Student Name:
DNI or PASSPORT:

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística i Investigació Operativa

CURS 16-17 Q1 – FINAL EXAM

Anàlisi de Dades de Transport i Logística (ADTL) .

(Data: 20/1/2017 17:00-21:00 h Lloc: Aula H9.1

Professor responsable:	Lídia Montero Mercadé - Edifici C5 D217 CAMPUS NORD UPC
Normativa de l'examen:	NO APUNTS TEORIA, NI COMANDES DE R. SI - TAULES ESTADÍSTIQUES ES POT DUR CALCULADORA
Durada :	3h
Sortida de notes:	Abans 25 /1/17 al WEB de l'assignatura.
Revisió:	El 25 /1/17 a les 13 hores (C5-217).

Problem 1: Gross Revenue of Hollywood films

IMDB dataset contents data about 470 Hollywood films from the last decade (www.imdb.com) and includes variables (data provided by JA Sánchez):

movie_title:	Name of the film
gross:	Gross revenue (million de $) - Target
budget:	Budget (million $)
duration:	Duration (minutes)
title_year:	Year
actor1_fl:	Popularity of the first actor (number of "Likes" in Facebook)
actor2_fl:	Popularity of the second actor (number of "Likes" in Facebook)
actor3_fl:	Popularity of the third actor (number of "Likes" in Facebook)
cast_fl:	Average casting popularity (number of "Likes" in Facebook)
faces_poster:	Number of actors appearing in promotion card
Genre:	Film category (Comedy, Drama, Action, Horror)

1. Describe the profile for the gross revenue target using the tools available in FactoMineR package

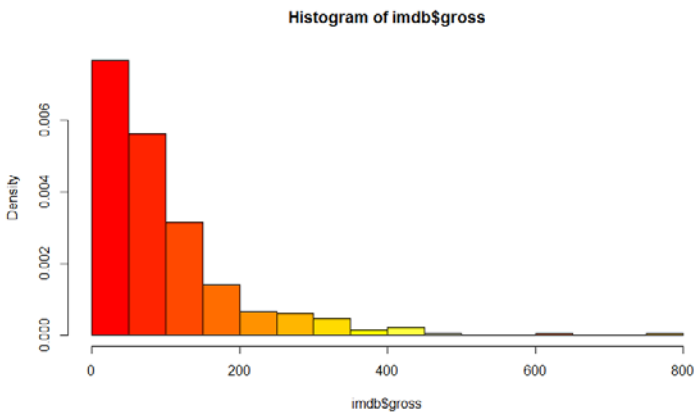
```
> condes(imdb[,2:12],1)
$quanti
           correlation      p.value
budget       0.6835283 5.555625e-66
duration     0.4335185 5.881788e-23
actor3_fl    0.3716480 7.676382e-17
cast_fl      0.3584781 1.066289e-15
actor2_fl    0.3504406 5.008197e-15
actor1_fl    0.2192484 1.594224e-06
…
```

```
$quali
              R2      p.value
Genre 0.256727 8.352179e-30

$category
        Estimate      p.value
Action  85.80388 1.623827e-29
Horror -40.48026 1.370002e-04
Drama  -36.00226 1.587742e-07
```

Budget, duration, actor3_fl, cas_fl, actor2_fl and actor1_fl are numeric variables positively related to the target (gross revenue), in decreasing order, for all of them, the Pearson correlation is different to 0. Genre is globally related to the target also. The mean gross revenue for Genre-

Average Action film revenues are 85. 8 unit significantly over the global mean revenue, while Genre-Horror and Genre-Drama average revenues are significantly less than the overall mean revenue, 40.5 and 36 million $ below.

