

Advanced Statistical Modelling: Linear Models

Joel Cantero Priego and Ricard Meyerhofer Parra

12/10/2019

Introduction

In this assignment, we are going to use the IMDB dataset. This IMDB dataset, contains information of 940 films released between 2000 and 2016. The data has been obtained from the IMDB's webpage. The following is a list where we can see all the variables of the dataset:

Variable name	Description	Values
movietitle	Director of the given title	String
gross	Gross in dollars	Integer
budget	Budget in dollars	Integer
duration	Film duration in minutes	Integer
titleyear	The release year of the title	Integer
directorfl	Director Facebook likes	Integer
actor1fl	Actor 1 Facebook likes	Integer
actor2fl	Actor 2 Facebook likes	Integer
actor3fl	Actor 3 Facebook likes	Integer
castfl	Cast Facebook likes	Integer
facenumber_in_poster	Number of faces that appears in the poster	Integer
genre	Genre film	Action/Comedy/Drama/Terror

As we can see we have that all our variables are numerical in exception genre. This dataset is complete which means that it has no missing values. However, this does not imply that there are no outliers.

As required in the assignment, we are going to create a categorical variable: **yearcat** which is the categorical substitution of titleyear with 3 levels: 2000-2005, 2006-2010 and 2011-2016. Therefore, we will have two categorical variables (genre and titleyear).

```
dataset$yearcat<-cut(dataset$titleyear, c(2000,2005,2010,2016),  
                      include.lowest = TRUE,  
                      labels=c("2000-2005", "2006-2010", "2011-2016"))
```

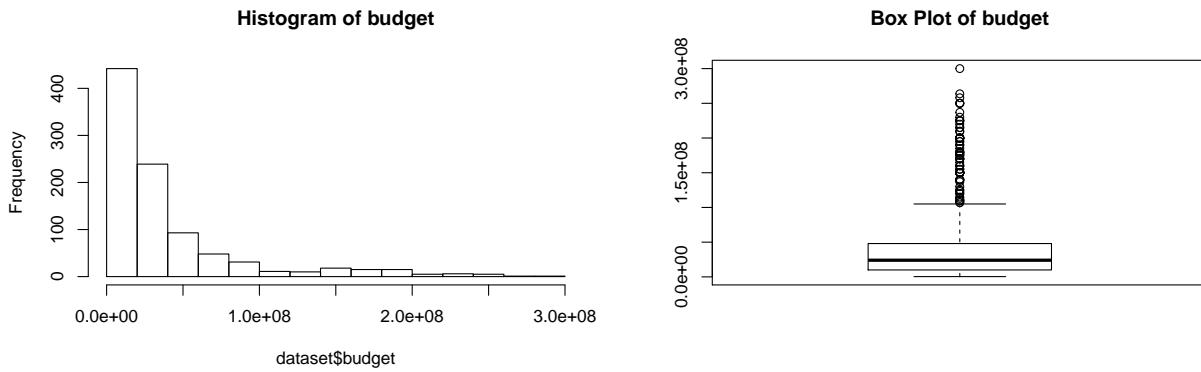
Exploratory Data Analysis

In this section we are going to focus in explaining the most interesting conclusions of our data, perform an univariate and multivariate analysis of the variables in order to find outliers and see the relationship and structure of the dataset variables. We are also going to modify some variables in order to make the linear model perform better on them.

Univariate analysis

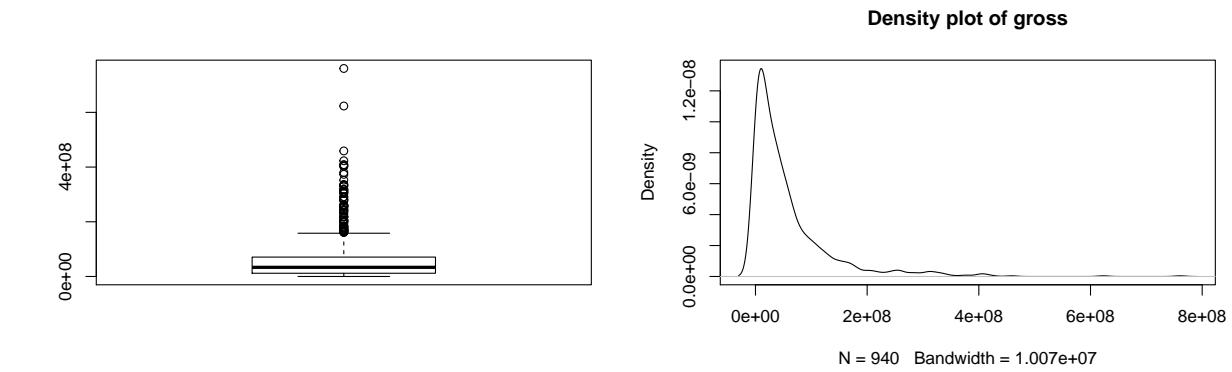
For univariate analysis, histograms and box plot were obtained for the numerical variables and bar charts were generated for the categorical variables.

