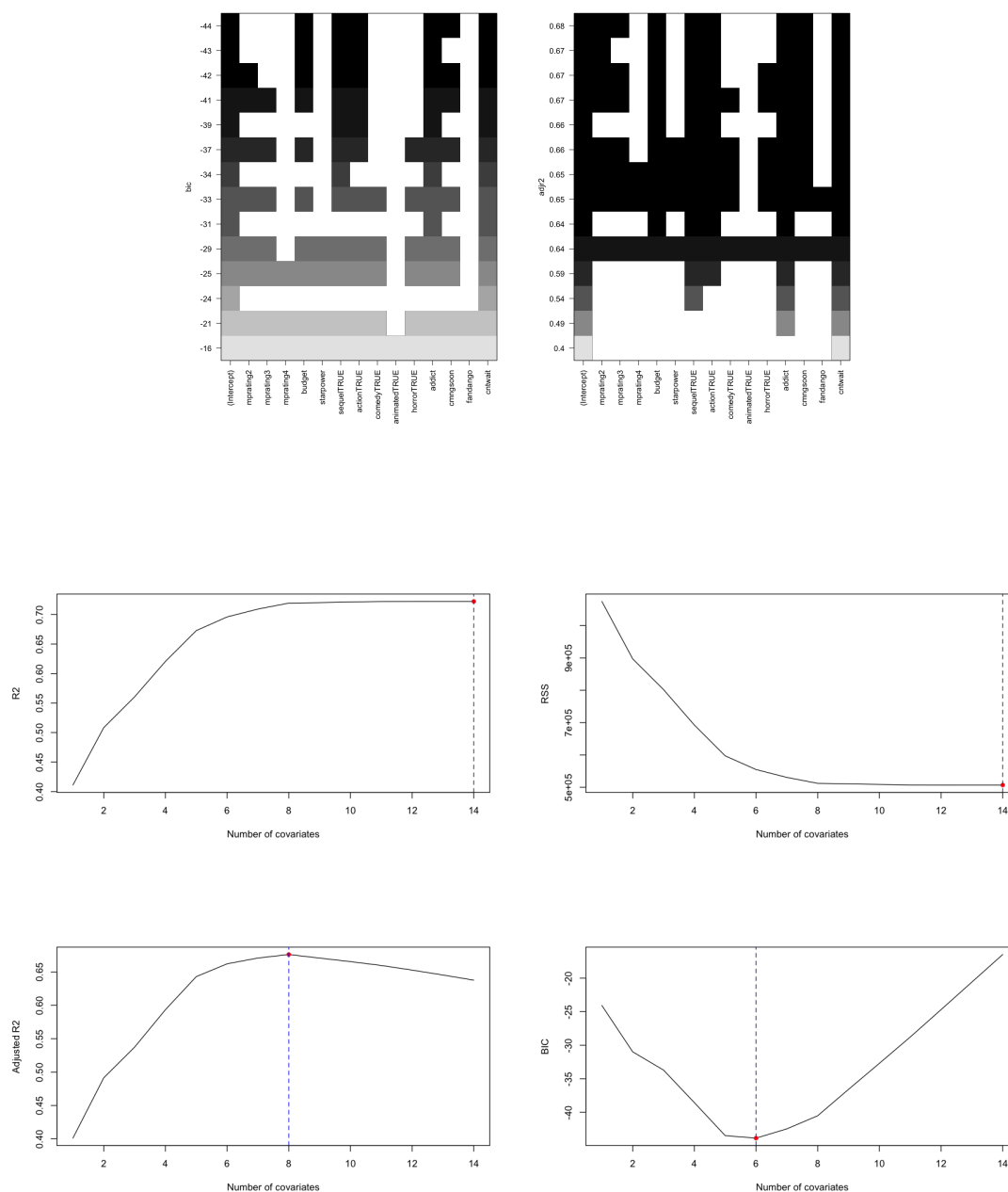
As you can see, forward and backward give us the same amount of covariates based on BIC. In fact in this case we have 6 variables.

