

# 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.

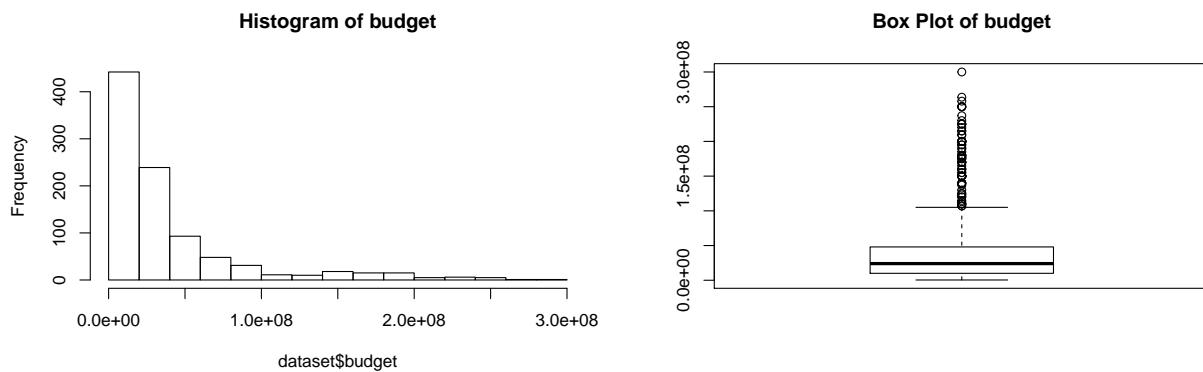
### Movie title

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.



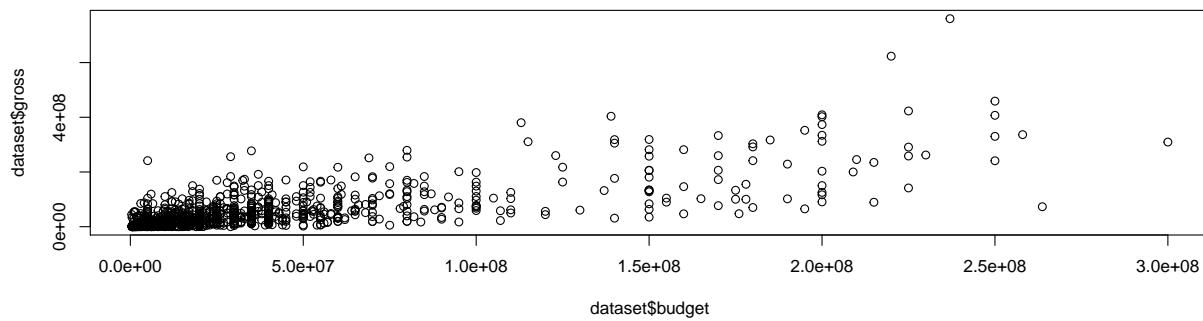
## Budget

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 these numbers can appear to be, we have revised them by looking at the budget of these two movies on the internet and are correct. Note that this does not imply that all the budgets we have are correct but it implies that we have to deal with such a range of different values in a same variable.



### Relationship between budget and gross

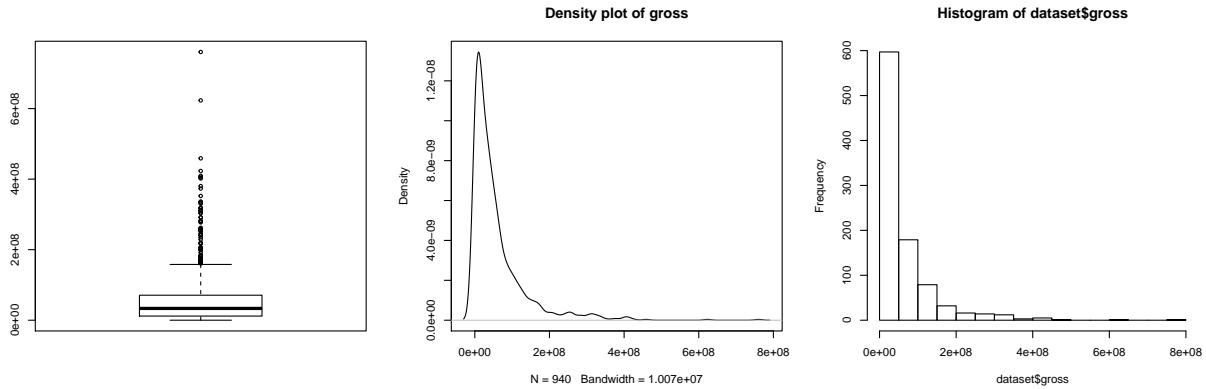
We can see that the revenue is correlated with the budget as we can see in the plot below.



