Marcel Pons Cloquells

Report 1. Linear Models

1. IMDB Dataset

In this project, data from the IMDb[1] film database will be analyzed. The dataset obtained from this database consists of 940 films released between 2000 and 2016 and several variables associated with them, which are represented in the following table:

VARIABLE	DESCRIPTION
movietitle	Title of the movie.
gross	Total income earned from theatres. In dollars.
budget	Cost of the production of the movie. In dollars.
duration	Film duration in minutes.
titleyear	Release year of the film (200-2016)
directorfl	Director Facebook likes.
actor1fl	Actor 1 Facebook likes.
actor2fl	Actor 2 Facebook likes.
actor3fl	Actor 3 Facebook likes.
castfl	Cast Facebook likes.
facenumberinposter	Number of faces that appear in the poster.
genre	Action / Comedy / Drama / Terror

As can be seen, all variables (10) are of continuous type except the categorical variable *genre*. In the analysis, *movietitle* will be used as row names, not as a variable.

2. Objective

The objective of the analysis consists of first performing an exploratory data analysis of the dataset in order to get insightful information about the relations that can exist between variables.

Then, a linear regression model will be built having as response variable the *gross* of each film and as explanatory variables the ones that happen to be significant for the model. For the selection of these significant variables, the stepwise procedure using the BIC criterion will be used. Besides, the presence of multicollinearity will be assessed using the Variance Inflation Factor (VIF). Finally, the assumptions for the validation of the model will be analyzed and the model will be interpreted.

[1] https://www.imdb.com/

3. Exploratory Data Analysis

A basic description of the IMDb dataset is provided using the `summary` function in R.

```
     gross                budget              duration       titleyear
 Min.   :     3330   Min.   :   400000   Min.   : 74.0   Min.   :2000
 1st Qu.: 11816543   1st Qu.: 10000000   1st Qu.: 95.0   1st Qu.:2004
 Median : 33428175   Median : 24000000   Median :104.0   Median :2008
 Mean   : 57813237   Mean   : 40484550   Mean   :108.9   Mean   :2008
 3rd Qu.: 70756664   3rd Qu.: 48000000   3rd Qu.:119.0   3rd Qu.:2012
 Max.   :760505847   Max.   :300000000   Max.   :280.0   Max.   :2016
   directorfl         actor1fl            actor2fl            actor3fl
 Min.   :    0.0   Min.   :     0.0   Min.   :     0.0   Min.   :    0.0
 1st Qu.:   11.0   1st Qu.:   831.5   1st Qu.:   462.5   1st Qu.:  255.0
 Median :   56.0   Median :  2000.0   Median :   756.0   Median :  501.0
 Mean   :  757.2   Mean   :  9006.8   Mean   :  2391.7   Mean   :  891.1
 3rd Qu.:  189.8   3rd Qu.: 13000.0   3rd Qu.:  1000.0   3rd Qu.:  748.2
 Max.   :22000.0   Max.   :640000.0   Max.   :137000.0   Max.   :19000.0
     castfl        facenumber_in_poster    genre
 Min.   :     0   Min.   : 0.000       Action:112
 1st Qu.:  2422   1st Qu.: 0.000       Comedy:365
 Median :  4868   Median : 1.000       Drama :330
 Mean   : 13466   Mean   : 1.624       Terror:133
 3rd Qu.: 17659   3rd Qu.: 2.000
 Max.   :656730   Max.   :31.000
```

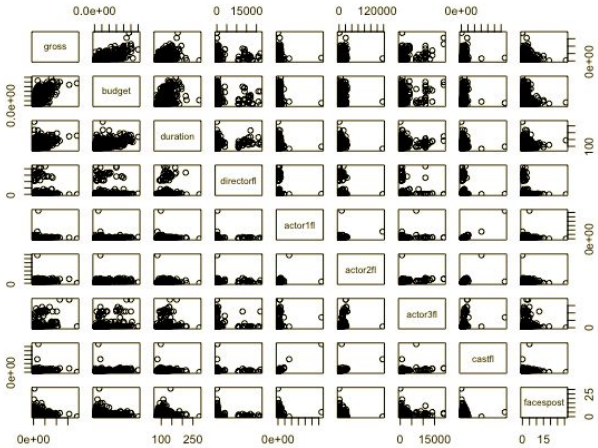
It can be seen that the variables *budget* and *gross* present a high range of magnitudes (from 10^5 to 10^9 in *budget* and from 10^3 to 10^9 in *gross*). Due to this high variability, it could be better for the regression analysis to have at least the response variable *gross* in `log10()` scale, otherwise this high variability might affect the validation of the model. Other variables, like *actor1fl* and *actor2fl,* also seem to present high variability, hence when performing exploratory visualizations, some movies will be far away from the others, complicating the interpretation.