6.0.5 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

6.0.6 PRINCIPAL COMPONENT ANALYSIS

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

```
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           175.9    132.3    132.4    130.2    135.2    135.2    124.6
adjCV        175.9    131.8    132.0    129.7    134.7    136.0    123.3
       7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
CV       123.2    120.9    117.3     119.4     122.9     120.8     122.4
adjCV    122.1    119.9    115.1     118.2     121.6     119.2     120.7
       14 comps
CV        125.2
adjCV     123.3

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    46.96    47.89    50.58    51.08    51.21    62.12    65.20    66.20
     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    69.21     69.34     70.66     71.97     72.06     72.22
```

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 : 1 We also have that the value of MSE is reported below.

```
       (Intercept)  1 comps
CV           30924    17507
adjCV        30924    17376
```
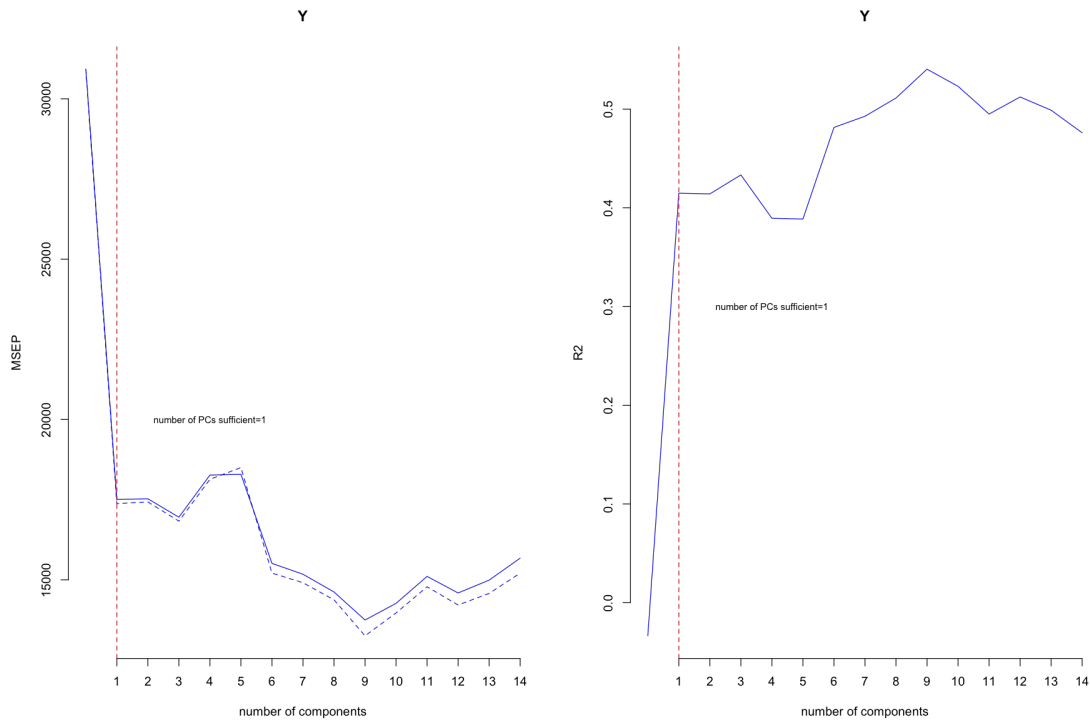