2. Can the gross revenue target assumed to be normally distributed? Justify your answer.



Clearly, the profile does not belong to normally distributed data. Additionally, the Shapiro-Wilk normality test shows a pvlue very low, thus rejecting the normality hypothesis. Consistent with the visual assessment.

```
> shapiro.test(imdb$gross)
Shapiro-Wilk normality test data:  gross
W = 0.76822, p-value < 2.2e-16
```

Student Name:
DNI or PASSPORT:

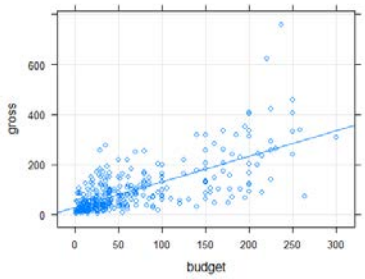
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

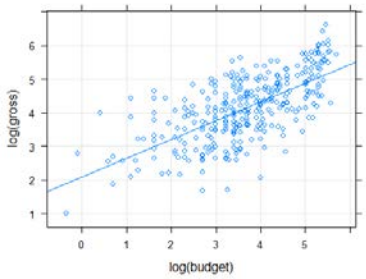
```
> Boxplot(imdb$gross~imdb$Genre,col=heat.colors(4),labels=row.names(imdb))
 [1] "How the Grinch Stole Christmas " "The Hangover Part II "
 [3] "Despicable Me "                  "Meet the Fockers "
 [5] "The Hangover "                   "My Big Fat Greek Wedding "
 [7] "Cinderella "                     "Twilight "
 [9] "The Blind Side "                 "Avatar "
[11] "The Avengers "                   "The Da Vinci Code "
[13] "Ocean's Eleven "
>
> kruskal.test(imdb$gross~imdb$Genre)

        Kruskal-Wallis rank sum test data:  imdb$gross by imdb$Genre
Kruskal-Wallis chi-squared = 87.427, df = 3, p-value < 2.2e-16

> pairwise.wilcox.test(imdb$gross,imdb$Genre)

        Pairwise comparisons using Wilcoxon rank sum test data:  imdb$gross and imdb$Genre

       Comedy  Drama   Action
Drama  2.1e-05 -       -
Action 1.9e-07 1.3e-14 -
Horror 5.8e-05 0.88    6.6e-12

P value adjustment method: holm
```

3. Are all the average gross revenue per Genre equal to the global gross revenue? If it is not the case, indicate those pairs of levels that produce different average revenues.

According to Kruskal Wallis test on means for gross revenues on gemre, there exists any category with a different gross mean. In fact, the pairwise test indicates that mean revenue for Drama and Horror films are not significantly different, but any other pair of categories have different mean revenues.

Modelling of gross revenue target is intended using the available variables. A subset of films premiered after 2005 is selected. Firstly, the linear association between the gross revenue (target) and the budget is analyzed. Two models are calculated, one with the original variables and a second one using logarithmic transformation for both variables.





Model_1

```
lm(formula = gross ~ budget, data = imdb)

Residuals:
     Min       1Q   Median       3Q      Max
-225.86   -31.38   -11.62    23.75   488.97

Coefficients:
             Estimate  Std. Error  t value Pr(>|t|)
(Intercept) 28.50249     5.58396    5.104 5.71e-07 ***
budget       1.02546     0.05914   17.341  < 2e-16 ***

Residual standard error: 70.91 on 320 degrees of freedom
Multiple R-squared:  0.4844, Adjusted R-squared:  0.4828
F-statistic: 300.7 on 1 and 320 DF,  p-value: < 2.2e-16
```

Model_2

```
lm(formula = log(gross) ~ log(budget), data = imdb)

Residuals:
     Min       1Q   Median       3Q      Max
-2.27501  -0.47714  0.00278  0.47987  1.72540

Coefficients:
             Estimate  Std. Error  t value Pr(>|t|)
(Intercept)  2.10093     0.13268    15.83   <2e-16 ***
log(budget)  0.55831     0.03466    16.11   <2e-16 ***

Residual standard error: 0.7232 on 320 degrees of freedom
Multiple R-squared:  0.4478, Adjusted R-squared:  0.4461
F-statistic: 259.5 on 1 and 320 DF,  p-value: < 2.2e-16
```

4. Discuss pros and cons for both modelling options. Which model is suitable to your data? Justify the answer. Calculate the additional gross revenue to be expected for a 100 million dollar film, when increasing the budget by 20%. Make the estimation using both models.