We can see usual values for the Budget, with a mean of 40,484,550\$.



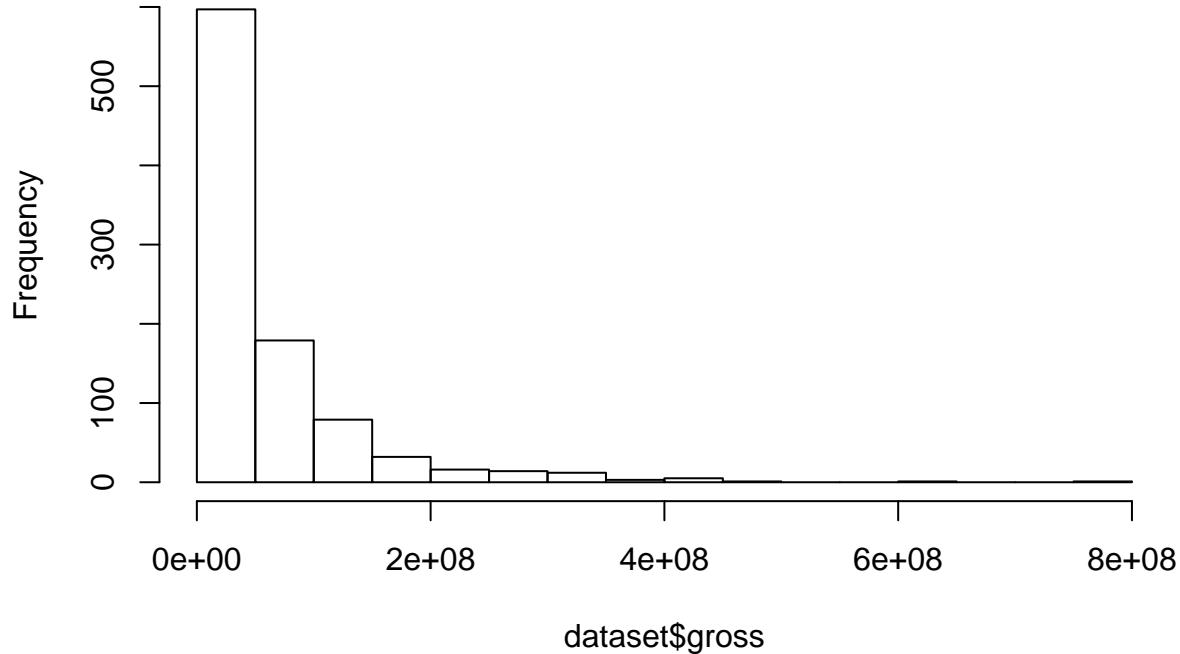
```
# Gross variable exploration
summary(dataset$gross)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 3330 11816543 33428175 57813237 70756664 760505847
par(mfrow=c(1,2))
boxplot(dataset$gross)
plot(density(dataset$gross), main="Density plot of gross")
```



```
hist(dataset$gross)
```

Histogram of dataset\$gross



Sample mean

```
(m=mean(dataset$gross))
```

```
## [1] 57813237
```

Sample Sd

```
cat("Sample Sd\n")
```

```
## Sample Sd
```

```
(s=sd(dataset$gross))
```

```
## [1] 77110518
```

LogLik

```
(loglik=sum(log(dnorm(dataset$gross,mean=m, sd=s))))
```

```
## [1] -18404.41
```

AIC

```
(AIC=-2*loglik+2*2)
```

```
## [1] 36812.81
```

Plot Quantiles-Quantiles (Empirical-Theoretical)