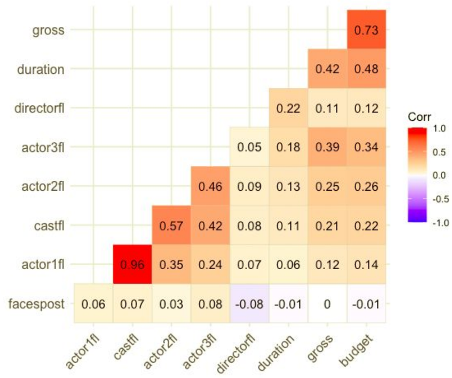
On the other hand, the continuous variable *titleyear* will be transformed into a categorical variable (*yearcat*) consisting of year intervals (2000-2005, 2006-2010, and 2011-2016). This transformation will help to determine if there are differences in the values of the continuous variables between year intervals and also will allow us to study the interaction between years and the other variables.

```
# Years go from 2000 to 2016
Years_bins <- c(2005, 2010)
Years_modalities <- c("2000-2005","2006-2010","2011-2016")
df$titleyear[which(df$titleyear<=Years_bins[1])] <- Years_modalities[1]
df$titleyear[which(df$titleyear>Years_bins[1] &
df$titleyear<=Years_bins[2])] <- Years_modalities[2]
df$titleyear[which(df$titleyear>Years_bins[2])] <- Years_modalities[3]

df$titleyear <- as.factor(df$titleyear)
df <- rename(df, yearcat = titleyear)
```

The number of films in each interval is well balanced, with 342 films between 2000-2005, 305 between 2006-2010, and 293 between 2011-2016.

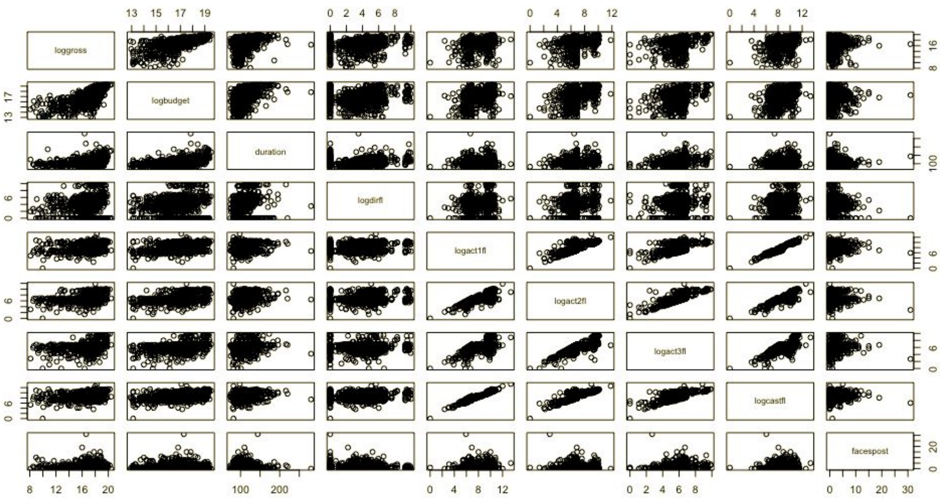
Once the *titleyear* is changed to categorical type, we proceed to assess the relations that exist between the continuous variables. For this purpose, two plots are made using the functions `pairs(~,df)` and `ggcorrplot()`. The first function returns a grid plot consisting of two-dimensional scatter plots between all the continuous variables and the second one returns a plot with the Pearson correlation coefficients between them.