## 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 millions of dollars.

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

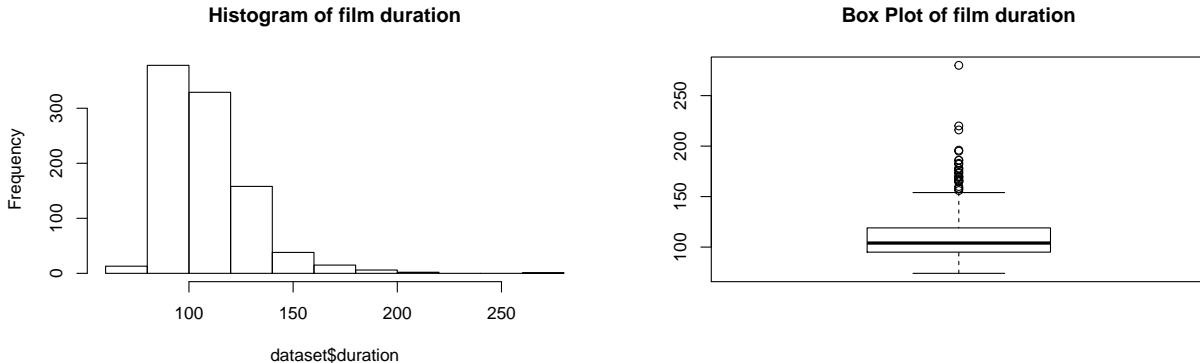


As we have just seen values from budget and gross are in a bigger scale than the rest of our data. This is a problem when performing a linear model since it adds complexity to the model. In order to avoid so, we are going to scale those variables. We decided to apply  $\log_{10}$  because it is easier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

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.

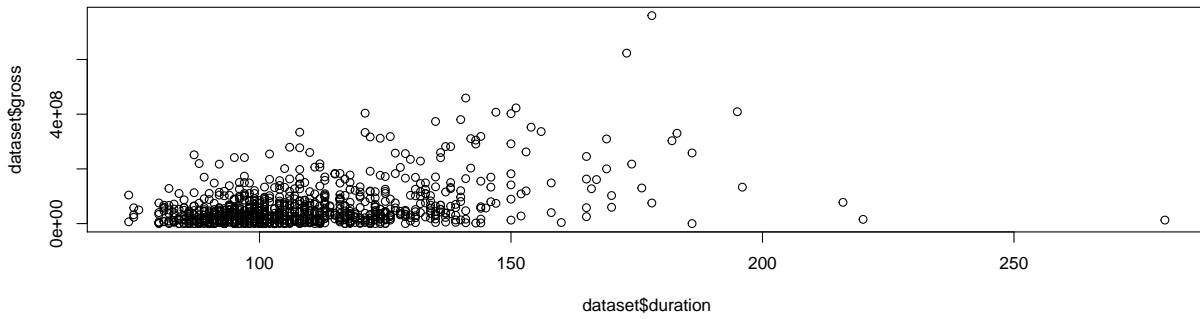
## Duration

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.