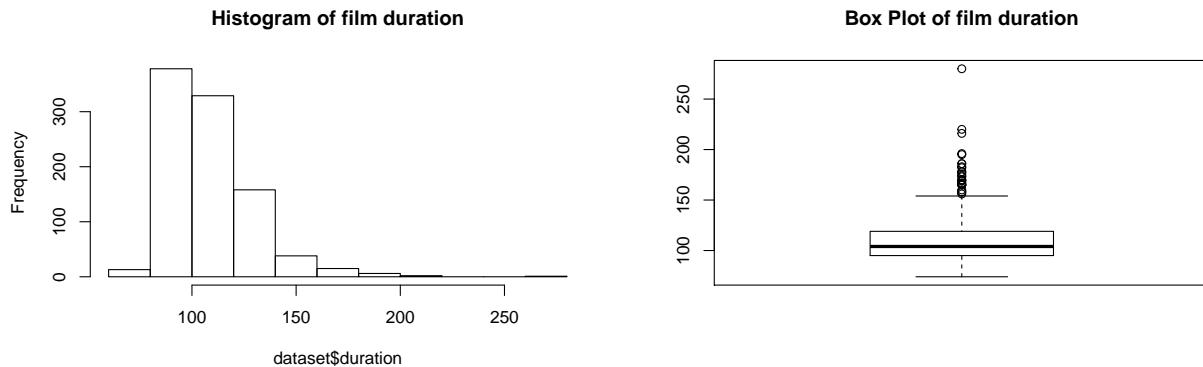
```
qqplot(qnorm(ppoints(500),mean=m,sd=s),dataset$gross) qnorm(dataset$gross) ````
```

In gross, our target variable, we see that there is a certain tendency to normality centred around 60.000.000\$, The highest grossing film is Avatar (760,505,847\$) and the lowest is Mi America (3330\$).

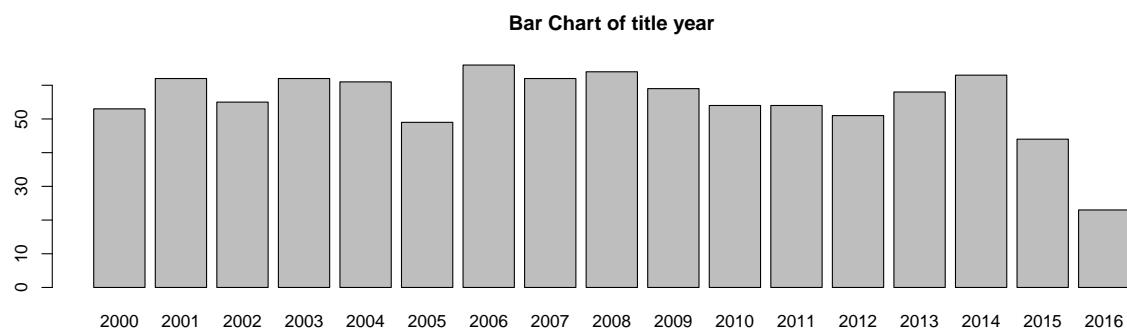
As we have just seen values from budget and gross are in a bigger scale than the rest of our data. This is a problem when performing a linear model since it adds complexity to the model. In order to avoid so, we are going to scale those variables. We decided to apply \log_{10} because it is easier to interpret later when showing (insert justification).

```
#consider later when it increases that we have this variables in logarithm
#log G = bo + b1 log B + .....
#if we add one unit, we have to consider that we are in logarithm and we should express in exponential.
#is not the same if we consider without log and with log, so we have to consider.
#if we multiply per 10 is easier to later show.
dataset$gross <- log10(dataset$gross)
dataset$budget <- log10(dataset$budget)
```

In duration film, we see that there is a certain tendency to normality centred around 100 minutes, we consider it as usual. There is a strange observation of 280 minutes for “Gods and General” film. After check it, we can say that it is not an error but an extrem value.