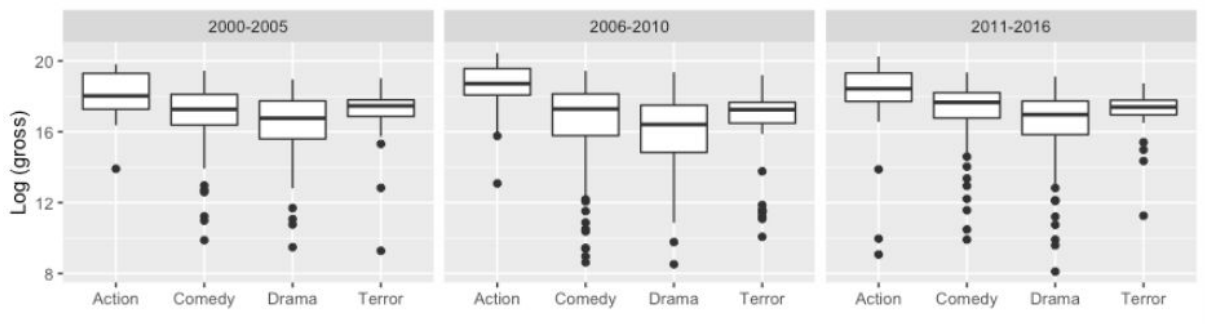
From the correlation plot, it can be seen that the response variable to be studied in the linear model, the *gross* of a film, is quite linearly correlated with its *budget* (0.73), to some extent with its *duration* (0.48), and a little with the actors' Facebook likes (~0.34).

Moreover, it seems that correlations between Facebook likes variables are present, especially between *castfl* and *actor1fl* (0.96). On the other hand, the `pairs()` plot does not seem to give enough visual information due to the high range of scale in some variables (as commented before), hence the visualization of the linear trends between variables would be more appreciable if they were in log scale.



In fact, now in a logarithmic scale very clear linear trends can be seen, especially between the variables related to Facebook likes. Therefore, these variables may be candidates to present multicollinearity when performing the linear regression.

With respect to the categorical variables, it could be interesting to study if significant differences in the response variable *gross* appear between the different levels of both *genre* and *yearcat.* For this reason, a box plot will be performed using the *gross* in the logarithm scale.



It seems that few differences in *gross* are observed between different year intervals (in all *genres*). Regarding the *genre* of the film, significant visual differences cannot be appreciated either. We can see that action related films, in general, obtain a higher *gross* than the other genres. On the other hand, the other genres seem to obtain similar *gross*.

4. Linear Regression Analysis

The exploratory data analysis has helped us to get some insights about the data and the relations between the variables. Now, a linear regression model will be built having as response variable the *gross* of a film.

First of all, we will fit a complete model with all the numerical variables `(. -yearcat-genre)` with their interaction with *genre* and *yearcat* `*(genre+yearcat)` as well as the categorical variables[2] plus their interaction `(genre:yearcat)`. Since the *gross* variable presented a high range of values and thus possibly affecting the validation of the model, it would be better to model it on a logarithmic scale.

```
model <- lm(log(gross)~(. -yearcat-genre)*(genre+yearcat)+genre:yearcat,
data=df)
summary(model)
```

The complete model results to be more explanatory than the null model without variables due to the fact that the ANOVA test which compares both models (Omnibus test) returns a p-value lower than 0.05. However, this model does not explain well enough the *gross* of a film with these explanatory variables and interactions (only 36% according to the adjusted R-squared).

```
Residual standard error: 1.628 on 880 degrees of freedom
Multiple R-squared:  0.4056, Adjusted R-squared:  0.3657
F-statistic: 10.18 on 59 and 880 DF,  p-value: < 2.2e-16
```

[2] Dummy variables are encoded with the treatment contrast (default).

With respect to the variables and coefficients (which can be consulted in Annex), *budget* and *duration* seem to be significant for the model (p-value < 0.05). Moreover, several significant interactions between continuous and categorical variables appear, such as *budget:genreComedy* and *actor1fl:genreDrama*. Interactions between some levels between the two categorical variables are important as well, like *yearcat2011-2016:genreDrama*.

The complete model results to be weak for explaining the response variable and it has a lot of variables and interactions that happen to be not significant. This low performance may be due to these coefficients that are not significant to explain the *gross* of a film. Furthermore, multicollinearity could occur between some variables like *castfl* and *actor1fl* which resulted to be highly correlated in the EDA. Therefore, it is necessary to perform some kind of selection of the most important variables for explaining the response variable.

To carry out this selection, the backward stepwise feature selection using the BIC criterion will be used (`step` function). The stepwise regression is a step-by-step iterative construction which aims to select the best explanatory variables to be used in a final model. The “backward” argument specifies that the stepwise starts with the complete model and from there it deletes one variable at a time, testing if the removed variable is statistically significant. If so, it keeps the variable and continues the iteration with the next variable.