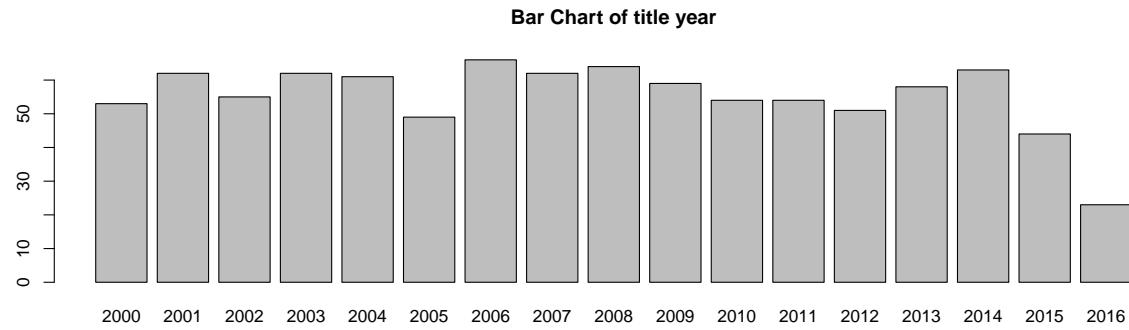
## Relationship between duration and gross

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



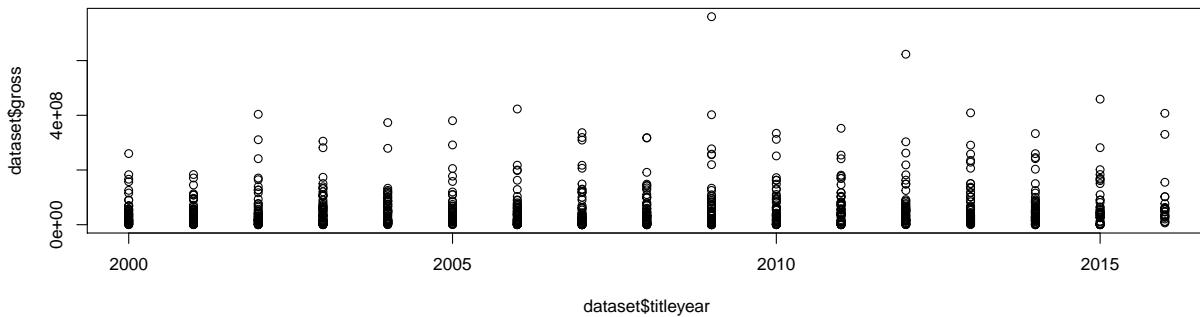
## Title Year

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.



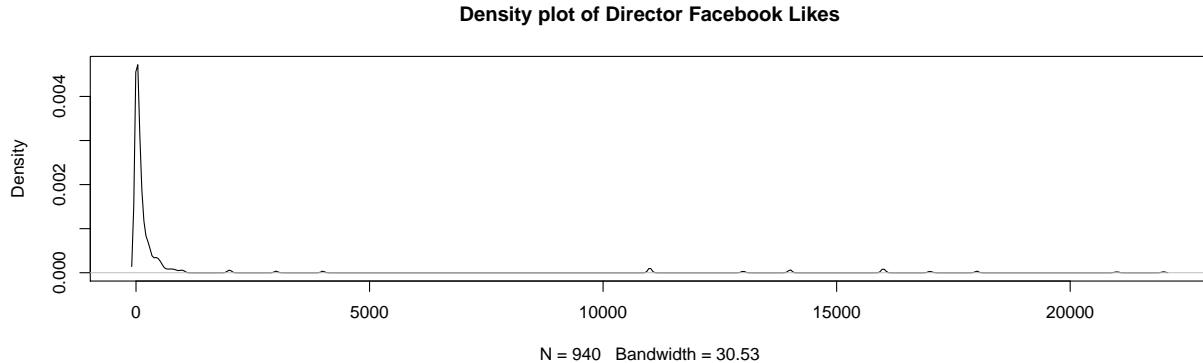
## Relationship between years and gross

We can see that there is not a strong correlation between the year of release and the gross obtained from those years. Even that there are some years better than others.