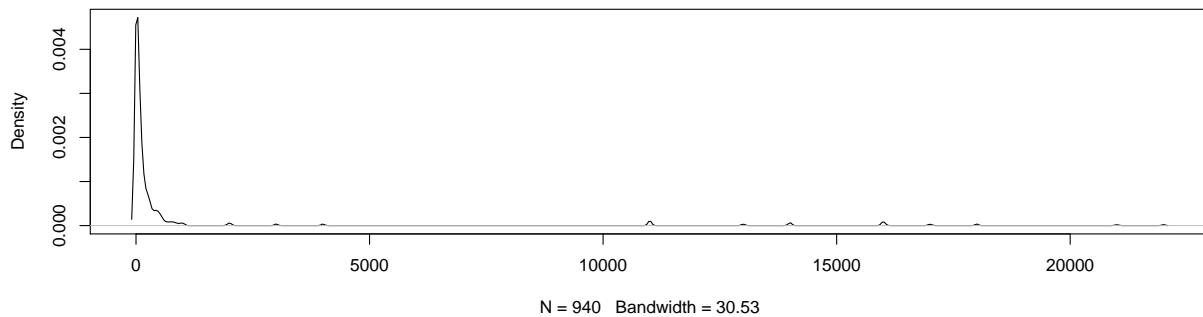


No problems for year, there is a certain expected balanced in years proportion.



In Director Facebook likes, we see that there is a value which appears in the majority of the cases: In this case, is the 0 value. Apart from this zero value, we see that small number of likes are more common than medium or higher number of likes.

Density plot of Director Facebook Likes

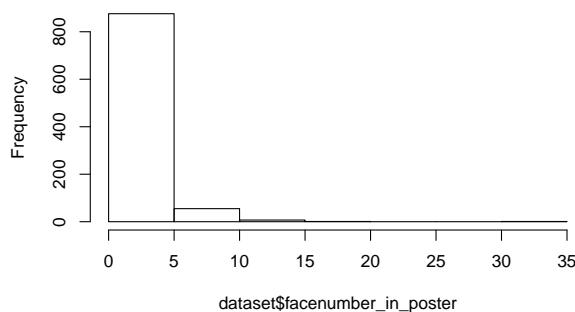


```
## [1] 0.1638298
```

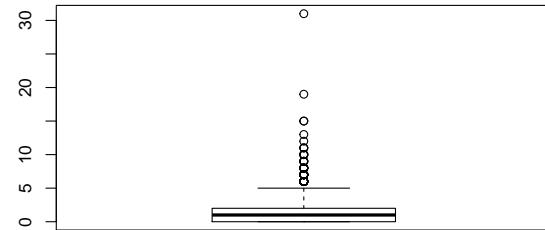
```
``
```

In face number in film poster, the mean is about 1,6 faces and we can observe an extrem value of 31 in "The Master". We can say once again that it is not an error but an extrem value.