The criteria for deciding if a variable stays in the model or not can be based on the AIC or the BIC. We will use the BIC (Bayesian Information Criterion), which is based on the likelihood function, which increases when adding more parameters, thus this index introduces a penalty term for the number of parameters in the model. If in a step the removal of a variable decreases the BIC, that variable is kept away from the model.

```
model <- step(model, direction = 'back', k = log(nrow(df)))
```

```
Start:  AIC=1265.47
log(gross) ~ ((budget + duration + yearcat + directorfl + actor1fl +
    actor2fl + actor3fl + castfl + facespost + genre) - yearcat -
    genre) * (genre + yearcat) + genre:yearcat

                     Df Sum of Sq    RSS    AIC
- yearcat:genre       6    38.599 2372.2 1239.8
- directorfl:genre    3     2.181 2335.7 1245.8
- actor3fl:genre      3     5.621 2339.2 1247.2
- duration:genre      3     8.248 2341.8 1248.2
- actor2fl:genre      3    12.202 2345.8 1249.8
- castfl:genre        3    12.210 2345.8 1249.8
- actor1fl:genre      3    12.248 2345.8 1249.8
- yearcat:actor1fl    2     0.452 2334.0 1252.0
- yearcat:directorfl  2     0.669 2334.2 1252.0
- yearcat:castfl      2     0.774 2334.3 1252.1
- yearcat:actor3fl    2     1.961 2335.5 1252.6
- yearcat:actor2fl    2     2.823 2336.4 1252.9
- budget:yearcat      2     8.615 2342.2 1255.2
- duration:yearcat    2    10.956 2344.5 1256.2
- yearcat:facespost   2    21.942 2355.5 1260.6
- facespost:genre     3    43.309 2376.9 1262.2
<none>                            2333.6 1265.5
- budget:genre        3    65.321 2398.9 1270.9
```

In the above image we can see the first step of the selection. The initial BIC is 1265.47 and the removal of all variables, except from *budget:genre*, decreases the index.

```
Step:  AIC=1012.85
log(gross) ~ budget + duration + castfl + genre + yearcat + budget:genre

               Df Sum of Sq    RSS    AIC
- castfl        1     5.404 2535.4 1008.0
<none>                      2530.0 1012.9
- duration      1    20.378 2550.4 1013.5
- yearcat       2    51.763 2581.8 1018.2
- budget:genre  3   137.187 2667.2 1042.0

Step:  AIC=1008.01
log(gross) ~ budget + duration + genre + yearcat + budget:genre

               Df Sum of Sq    RSS    AIC
<none>                      2535.4 1008.0
- duration      1    19.927 2555.3 1008.5
- yearcat       2    51.456 2586.9 1013.2
- budget:genre  3   137.279 2672.7 1037.0
```

The last steps of the function are shown above. The BIC has decreased from 1265.47 to 1008. We can see that the last variable to be removed is *castfl* and from there the deletion of the remaining variables increases the BIC, thus staying in the final model.

According to the stepwise function, the best model to explain *gross* is the one that consists of *budget, duration, castfl, genre, yearcat* and the interaction *budget:genre.* If the summary of this model is called, we obtain:

```
Call:
lm(formula = log(gross) ~ budget + duration + genre + yearcat +
    budget:genre, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4984 -0.5042  0.3013  1.0421  3.3325

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.495e+01  4.792e-01  31.192  < 2e-16 ***
budget              1.548e-08  2.395e-09   6.465 1.63e-10 ***
duration            9.815e-03  3.632e-03   2.702 0.007016 **
genreComedy        -1.056e-01  3.725e-01  -0.284 0.776789
genreDrama         -6.820e-01  3.750e-01  -1.819 0.069293 .
genreTerror         5.158e-01  3.973e-01   1.298 0.194553
yearcat2006-2010   -4.869e-01  1.312e-01  -3.711 0.000219 ***
yearcat2011-2016    3.333e-02  1.344e-01   0.248 0.804202
budget:genreComedy  2.365e-08  4.068e-09   5.814 8.39e-09 ***
budget:genreDrama   2.438e-08  4.561e-09   5.346 1.13e-07 ***
budget:genreTerror  1.020e-08  6.455e-09   1.580 0.114451
---
Residual standard error: 1.652 on 929 degrees of freedom
Multiple R-squared:  0.3542, Adjusted R-squared:  0.3472
F-statistic: 50.95 on 10 and 929 DF,  p-value: < 2.2e-16
```

As can be seen, the final model is much simpler than the complete one that we started from. The Omnibus test is still significant and all the variables selected are significant for the model (p-value < 0.05). However, the Adjusted R-squared has decreased a little (from 0.365 to 0.347), possibly because some variable that we removed explained a tiny part of the response variable but following the parsimony principle it was better to remove it.

