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

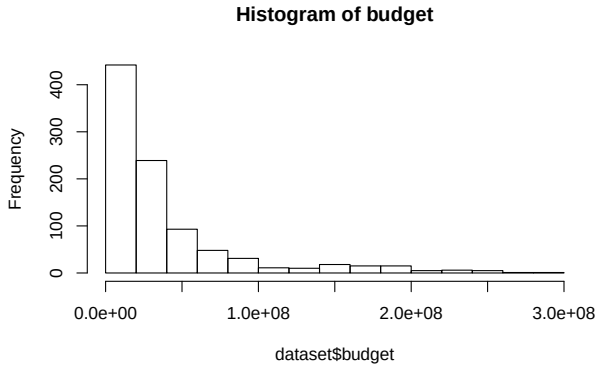
In this section we are going to focus in explaining the most interesting conclusions of our data, perform an univariate and multivariate analysis of the variables in order to find outliers and see how each of these variables is related with the gross. We are also going to modify some variables in order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

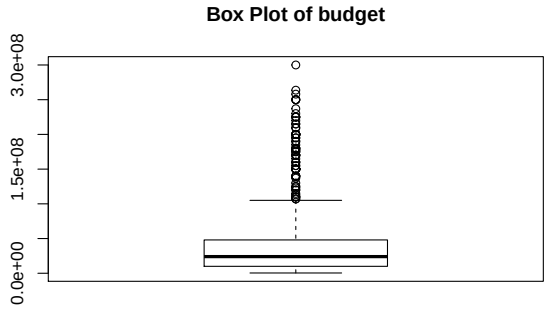
Movie title

We have performed a cloud of the most relevant words that appear in the movies. To do so, we have removed stopwords and punctuation. We could have also done a stemming process but is not so important for us to do so. We can see that the top words are words such as: man, love, movie, house, american, life, big, etc.

Budget

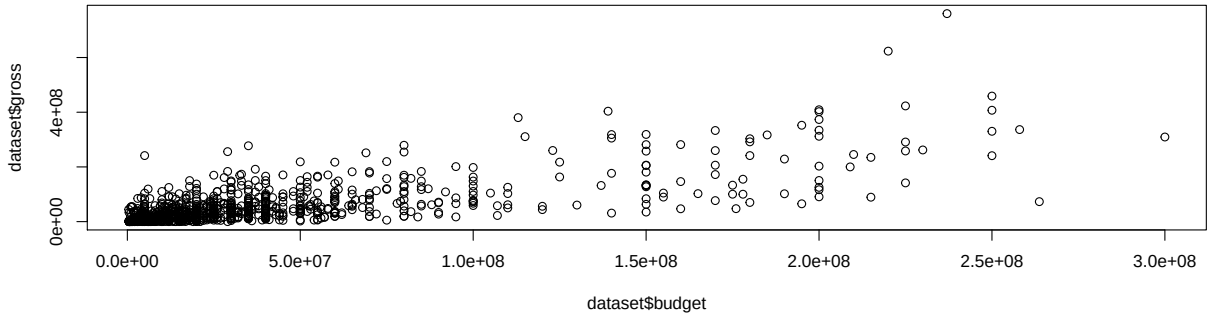
We can see that there is a very disperse amount of values regarding the budget that range from a minimum of 400 thousand dollars (Napoleon Dynamite) to 300 million dollars (Pirates of the Caribbean: At World's End). Despite how crazy this numbers can appear to be, we have revised them by looking at the budget of this two movies on the interent and are correct. Note that this does not imply that all the budgets we have are corrects but it implies that we have to deal with such a range of different values in a same variable.