Histogram of face number

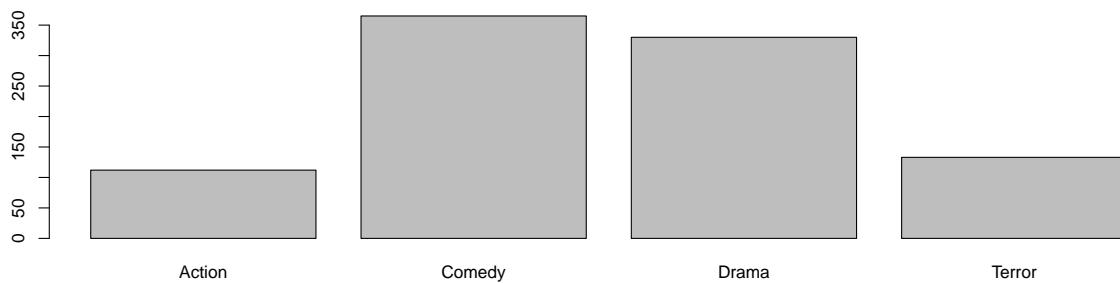


Box Plot of face number

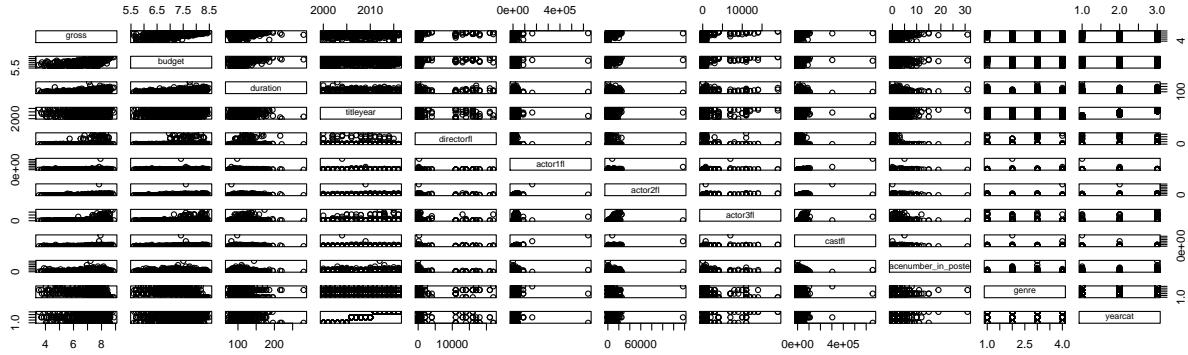


In genre film we can observe that there are more comedy and drama films than action and terror films.

Bar Chart of film genre



Thanks to cross-correlation matrix, we can see that there is a positive correlation between gross and budget variable (0,729). On the other hand there is positive correlation between Cast Facebook Likes and Actor Facebook likes.



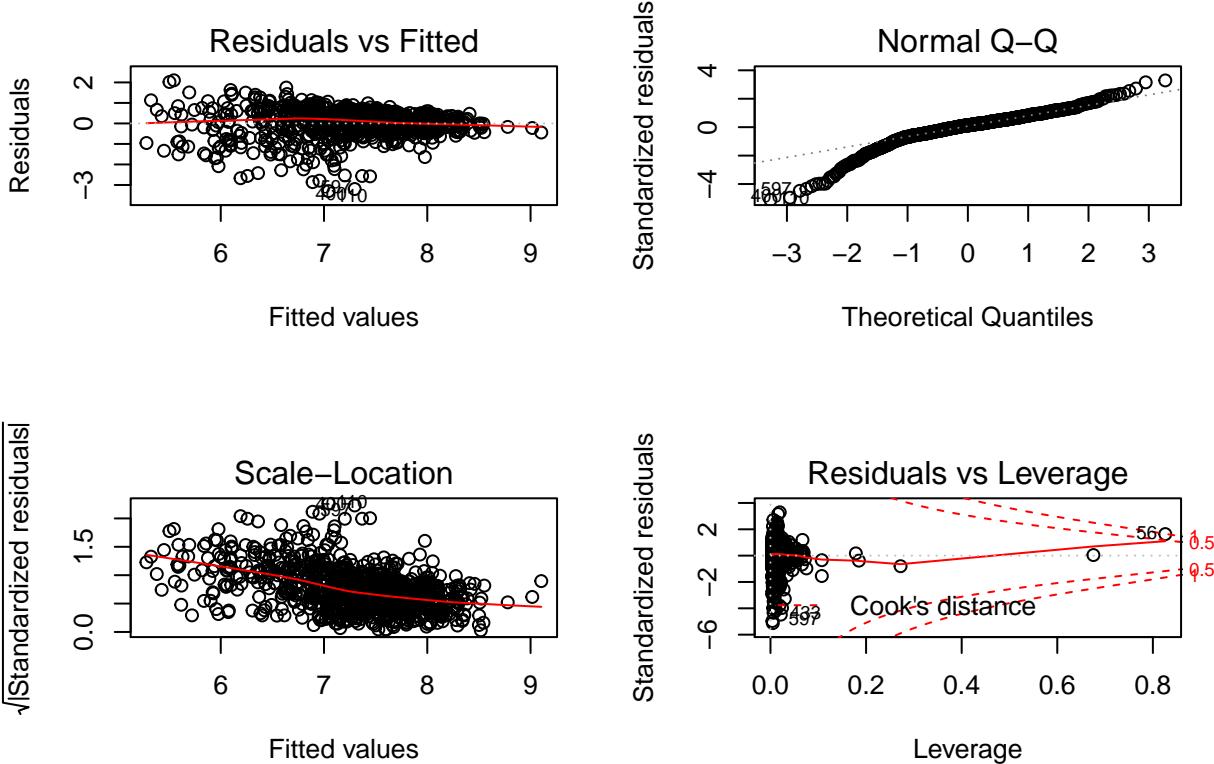
	gross	budget	duration	directorfl	actor1fl	actor2fl	actor3fl	castfl	facenumber_in_poster	yearcat
gross	1.0000000	0.6723773	0.2841011	0.1095329	0.0980622	0.1712539	0.2185811	0.1506855	-0.0045927	
budget	0.6723773	1.0000000	0.4201833	0.1374398	0.1391303	0.2309626	0.2672052	0.2043609	0.0274253	
duration	0.2841011	0.4201833	1.0000000	0.2152645	0.0649505	0.1288276	0.1809332	0.1070346	-0.0123550	
directorfl	0.1095329	0.1374398	0.2152645	1.0000000	0.0660325	0.0940824	0.0453623	0.0825869	-0.0843436	
actor1fl	0.0980622	0.1391303	0.0649505	0.0660325	1.0000000	0.3491797	0.2377791	0.9618000	0.0551944	
actor2fl	0.1712539	0.2309626	0.1288276	0.0940824	0.3491797	1.0000000	0.4591963	0.5725142	0.0258751	
actor3fl	0.2185811	0.2672052	0.1809332	0.0453623	0.2377791	0.4591963	1.0000000	0.4160769	0.0847390	
castfl	0.1506855	0.2043609	0.1070346	0.0825869	0.9618000	0.5725142	0.4160769	1.0000000	0.0650337	
facenumber_in_poster	-0.0045927	0.0274253	-0.0123550	-0.0843436	0.0551944	0.0258751	0.0847390	0.0650337	1.0000000	