The last thing that needs to be assessed in the model is the presence of multicollinearity, that is, when there is a strong correlation between explanatory variables which can adversely affect the regression results. The VIF (Variance Inflation Factor) detects this multicollinearity, with values >5 expressing high correlation between variables[3].

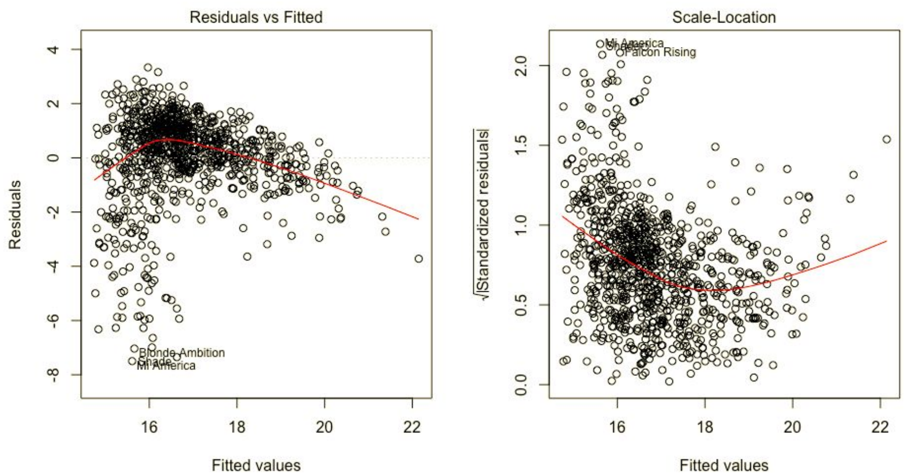
```
vif(model)

                  GVIF Df GVIF^(1/(2*Df))
 budget        4.869460  1        2.206685
 duration      1.962188  1        1.400781
 genre        21.262966  3        1.664450
 yearcat       1.057829  2        1.014154
 budget:genre 10.349120  3        1.476218
```

After calling the `vif()` function, we can see that *genre* presents a high VIF (although the corrected GVIF with the degrees of freedom is not high). We could think of removing it from the model; nevertheless, this variable has a significant interaction with budget and it would be better to keep it[4].

5. Validation

Once the final model is obtained, it has to be validated, that is, we have to check if the selected variables are acceptable as descriptors of the data. For this reason, four assumptions for a linear regression model have to be met: linearity, homoscedasticity, independence and normality. The validation of these assumptions can be assessed by visualizing the residuals in different ways which the `plot(model)` allows.

Linearity and homoscedasticity



The plots in the image above correspond to the outputs of `plot(model, 1)` and `plot(model,3)`. These plots show whether the residuals (normal residuals in left plot and standardized residuals in right plot) are spread equally along the ranges of the predictor variables without following any patterns, meaning constant variances and thus

---

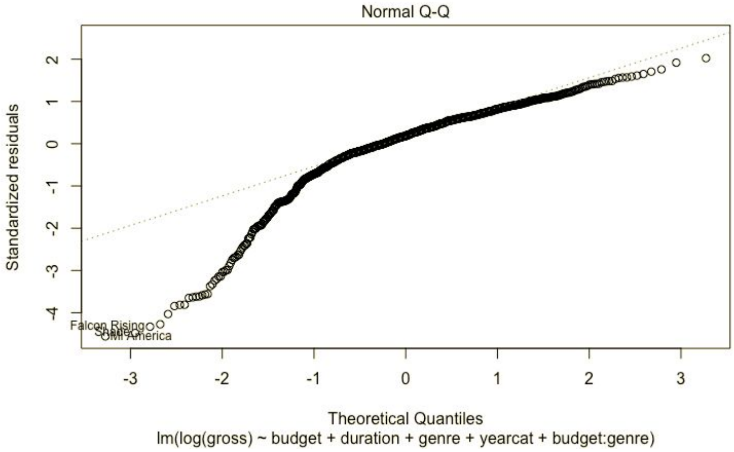
[3] VIF of the complete model can be consulted in the Annex.

[4] If we remove this variable and build a new regression model without it, there is a significant drop in the adjusted R-squared (to 0.2648) and other variables are no longer significant.

homoscedasticity (straight red line). Linearity also can be assessed looking for linear trends in the residuals.