### **Directorfl, Actor1fl, Actor2fl, Actor3fl, Castfl**

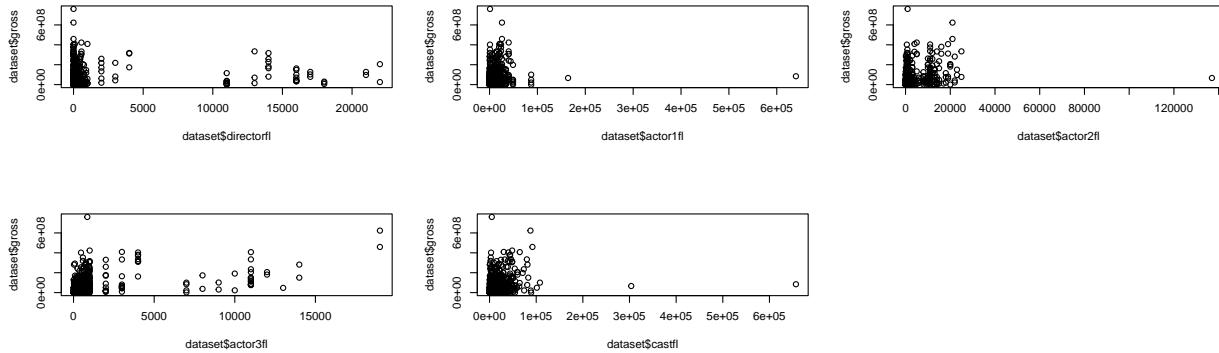
In Director Facebook likes, we see that there is a value which appears in the majority of the cases: In this case, is the 0 value. Apart from this zero value, we see that small number of likes are more common than medium or higher number of likes.



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

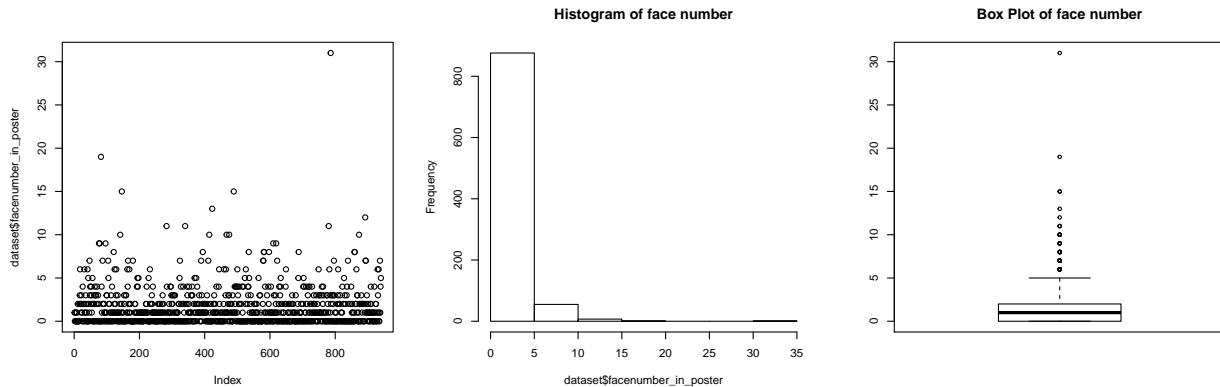
### **Relationship of Directorfl, Actor1fl, Actor2fl, Actor3fl, Castfl with gross**

We can see a trend where the more likes Directors, Acts or Cast are they tend to have more revenue.

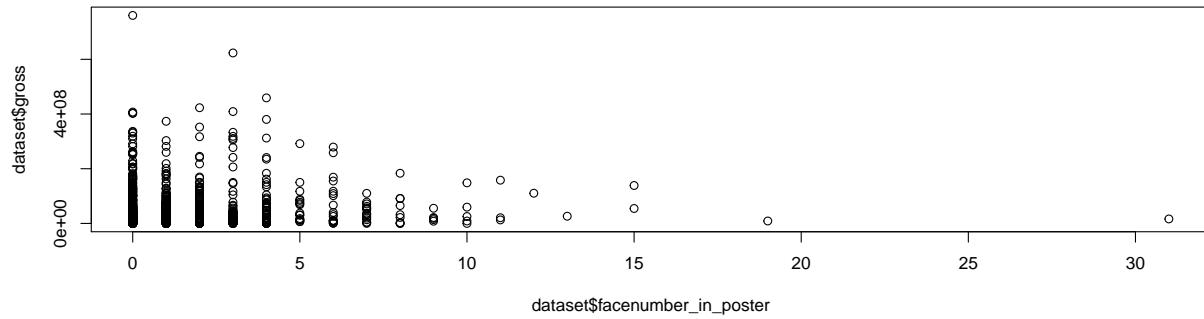


### **Facenumber in poster**

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.