Relationship between budget and gross

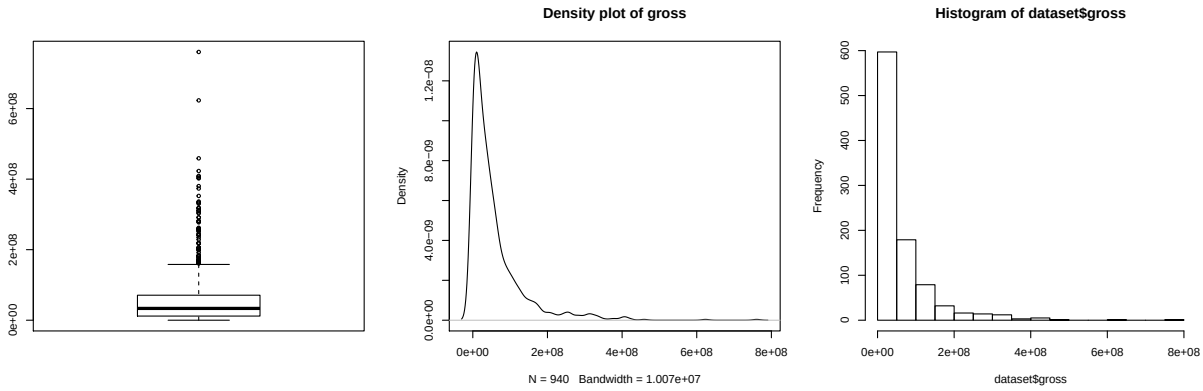
We can see that the revenue is correlated with the budget as we can see in the plot below:



Gross

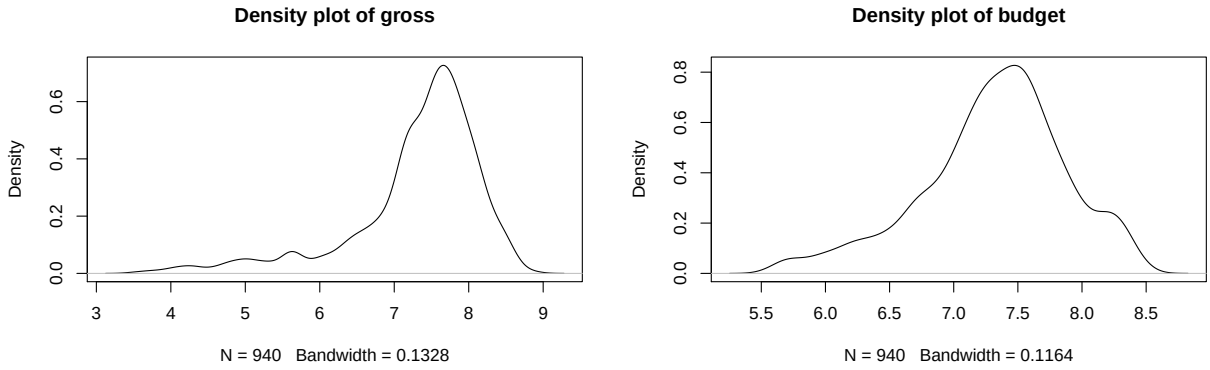
If we take a look at the gross variable, we can see that in a similar fashion than with budget, we have a range of values that can go from 3330$ with Mi America to Avatar with 760 milions of dollars.

```
par(mfrow=c(1,3))
boxplot(dataset$gross)
plot(density(dataset$gross), main="Density plot of gross")
hist(dataset$gross)
```



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 interprate later when showing (insert justification).

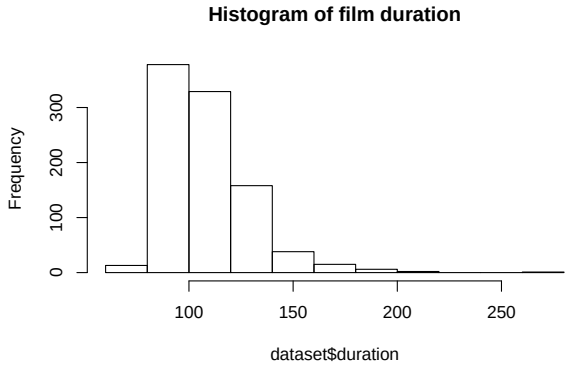
```
dataset$gross <- log10(dataset$gross)
dataset$budget <- log10(dataset$budget)
par(mfrow=c(1,2))
plot(density(dataset$gross), main="Density plot of gross")
plot(density(dataset$budget), main="Density plot of budget")
```

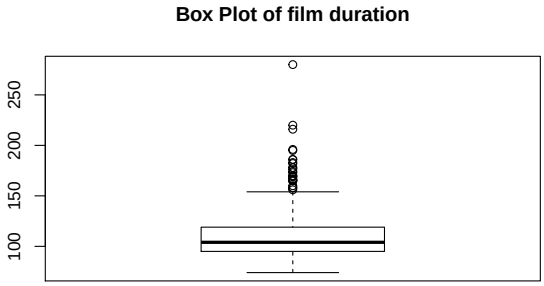


Now we can see that even it does not follow a normal distribution completely it starts to look like one and what is more important is that the range of values is smaller for both, budget and gross.