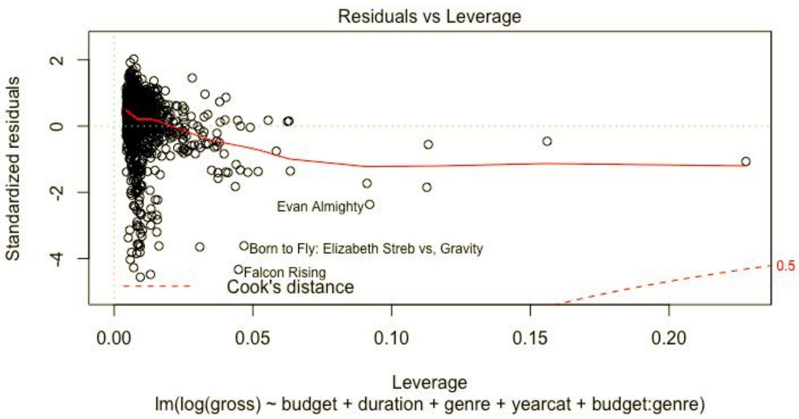
Therefore, by looking at the plots of our model, it seems that the residuals do not follow linearity either homoscedasticity. We can see that the red line is not straight, it has a sort of quadratic relationship; and the residuals are not randomly distributed.

Normality of residuals



With `plot(model, 2)` we can assess if the residuals follow a normal distribution. In the case of our model, it looks like they are not following the desired distribution due to the fact that there are several observations that fall far away from the theoretical line[5].

Presence of outliers



The last plot that we assess is the Residuals vs Leverage (`plot(model,5)`). This plot helps to detect the presence of influential points and outliers, which are the observations that are significantly far away from the others. We can see some films that are far away from the others, such as *Rent* (furthest point) and *Evan Almighty*. The presence of these films can influence the model and the other assumptions (which happened to be rejected)

[5] If we perform a normality test like the Shapiro-Wilk for the residuals, its p-value is lower than 0.05 and therefore the normality assumption is not met.

but they seem to be extreme values that appear in the film industry, not errors of measure. Therefore, if we remove them we will lose information about the reality of this sector.

6. Interpretation

The final linear regression model for the *gross* variable has as explanatory variables *budget* and *duration* as continuous, *genre* and *yearcat* as categorical; and the interaction between *budget* and *genre.* The genreAction and yearcat2000-2006 are integrated inside the intercept (since we are using the treatment contrast).

```
Call:
lm(formula = log(gross) ~ budget + duration + genre + yearcat +
    budget:genre, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4984 -0.5042  0.3013  1.0421  3.3325

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.495e+01  4.792e-01  31.192  < 2e-16 ***
budget              1.548e-08  2.395e-09   6.465 1.63e-10 ***
duration            9.815e-03  3.632e-03   2.702 0.007016 **
genreComedy        -1.056e-01  3.725e-01  -0.284 0.776789
genreDrama         -6.820e-01  3.750e-01  -1.819 0.069293 .
genreTerror         5.158e-01  3.973e-01   1.298 0.194553
yearcat2006-2010   -4.869e-01  1.312e-01  -3.711 0.000219 ***
yearcat2011-2016    3.333e-02  1.344e-01   0.248 0.804202
budget:genreComedy  2.365e-08  4.068e-09   5.814 8.39e-09 ***
budget:genreDrama   2.438e-08  4.561e-09   5.346 1.13e-07 ***
budget:genreTerror  1.020e-08  6.455e-09   1.580 0.114451
---
Residual standard error: 1.652 on 929 degrees of freedom
Multiple R-squared:  0.3542, Adjusted R-squared:  0.3472
F-statistic: 50.95 on 10 and 929 DF,  p-value: < 2.2e-16
```

The coefficients try to reflect the linear relationship between the explanatory variable and the response one (*gross*), for example, an increase by one unit of duration (i.e., increment the film by one minute), would imply an increase on log(*gross*) of 0.0098. On the other hand, if the *genre* of a film is Drama, according to the model, its log(*gross)* would be 0.682 lower than the Intercept (corresponding to action genre).