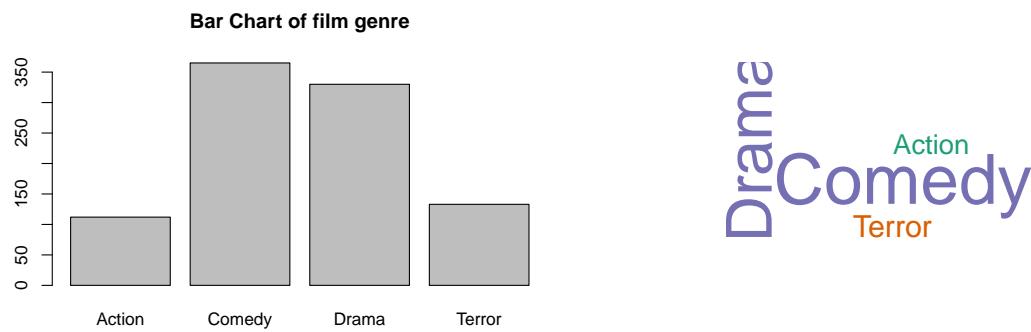


### Relationship of Facenumber with Gross



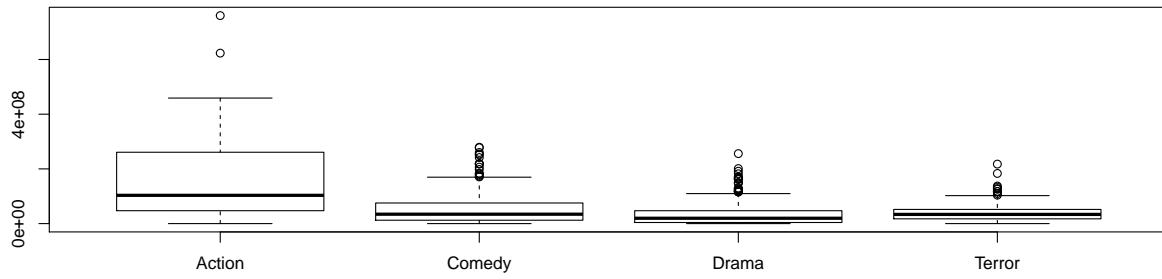
### Genre

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



### Relationship of Genre with Gross

We can see that action films from the dataset tend to have a slightly more income than drama, terror and comedy films. We also can see that there are more outliers in the comedy genre.