Let's look how much variance is explained by the 14 components in the plot below. While the explained variance for our 1 PCs obtained before is 25% that is a bit low. So we will also consider 10 components that are the same number obtained by lasso which lead to 95% of explained deviance.
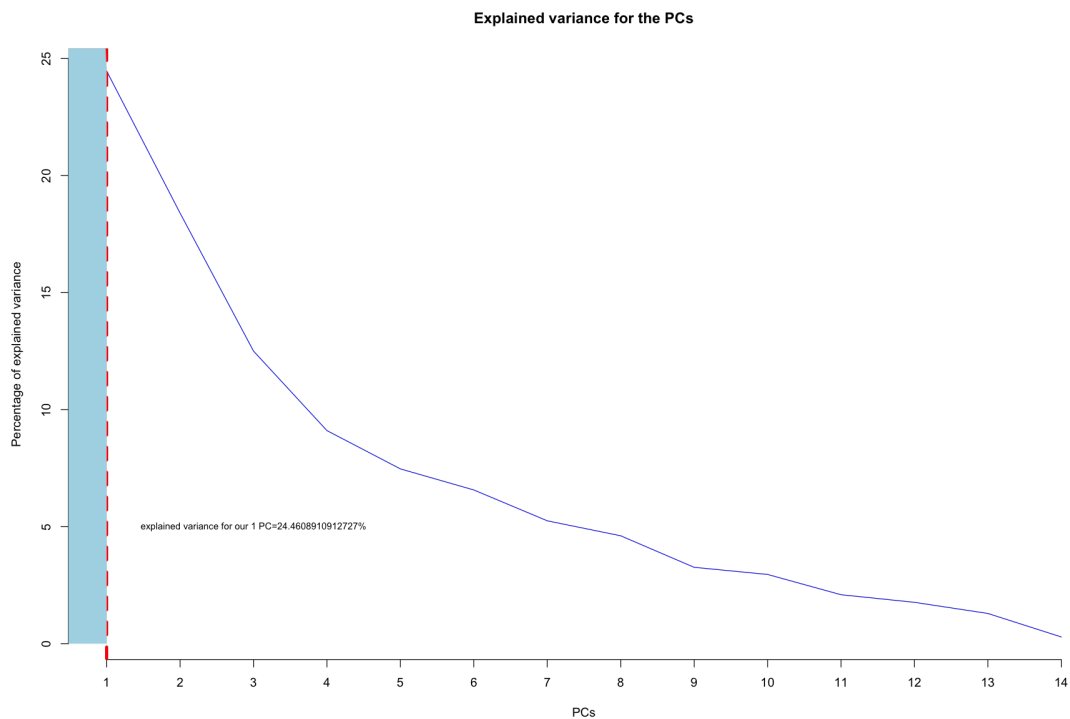
Let's now plot Plot the regression coefficients associated to the models with increasing PCs, from 1 to 10 We see that we have our 10 models. We look for the picks. As picks are higher as our model is better. The model with 10 comps give us the largest amount of informations (higher explained deviance).
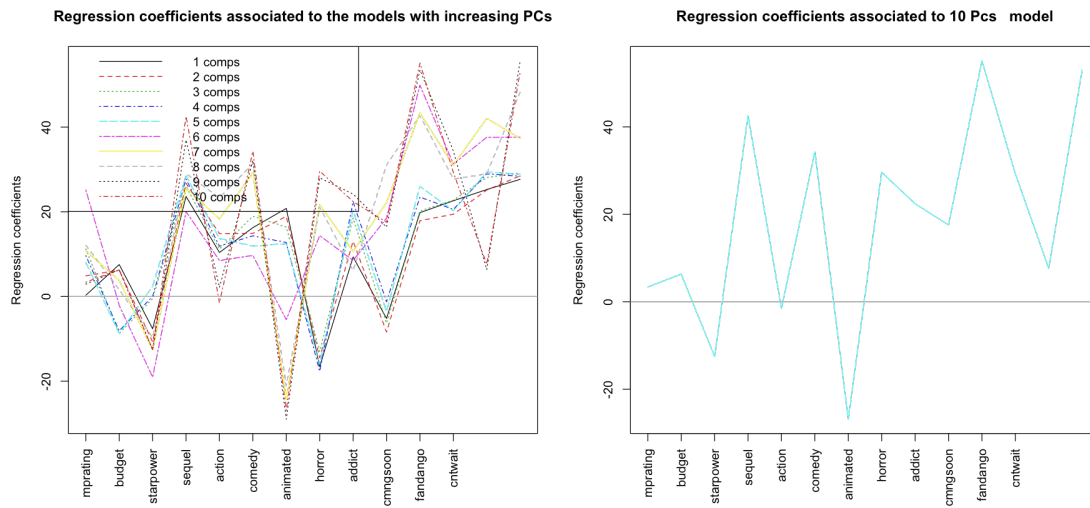
Let's evaluate the presence of of groups of observations or outliers through the scores (for the first 5 PC components but the same consideration are for the others). We do not see a kind of trend in the groups. So PCA is satisfactory.
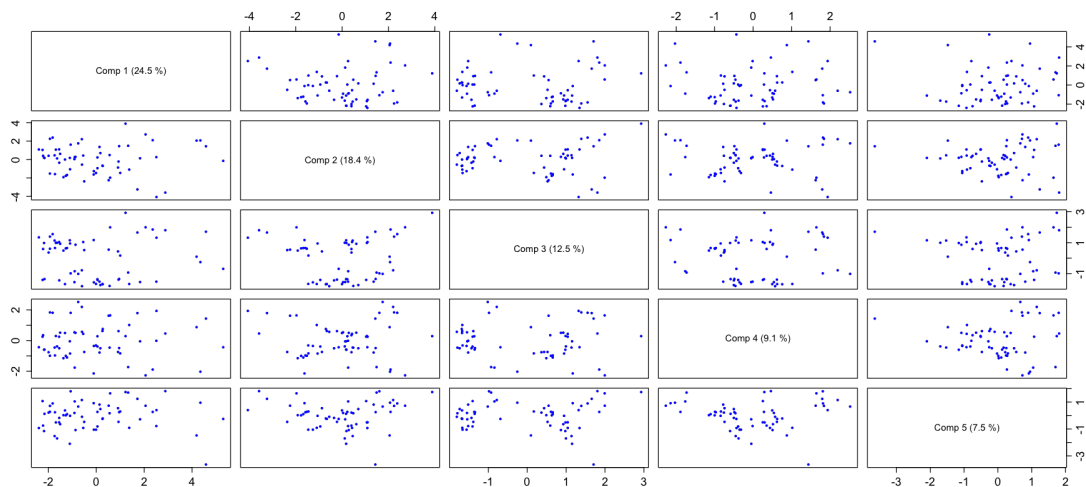
Finally, evaluate the predictions from the model. Values around the bisector does suggest a good behavior of the model.