We could try to predict the *gross* of a film with the explanatory variables. To that end, we create instances of two films with their corresponding values for all variables. Then, we use the `predict` function for a single new observation ("prediction") and for the several films with these values ("confidence").

	gross	budget	duration	yearcat	directorfl	actor1fl	actor2fl	actor3fl	castfl	facesoost	genre
1	64001297	1.5e+07	94	2006-2010	87	17000	975	569	20154	1	Comedy
2	127968405	8.0e+07	138	2006-2010	17000	29000	223	163	29585	0	Terror

We have to take into account that the `predict` function will use the logarithm of the *gross*, so the output of the predict will have to be exponentiated to revert the logarithmic transformation.

```
exp(predict(model,newdata=dfnew,interval="prediction"))

       fit       lwr        upr
 1  7758493  300154.4  200544149
 2 96497407 3450765.9 2698458748

exp(predict(model,newdata=dfnew,interval="confidence"))

       fit      lwr       upr
 1  7758493  6005172  10023727
 2 96497407 44952954 207144331
```

As we can see, the prediction is not very precise. In the first film, which had an actual *gross* of 64.001.297 dollars, our model predicted a *gross* of $7.758.493, which is significantly lower than the real value. With respect to the second film, the model performs better, predicting a *gross* of 96.497.407 dollars being the actual *gross* $127.968.405.

Regarding the intervals, we can see that the both intervals for "prediction" and "confidence" are very large, being the "prediction" one larger as expected.

7. Conclusions

To sum up, it can be said that the linear regression model that we built was not precise enough to be useful both for inference statistics nor predicting.

That low precision might be due to the fact that the explanatory variables only come to explain 34.72% of the response variable (Adjusted R-squared). Furthermore, the model does not fulfill the homoscedasticity, normality and linearity assumptions, making it an invalid model.

For the IMDB dataset, it could be better to use other models (e.g., polynomial regression) than linear regression because the relationship of the *gross* of a film might not be linear with the other variables.

8. Annex

Coefficients of the complete model:

```
Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   1.669e+01  1.117e+00  14.950  < 2e-16 ***
budget                        1.033e-08  5.204e-09   1.986 0.047349 *  
duration                      2.974e-04  1.055e-02   0.028 0.977518    
directorfl                    1.970e-05  5.526e-05   0.357 0.721536    
actor1fl                      5.899e-05  1.562e-04   0.378 0.705715    
actor2fl                      1.629e-04  1.672e-04   0.974 0.330132    
actor3fl                      1.868e-04  2.699e-04   0.692 0.488870    
castfl                       -7.426e-05  1.543e-04  -0.481 0.630472    
facespost                     2.805e-02  1.081e-01   0.260 0.795284    
genreComedy                  -2.534e+00  1.249e+00  -2.028 0.042811 *  
genreDrama                   -1.648e+00  1.156e+00  -1.426 0.154291    
genreTerror                   5.790e-01  1.582e+00   0.366 0.714469    
yearcat2006-2010             -2.486e+00  1.078e+00  -2.306 0.021358 *  
yearcat2011-2016             -2.219e+00  1.012e+00  -2.192 0.028608 *  
budget:genreComedy            2.437e-08  5.432e-09   4.485 8.24e-06 ***
budget:genreDrama             2.442e-08  6.249e-09   3.907 0.000101 ***
budget:genreTerror            1.955e-08  9.754e-09   2.004 0.045378 *  
budget:yearcat2006-2010       2.814e-09  5.242e-09   0.537 0.591552    
budget:yearcat2011-2016       9.347e-09  5.328e-09   1.754 0.079734 .  
duration:genreComedy          1.512e-02  1.214e-02   1.246 0.213053    
duration:genreDrama           1.843e-03  1.062e-02   0.174 0.862232    
duration:genreTerror         -7.103e-03  1.608e-02  -0.442 0.658774    
duration:yearcat2006-2010     1.905e-02  9.521e-03   2.001 0.045653 *  
duration:yearcat2011-2016     3.435e-03  8.367e-03   0.410 0.681553    
directorfl:genreComedy       -6.841e-05  7.682e-05  -0.890 0.373469    
directorfl:genreDrama        -2.254e-05  5.682e-05  -0.397 0.691720    
directorfl:genreTerror       -3.224e-05  8.050e-05  -0.401 0.688853    
yearcat2006-2010:directorfl   2.284e-05  4.546e-05   0.502 0.615482    
yearcat2011-2016:directorfl   1.237e-05  5.234e-05   0.236 0.813173    
actor1fl:genreComedy         -1.667e-04  1.509e-04  -1.105 0.269545    
actor1fl:genreDrama          -3.139e-04  1.471e-04  -2.134 0.033113 *  
actor1fl:genreTerror         -7.948e-05  2.697e-04  -0.295 0.768311    
yearcat2006-2010:actor1fl    -1.664e-05  1.642e-04  -0.101 0.919285    
yearcat2011-2016:actor1fl    -5.811e-05  1.474e-04  -0.394 0.693498    
actor2fl:genreComedy         -2.366e-04  1.640e-04  -1.443 0.149463    
actor2fl:genreDrama          -3.202e-04  1.554e-04  -2.061 0.039584 *  
actor2fl:genreTerror         -9.708e-05  2.718e-04  -0.357 0.721073    

yearcat2006-2010:actor2fl    -1.063e-04  1.765e-04  -0.602 0.546997    
yearcat2011-2016:actor2fl    -1.644e-04  1.595e-04  -1.031 0.302962    
actor3fl:genreComedy         -2.564e-04  2.372e-04  -1.081 0.280097    
actor3fl:genreDrama          -3.283e-04  2.402e-04  -1.366 0.172155    
actor3fl:genreTerror         -6.434e-05  4.662e-04  -0.138 0.890265    
yearcat2006-2010:actor3fl    -7.693e-06  2.618e-04  -0.029 0.976566    
yearcat2011-2016:actor3fl    -1.712e-04  2.402e-04  -0.713 0.476258    
castfl:genreComedy            1.829e-04  1.494e-04   1.224 0.221399    
castfl:genreDrama             3.067e-04  1.448e-04   2.118 0.034455 *  
castfl:genreTerror            9.496e-05  2.668e-04   0.356 0.721942    
yearcat2006-2010:castfl       2.159e-05  1.638e-04   0.132 0.895171    
yearcat2011-2016:castfl       7.533e-05  1.463e-04   0.515 0.606649    
facespost:genreComedy         2.001e-02  1.037e-01   0.193 0.846957    
facespost:genreDrama         -6.705e-02  1.052e-01  -0.638 0.523952    
facespost:genreTerror        -4.813e-01  1.588e-01  -3.032 0.002504 ** 
yearcat2006-2010:facespost   -1.727e-01  7.339e-02  -2.353 0.018831 *  
yearcat2011-2016:facespost    1.286e-02  6.042e-02   0.213 0.831545    
yearcat2006-2010:genreComedy  3.780e-01  6.521e-01   0.580 0.562298    
yearcat2011-2016:genreComedy  1.521e+00  6.285e-01   2.420 0.015738 *  
yearcat2006-2010:genreDrama   1.229e-01  7.071e-01   0.174 0.861996    
yearcat2011-2016:genreDrama   1.755e+00  6.696e-01   2.620 0.008935 ** 
yearcat2006-2010:genreTerror -3.845e-01  7.417e-01  -0.518 0.604312    
yearcat2011-2016:genreTerror  1.905e+00  7.177e-01   2.654 0.008102 ** 
---
```