Duration

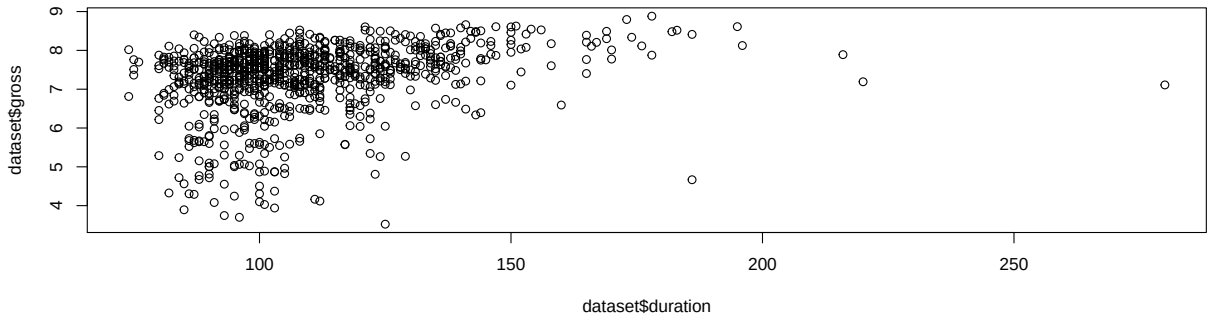
In duration film, we see that there is a certain tendency to normality centered around 100 minutes, we consider it as usual. There is a strange observation of 280 minutes for "Gods and General" film. After we check it, we can say that it is not an error but an extreme value.





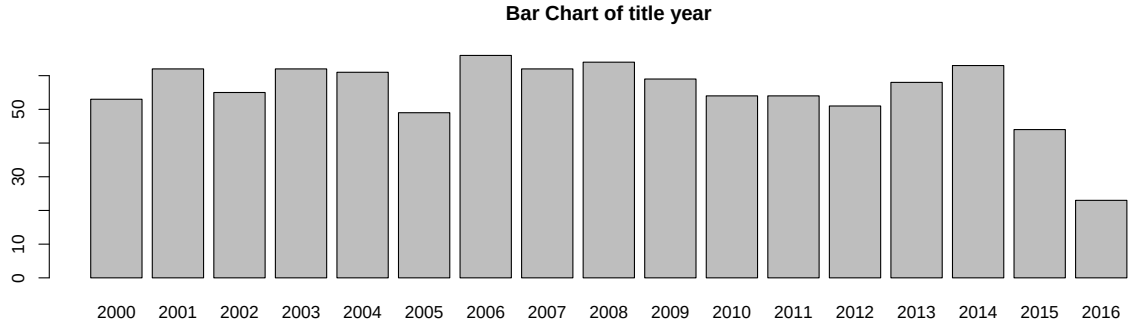
Relationship between duration and gross

We can see that there is not a clear correlation between the duration of a film and it's revenue since we can see that transversally to the duration, we have the similar results in gross.

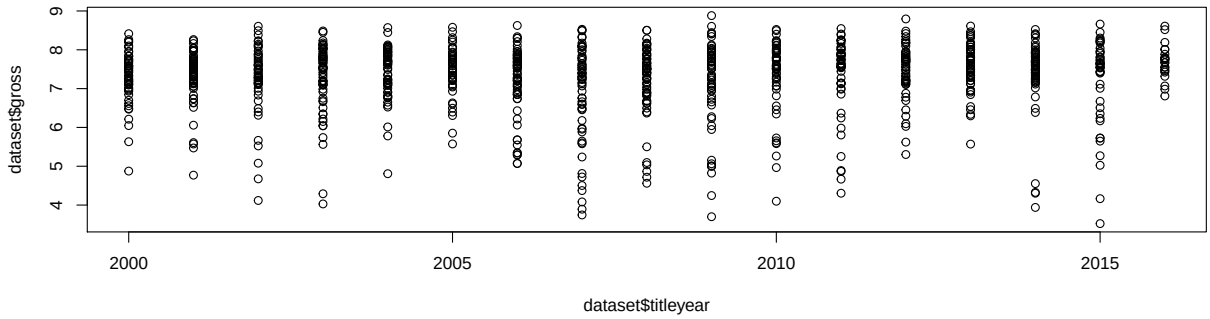


Title Year

No problems for year, there is a certain expected balanced in years proportion even that we can see a decay in the number of films for 2016.