The model with untransformed variables allows a simpler interpretation. The predictor coefficient can be interpreted directly as the absolute change in response for each $ 1 million increase in the movie budget. However, the graph shows a clear increase in the variance of the residuals, as the explanatory variable increases, and consequently the predictions. This would invalidate the model and, in case of not applying the transformations, would force to adjust a model with heteroscedasticity (GLS) instead of OLS model.

E (Gross | Budget) = 28.5 + 1.025 * Budget

On the other hand, the adjusted model with both transformed variables presents better characteristics to validate the premises, since the conditional variance seems more constant. However, the interpretation is more complicated. The coefficient acquires the role of a budget power.

E (Gross | Budget) = exp (2.1) * Budget ^ 0.558

In the absence of checking other premises for which other graphs / tests are needed (such as the normality of residues), the second model would be the one with the best validation. The comparison of two models with different variable response can not be done either with R2 or with R2-adj, since they are not comparable since the variability of the response is measured even in different units (the first in millions Of dollars and the second in logarithm of millions of dollars).

In the first case:

E (Gross | Budget = 100) = 28.5 + 1.025 * 100 = 131

E (Gross | Budget = 120) = 28.5 + 1.025 * 120 = 151.5

Increase = 151.5-131 = 20.5 million dollars

For the second model:

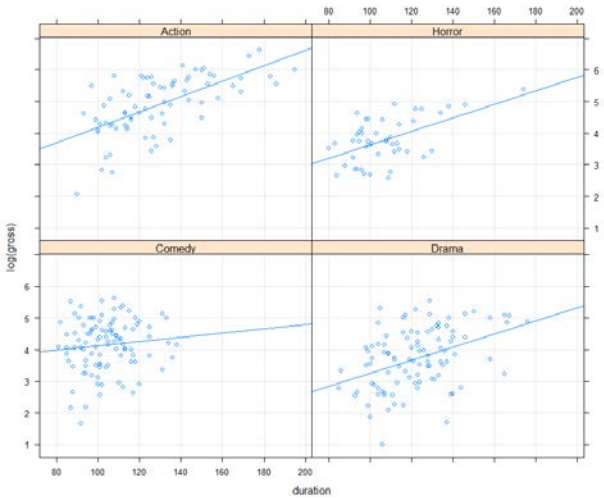
E (Gross | Budget = 100) = exp (2.1) * 100 ^ 0.558 = 106.6

E (Gross | Budget = 120) = exp (2.1) * 120 ^ 0.558 = 118.1

Increase = 118.1-106.6 = 11.5 million dollars

(In this last calculation, assuming the log-normal model, the effect of the residual variance in the calculation of the expected value should be included)

Irrespective of the answer to the first question, it is decided to work with the logarithmic scale of both variables. Next, the relationship between the gross revenue and the duration according to film categories is studied.



The following models are calculated and will be discussed:

Student Name:
DNI or PASSPORT:

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

Model_3

```
lm(formula = log(gross) ~ log(budget) + duration + Genre,
data = imdb)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.398664   0.246849   5.666 3.29e-08 ***
log(budget)  0.464433   0.051939   8.942  < 2e-16 ***
duration     0.010588   0.002404   4.404 1.45e-05 ***
GenreDrama  -0.413373   0.114972  -3.595 0.000376 ***
GenreAction -0.212875   0.130559  -1.630 0.103995
GenreHorror -0.051585   0.130003  -0.397 0.691783

Residual standard error: 0.7015 on 316 degrees of freedom
Multiple R-squared:  0.4869, Adjusted R-squared:  0.4788
F-statistic: 59.98 on 5 and 316 DF,  p-value: < 2.2e-16
```