Fitting the complete model

```
#Linear Regression
summary(m1<-lm(gross ~ budget + duration + titleyear + directorfl + actor1fl + actor2fl + actor3fl + ca

##
## Call:
## lm(formula = gross ~ budget + duration + titleyear + directorfl +
##     actor1fl + actor2fl + actor3fl + castfl + facenumber_in_poster +
##     genre, data = dataset)
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -3.2988 -0.2709  0.0808  0.3631  2.1022
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.055e+00 9.389e+00 -0.858 0.3912
## budget       1.058e+00 4.823e-02 21.941 < 2e-16 ***
## duration     2.522e-03 1.313e-03  1.920 0.0552 .
## titleyear    3.530e-03 4.666e-03  0.757 0.4495
## directorfl   5.607e-06 7.256e-06  0.773 0.4399
## actor1fl    -2.574e-05 1.867e-05 -1.379 0.1683
## actor2fl    -2.589e-05 1.940e-05 -1.335 0.1823
## actor3fl    -1.606e-05 3.131e-05 -0.513 0.6082
## castfl       2.574e-05 1.871e-05  1.376 0.1693
## facenumber_in_poster -1.183e-02 9.125e-03 -1.296 0.1953
## genreComedy  3.167e-01 8.062e-02  3.928 9.19e-05 ***
## genreDrama   1.264e-01 8.166e-02  1.548 0.1219
## genreTerror  4.803e-01 9.333e-02  5.147 3.24e-07 ***
## ---
```

```
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6456 on 927 degrees of freedom
## Multiple R-squared: 0.4782, Adjusted R-squared: 0.4714
## F-statistic: 70.79 on 12 and 927 DF, p-value: < 2.2e-16
op<-par(mfrow=c(2,2))
plot(m1)
```



```
par(op)

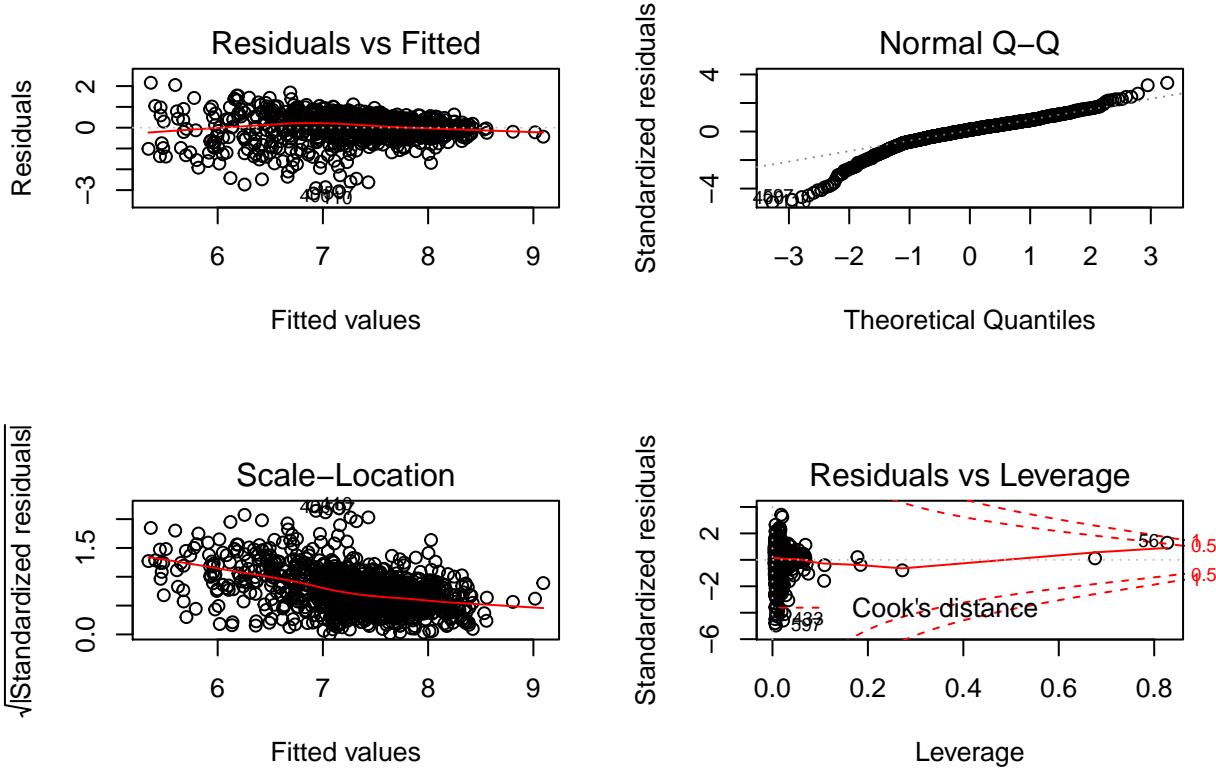
summary(m2<-lm(gross ~ budget + duration + yearcat + directorfl + actor1fl + actor2fl + actor3fl + cast

## 
## Call:
## lm(formula = gross ~ budget + duration + yearcat + directorfl +
##     actor1fl + actor2fl + actor3fl + castfl + facenumber_in_poster +
##     genre, data = dataset)
##
## Residuals:
##      Min    1Q   Median    3Q   Max
## -3.1652 -0.2552  0.0829  0.3710  2.1596
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           -9.347e-01  3.539e-01 -2.641 0.008404 **  
## budget                  1.064e+00  4.780e-02 22.256 < 2e-16 ***
```

```

## duration           2.205e-03  1.303e-03  1.692  0.090900 .
## yearcat2006-2010 -1.765e-01  5.062e-02 -3.487  0.000511 ***
## yearcat2011-2016  4.854e-02  5.245e-02  0.925  0.354990
## directorfl        6.145e-06  7.185e-06  0.855  0.392649
## actor1fl          -2.412e-05  1.848e-05 -1.305  0.192269
## actor2fl          -2.377e-05  1.921e-05 -1.237  0.216255
## actor3fl          -1.612e-05  3.100e-05 -0.520  0.603288
## castfl            2.421e-05  1.852e-05  1.307  0.191532
## facenumber_in_poster -1.485e-02  9.051e-03 -1.641  0.101084
## genreComedy       3.222e-01  7.966e-02  4.044  5.69e-05 ***
## genreDrama         1.391e-01  8.088e-02  1.720  0.085678 .
## genreTerror        4.940e-01  9.244e-02  5.345  1.14e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6392 on 926 degrees of freedom
## Multiple R-squared:  0.4891, Adjusted R-squared:  0.4819
## F-statistic: 68.19 on 13 and 926 DF,  p-value: < 2.2e-16
op<-par(mfrow=c(2,2))
plot(m2)

```



```
par(op)
```

Fitting the complete model

```
#step(m1, scope = list(lower=m1,upper=m2), direction="both", criterion = "BIC", k=log(940))
```