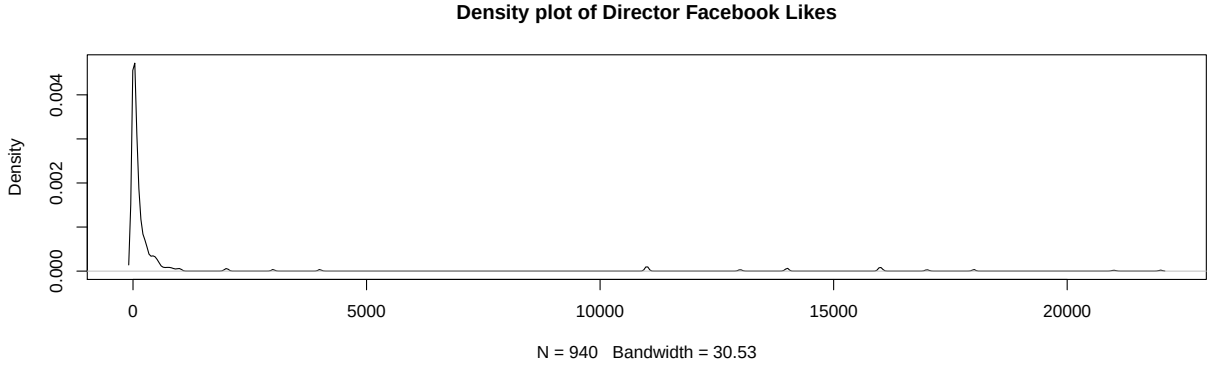


Relationship between years and gross



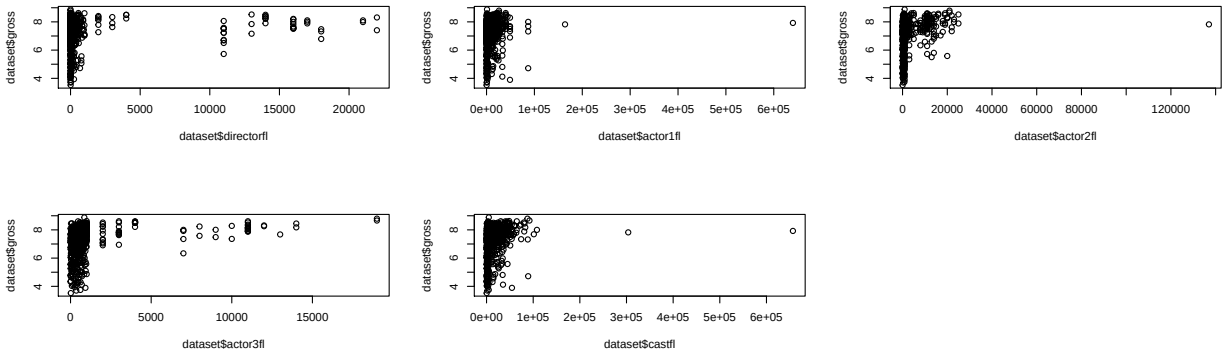
Directorfl, Actor1fl, Actor2fl, Actor3fl, Castfl

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.



We can see that the other variables Actor1fl, Actor2fl, Actor3fl, Castfl follow a similar fashion.

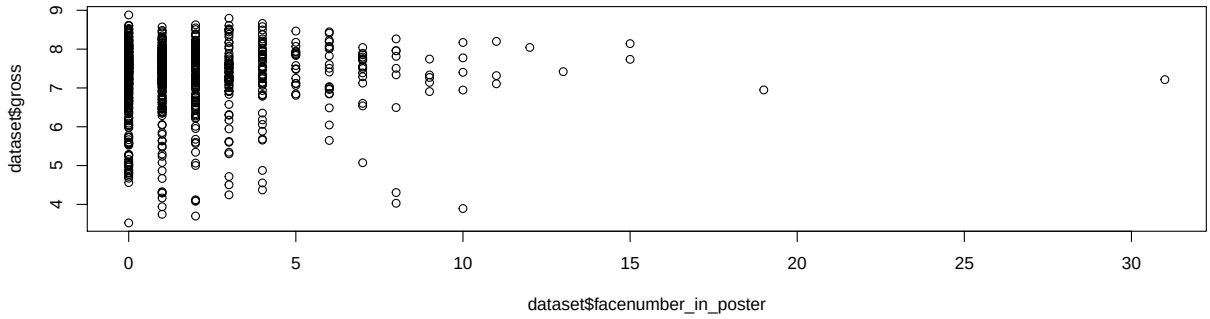
Relationship of Directorfl, Actor1fl, Actor2fl, Actor3fl, Castfl with gross