Predictions PCA
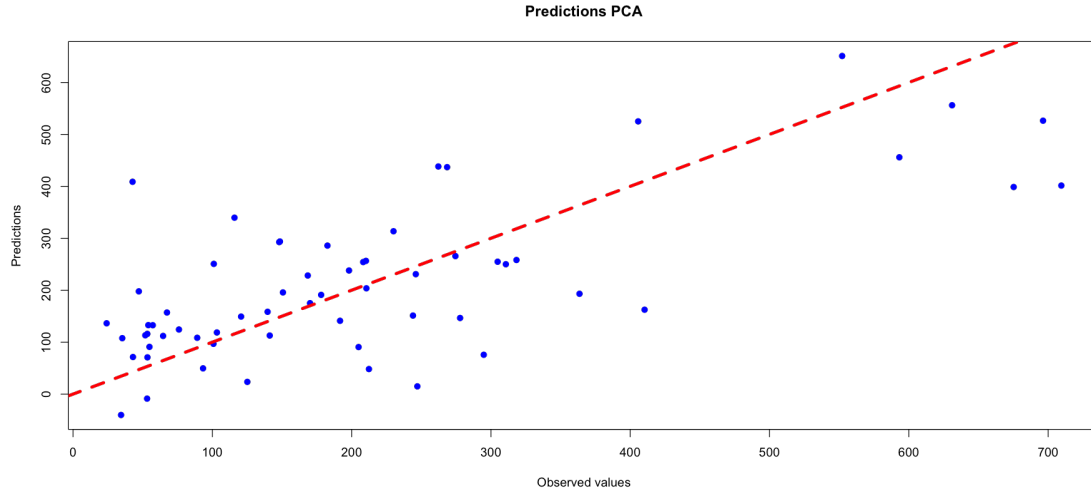
Finally we compute the MSE considering 10 components which is equal to: 9173.

7 Conclusion Point 2

Base on MSE we have:

- MSE for lasso is:16201
- MSE for linear model is:16332
- MSE for ridge is:16203
- MSE for PCA is: 9173

So based on MSE the best approach is with the PCA. In particular we see that the 10 PC depends on sequel, comedy, horror and fandango. By the way PCA usually is used for clustering purpose so we might pay attention about these results. Considering the variable selection lasso usually is a better approach and in this case it gives a model without the following covariates:: comedy, animated , fandango, starpower and mprating3. As concern the automatic selection since the residual of the selection are not good I decided to not consider it.