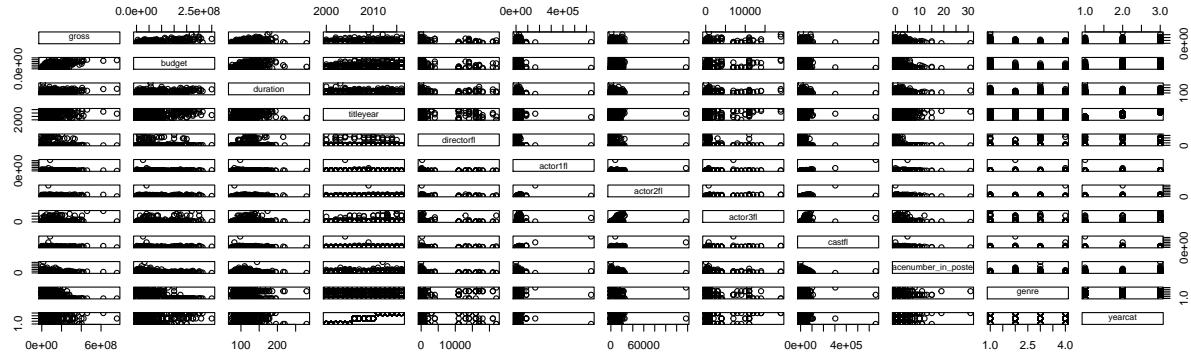


## Correlation matrix

In this section we can see the correlation between the variables. It is a bit of what we have seen but in this case is not only focused with the gross but all the variables. Furthermore, by doing the correlation of the variables, we have a numeric value that says us how correlated two variables are which we did not quantify when doing the exploratory analysis.

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.

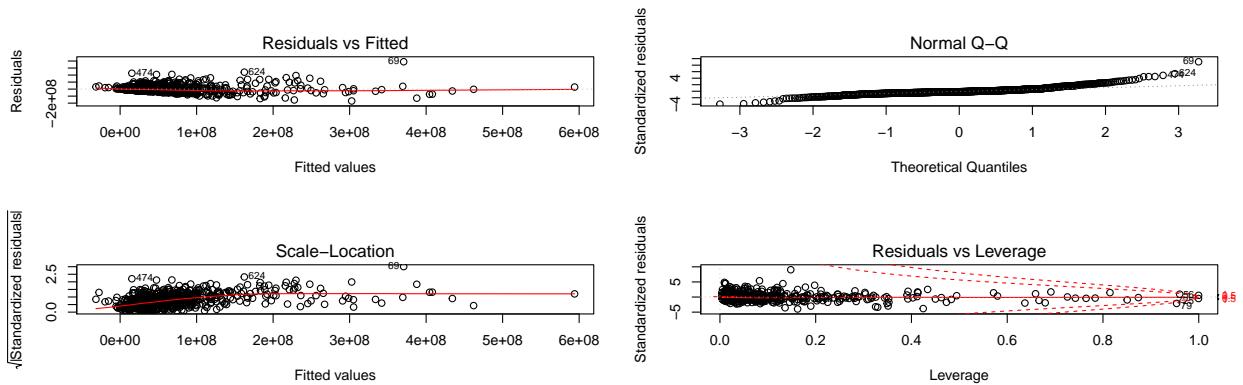
## EXPLICAR MÉS



	gross	budget	duration	directorfl	actor1fl	actor2fl	actor3fl	castfl	facenumber_in_poster
gross	1.000000	0.7295407	0.4169113	0.1135821	0.1203821	0.2516909	0.3897209	0.2117592	0.0029669
budget	0.7295407	1.000000	0.4807725	0.1185898	0.1384697	0.2587875	0.3404662	0.2198807	-0.0065147
duration	0.4169113	0.4807725	1.000000	0.2152645	0.0649505	0.1288276	0.1809332	0.1070346	-0.0123550
directorfl	0.1135821	0.1185898	0.2152645	1.000000	0.0660325	0.0940824	0.0453623	0.0825869	-0.0843436
actor1fl	0.1203821	0.1384697	0.0649505	0.0660325	1.000000	0.3491797	0.2377791	0.9618000	0.0551944
actor2fl	0.2516909	0.2587875	0.1288276	0.0940824	0.3491797	1.000000	0.4591963	0.5725142	0.0258751
actor3fl	0.3897209	0.3404662	0.1809332	0.0453623	0.2377791	0.4591963	1.000000	0.4160769	0.0847390
castfl	0.2117592	0.2198807	0.1070346	0.0825869	0.9618000	0.5725142	0.4160769	1.000000	0.0650337
facenumber_in_poster	0.0029669	-0.0065147	-0.0123550	-0.0843436	0.0551944	0.0258751	0.0847390	0.0650337	1.000000