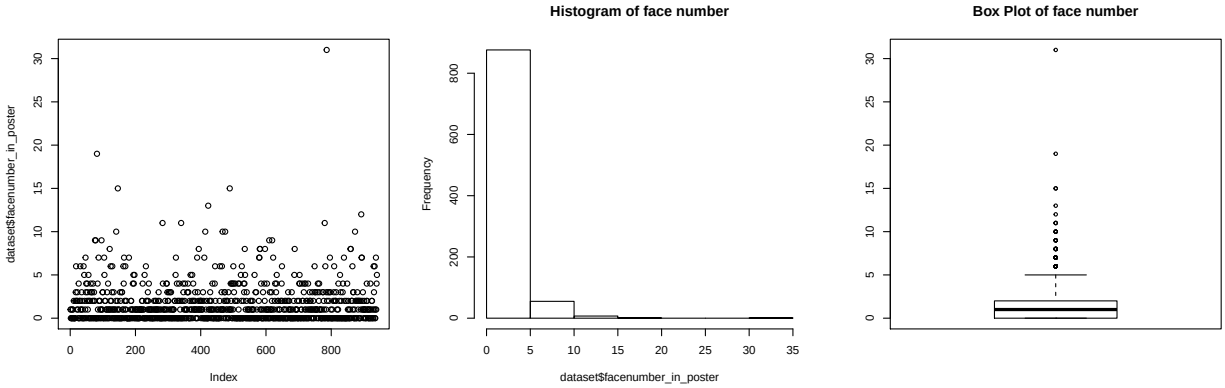


Facenumber in poster

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 extreme value.

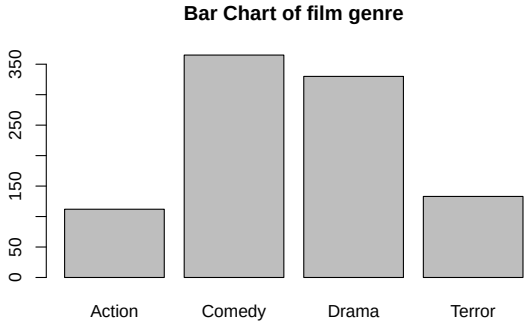
Relationship of Facenumber with Gross





Genre

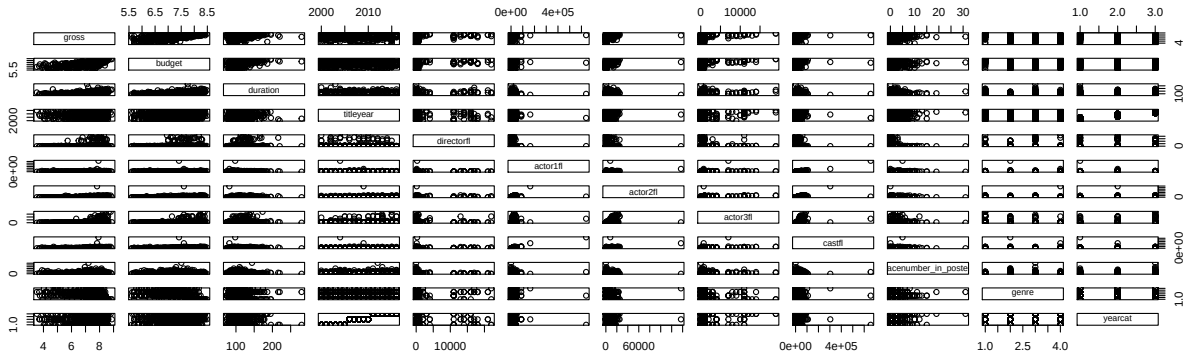
In genre film we can observe that there are more comedy and drama films than action and terror films.