VIF of the complete model

```
                        GVIF Df GVIF^(1/(2*Df))
budget          2.367009e+01  1        4.865191
duration        1.703860e+01  1        4.127784
directorfl      9.775315e+00  1        3.126550
actor1fl        4.980248e+03  1       70.570873
actor2fl        3.702661e+02  1       19.242300
actor3fl        9.964060e+01  1        9.982014
castfl          6.562625e+03  1       81.010032
facespost       2.461760e+01  1        4.961613
genre           1.923716e+05  3        7.597841
yearcat         4.465735e+03  2        8.174726
budget:genre    6.816824e+01  3        2.021143
budget:yearcat  1.439920e+02  2        3.464054
duration:genre  3.103337e+05  3        8.228200
duration:yearcat 4.122458e+03  2       8.012888
directorfl:genre 1.188578e+01  3       1.510676
yearcat:directorfl 4.264142e+00  2     1.437004
actor1fl:genre  1.677454e+09  3       34.470038
yearcat:actor1fl 2.893494e+05  2      23.192932
actor2fl:genre  1.223223e+06  3       10.341517
yearcat:actor2fl 1.932347e+04  2      11.790203
actor3fl:genre  3.119180e+03  3        3.822435
yearcat:actor3fl 8.300133e+02  2       5.367490
castfl:genre    8.579410e+09  3       45.245588
yearcat:castfl  1.680881e+06  2       36.006779
facespost:genre 6.014479e+01  3        1.979397
yearcat:facespost 1.114155e+01  2      1.826991
yearcat:genre   1.522937e+04  6        2.231294
```

As we saw in the Exploratory Data Analysis, the Facebook like variables are highly correlated with elevated VIF.