## Fitting the complete model

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



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

```
kable(summary(forwardModel)$coefficients, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="scale_down")
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.926539e+08	2.534451e+07	-7.601407	0.0000000
budget	8.626975e-01	5.979130e-02	14.428473	0.0000000
actor3fl	-1.338824e+04	4.910479e+03	-2.726463	0.0065223
duration	1.755780e+06	2.285396e+05	7.682607	0.0000000
genreComedy	1.605711e+08	3.241567e+07	4.953503	0.0000009
genreDrama	1.974540e+08	2.816839e+07	7.009773	0.0000000
genreTerror	1.871068e+08	3.735289e+07	5.009167	0.0000007
actor3fl:duration	1.533793e+02	3.791719e+01	4.045111	0.0000566
duration:genreComedy	-1.189038e+06	2.997022e+05	-3.967398	0.0000783
duration:genreDrama	-1.715470e+06	2.391787e+05	-7.172338	0.0000000
duration:genreTerror	-1.484417e+06	3.446298e+05	-4.307278	0.0000183

```
kable(summary(backwardModel)$coefficients, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="scale_down")
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-2.139399e+08	2.591216e+07	-8.2563544	0.0000000
budget	8.249304e-01	6.047660e-02	13.6404981	0.0000000
duration	1.946699e+06	2.349736e+05	8.2847576	0.0000000
directorfl	6.207606e+02	1.032026e+03	0.6014971	0.5476571
actor1fl	5.263282e+02	2.916016e+03	0.1804956	0.8568032
actor2fl	-5.464719e+03	1.752279e+03	-3.1186346	0.0018734
actor3fl	-5.693808e+03	2.408883e+03	-2.3636712	0.0183019
castfl	1.757356e+03	2.534712e+03	0.6933158	0.4882864
genreComedy	1.624669e+08	3.223146e+07	5.0406316	0.0000006
genreDrama	2.050181e+08	2.835780e+07	7.2296899	0.0000000
genreTerror	1.889485e+08	3.720772e+07	5.0782061	0.0000005
duration:actor1fl	-6.265962e+01	1.821487e+01	-3.4400266	0.0006078
duration:castfl	3.933400e+01	1.290360e+01	3.0482975	0.0023672
duration:genreComedy	-1.232875e+06	2.974126e+05	-4.1453361	0.0000371
duration:genreDrama	-1.797363e+06	2.421253e+05	-7.4232776	0.0000000
duration:genreTerror	-1.512632e+06	3.438891e+05	-4.3986032	0.0000122
directorfl:actor1fl	1.636947e+00	4.690924e-01	3.4896048	0.0005067
directorfl:actor2fl	1.462274e+00	4.639009e-01	3.1521262	0.0016733
directorfl:actor3fl	2.882406e+00	9.254108e-01	3.1147313	0.0018981
directorfl:castfl	-1.535387e+00	4.667675e-01	-3.2894048	0.0010421

Check the presence of multicollinearity. If there is some non-interaction multicollinearity in the model, make the corresponding corrections.

```
kable(car::vif(forwardModel), format="latex", booktabs=TRUE)
```

	GVIF	Df	GVIF^(1/(2*Df))
budget	3.541305	1	1.881835
actor3fl	37.395201	1	6.115162
duration	9.061749	1	3.010274
genre	70073.898079	3	6.420853
actor3fl:duration	38.483049	1	6.203471
duration:genre	75721.204208	3	6.504335

```
kable(car::vif(backwardModel), format="latex", booktabs=TRUE)
```

	GVIF	Df	GVIF^(1/(2*Df))
budget	3.731434	1	1.931692
duration	9.866015	1	3.141021
directorfl	3.980337	1	1.995078
actor1fl	2027.247202	1	45.024962
actor2fl	47.455214	1	6.888775
actor3fl	9.268605	1	3.044438
castfl	2066.844582	1	45.462562
genre	75813.253895	3	6.505652
duration:actor1fl	810.510092	1	28.469459
duration:castfl	588.713938	1	24.263428
duration:genre	82558.200118	3	6.598725
directorfl:actor1fl	308.695523	1	17.569733
directorfl:actor2fl	51.927199	1	7.206053
directorfl:actor3fl	31.930128	1	5.650675
directorfl:castfl	742.590291	1	27.250510

```
finalModel <- lm(gross ~ budget + duration, dataset)
```