Model_4

```
lm(formula = log(gross) ~ log(budget) + duration * Genre, data
= imdb)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.148385   0.591499   3.632 0.000328 ***
log(budget)         0.487804   0.053100   9.187  < 2e-16 ***
duration            0.002599   0.005488   0.474 0.636144
GenreDrama         -0.639919   0.760384  -0.842 0.400669
GenreAction        -2.118841   0.730155  -2.902 0.003972 **
GenreHorror        -0.210090   0.870846  -0.241 0.809520
duration:GenreDrama  0.003129   0.006741   0.464 0.642821
duration:GenreAction 0.016022   0.006450   2.484 0.013514 *
duration:GenreHorror 0.001842   0.008085   0.228 0.819909

Residual standard error: 0.6936 on 313 degrees of freedom
Multiple R-squared:  0.5031,    Adjusted R-squared:  0.4904
F-statistic: 39.62 on 8 and 313 DF,  p-value: < 2.2e-16
```

A Fisher test comparing Model_3 and Model_4 is also shown.

```
Analysis of Variance Table

Model 1: log(gross) ~ log(budget) + duration + Genre
Model 2: log(gross) ~ log(budget) + duration * Genre
     Res.Df     RSS       Df    Sum of Sq        F        Pr(>F)
1    --(1)--  --(2)--
2      313    --(3)--   --(4)--   --(5)--     --(6)--    0.01806 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5. Fill in the blanks --(1)-- to --(6)-- in the output above. A description of the calculation procedure has to be included for each value.

(1) Degrees of freedom from model 1: (N-p) = 322-6 = 316

(2) Residual sum of squares of model 1: RSS1 = (N-p) S12 = 316 * 0.70152 = 155.5043

(3) Residual sum of squares of model 2: RSS2 = (N-p) S22 = 313 * 0.69362 = 150.5783

(4) Degrees of freedom of the deviance test: df = df1-df2 = 316-313 = 3

(5) Sum of squares of the deviance test: RS = RSS1-RSS2 = 155.5043-10.5783 = 4.9260

(6) Statistic F of the deviance test: F = ((RSS1-RSS2) / q) / (RSS2 / df2) = (4,926 / 3) / (150.5783 / 313) = 3.41

The p-value of the test is 0.018 which is below the significance level of 5%, which indicates that there is significant statistical evidence that the second model is different from the first and that therefore the interaction in model 2 is Significant. It means that the effect on the revenue depending on the length of the film depends on the genre of the film. The representation of the models segmented by gender indicate that the slope of the adjusted line in each graph can be considered as not the same in the different cases.

6. Does the relationship between gross revenue and duration depend on film category? Justify the answer.

The comparison between the additive ancova model and the complete ancova model has a p.value of 0.018 < 0.05, thus it provides evidence for rejecting the null hypothesis of equivalence between both models, then the relationship between the revenue and the duration depends on film category.

Model 1: log(gross) ~ log(budget) + duration + Genre

```
Model 2: log(gross) ~ log(budget) + duration * Genre
    Res.Df     RSS      Df   Sum of Sq       F       Pr(>F)
1   --(1)--  --(2)--
2     313    --(3)--  --(4)--  --(5)--     --(6)--   0.01806 *
```

7. Write the equations for gross revenue according to Model_4.

For Genre = Comedy

LogGross = 2.148 + 0.00260 * duration + 0.4878 * log (Budget)

For Genre = Drama

LogGross = (2.148 -0.640) + (0.00260 + 0.00313) * duration + 0.4878 * log (Budget)

LogGross = 1.508 + 0.00573 * duration + 0.4878 * log (Budget)

For Genre = Action

LogGross = (2.148 -2.119) + (0.00260 + 0.01602) * duration + 0.4878 * log (Budget)

LogGross = 0.029 + 0.01861 * duration + 0.4878 * log (Budget)

For Genre = Horror

LogGross = (2.148 + 0.210) + (0.00260 + 0.00184) * duration + 0.4878 * log (Budget)

LogGross = 2.358 + 0.00444 * duration + 0.4878 * log (Budget)

Taking into account that the active contrast is Baseline type with the first category as a reference ("contr.treatment"), the coefficients of the categorical variable Genre are interpreted as the change in the coefficient of the ordinate at the origin between a film of the generus Comedy and each of the other genres. In the same way, the interaction is interpreted in term of change in the slope of the different genres with respect to the comedies. Of all the p-values associated with these two terms, only the intercept and slope associated with the action films are significant. Thus, in the case of action films, the difference in the intercept with respect to a comedy is -2.119 significant, and in the slope there is a significant increase of 0.016. For the rest of genres the significance of the difference in any parameter of the model with respect to the comedies has not been established.

Popularity of the main actors is included once duration and film category are already in the model for gross revenue. Model_5 and Model_6 below are discussed.

Model_5