Relationship of Genre with Gross

Correlation matrix

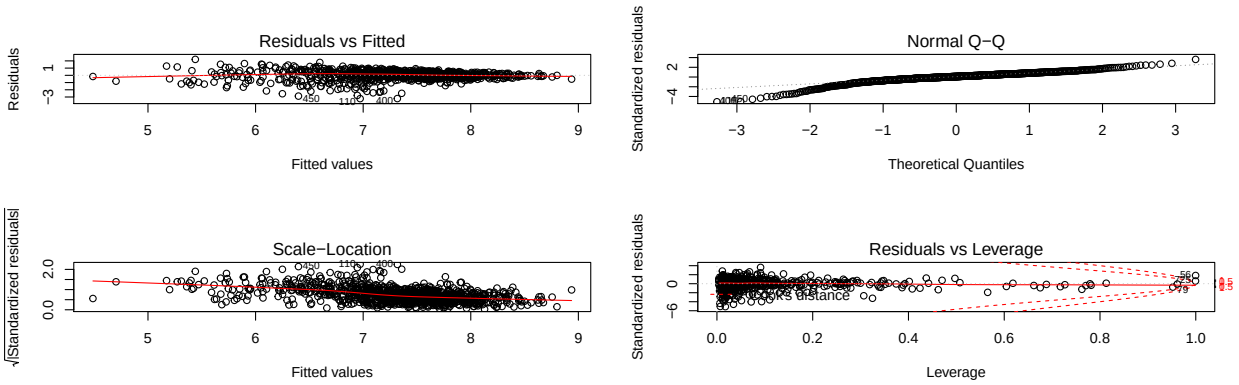
Thanks to cross-correlation matrix, we can see that there is a positive correlation between gross and budget variable (0,729). On the other hand there is positive correlation between Cast Facebook Likes and Actor Facebook likes.



	gross	budget	duration	directorfl	actor1fl	actor2fl	actor3fl	castfl	facenumber_in_poster
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

```
op<-par(mfrow=c(2,2))
plot(completeModel)
```



Use the stepwise procedure, by using the BIC criterion, to select the signicant variables

```
nullModel <- lm(gross ~ 1, dataset)
step(nullModel, scope = list(lower=nullModel,upper=completeModel), direction="both", criterion = "BIC",
```

```
## Start:  AIC=-217.44
## gross ~ 1
##
##                        Df Sum of Sq    RSS     AIC
## + budget                1    334.76 405.71 -776.14
```

```
## + duration              1     59.77 680.70 -289.70
## + genre                 3     60.74 679.73 -277.36
## + actor3fl              1     35.38 705.09 -256.61
## + actor2fl              1     21.72 718.75 -238.57
## + castfl                1     16.81 723.65 -232.18
## + directorfl            1      8.88 731.58 -221.94
## + actor1fl              1      7.12 733.34 -219.68
## <none>                             740.47 -217.44
## + facenumber_in_poster  1      0.02 740.45 -210.61
##
## Step:  AIC=-776.14
## gross ~ budget
##
##                         Df Sum of Sq    RSS     AIC
## + genre                  3     13.43 392.28 -787.25
## <none>                             405.71 -776.14
## + actor3fl               1      1.21 404.50 -772.10
## + facenumber_in_poster   1      0.39 405.31 -770.21
## + directorfl             1      0.22 405.49 -769.81
## + actor2fl               1      0.20 405.51 -769.76
## + castfl                 1      0.14 405.57 -769.61
## + actor1fl               1      0.02 405.69 -769.33
## + duration               1      0.00 405.71 -769.30
## - budget                 1    334.76 740.47 -217.44
##
## Step:  AIC=-787.25
## gross ~ budget + genre
##
##                         Df Sum of Sq    RSS     AIC
## <none>                             392.28 -787.25
## + duration               1     2.190 390.09 -785.66
## + actor3fl               1     2.070 390.21 -785.37
## + directorfl             1     0.538 391.74 -781.69
## + budget:genre           3     6.189 386.09 -781.66
## + actor2fl               1     0.403 391.87 -781.37
## + facenumber_in_poster   1     0.342 391.94 -781.22
## + castfl                 1     0.268 392.01 -781.04
## + actor1fl               1     0.045 392.23 -780.51
## - genre                  3    13.430 405.71 -776.14
## - budget                 1   287.447 679.73 -277.36
##
## Call:
## lm(formula = gross ~ budget + genre, data = dataset)
##
## Coefficients:
## (Intercept)       budget  genreComedy   genreDrama  genreTerror
##     -1.0951       1.1221       0.2388       0.1149       0.4323
```