```
lm(formula = log(gross) ~ log(budget) + actor1_fl +
actor2_fl + actor3_fl + cast_fl, data = imdb)

Residuals:
     Min       1Q   Median       3Q      Max
-2.15719 -0.44343  0.03302  0.47068  1.68968
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.174e+00  1.333e-01  16.311  < 2e-16 ***
log(budget)  5.157e-01  3.641e-02  14.163  < 2e-16 ***
actor1_fl   -6.606e-05  2.520e-05  -2.621  0.00919 **
actor2_fl   -4.703e-05  2.787e-05  -1.688  0.09248 .
actor3_fl   -6.456e-05  4.312e-05  -1.497  0.13529
cast_fl      6.129e-05  2.473e-05   2.478  0.01373 *

Residual standard error: 0.708 on 316 degrees of freedom
Multiple R-squared:  0.4773,Adjusted R-squared:  0.4691
F-statistic: 57.72 on 5 and 316 DF,  p-value: < 2.2e-16

> vif(Model_5)
log(budget)   actor1_fl   actor2_fl   actor3_fl     cast_fl
   1.151489   47.399810   14.493582   10.336643  122.549023
```

Model_6

```
lm(formula = log(gross) ~ log(budget) + actor1_fl +
actor2_fl + actor3_fl, data = imdb)

Residuals:
     Min       1Q   Median       3Q      Max
-2.19689 -0.44671  0.02993  0.47326  1.67544
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.179e+00  1.343e-01  16.217   <2e-16 ***
log(budget)  5.251e-01  3.651e-02  14.382   <2e-16 ***
actor1_fl   -4.646e-06  4.623e-06  -1.005    0.316
actor2_fl    1.487e-05  1.246e-05   1.194    0.233
actor3_fl    3.104e-05  1.941e-05   1.599    0.111


Residual standard error: 0.7137 on 317 degrees of
freedom
Multiple R-squared:  0.4672,Adjusted R-squared:  0.4604
F-statistic: 69.48 on 4 and 317 DF,  p-value: < 2.2e-16

> vif(Model_6)
log(budget)   actor1_fl   actor2_fl   actor3_fl
   1.139089    1.569572    2.849449    2.061186
```

8. According to vif() method output, what model would you use to interpret the popularity of actors appearing in the film? Justify the answer. According to Model_6 is there a significant relationship between the popularity of the actors and the revenue?

Variance Inflation Factors (VIFs) are statistics that allow the diagnosis of the presence of multicollinearity in the model. This phenomenon implies that the predictor variables can have a high correlation that prevents a simple interpretation of each coefficient of the model, keeping the rest of the predictors fixed (ceteris paribus). Likewise, the estimation in the presence of multicollinearity is no longer efficient. For each coefficient, the calculated VIF is a measure of how correlated this predictor is with the rest of predictors (If R2 is the coefficient of determination in terms of both 1 of the model that fits with each variable as if it were the answer based on the rest Of predictors, then VIF = 1 / (1-R2)). In the case where the predictors are independent (and there is no multicollinearity) the value of the VIFs will be 1. The further away from 1 indicates a greater presence of multicollinearity. A VIF of approximately 8 indicates that the multicollinearity problem may be important.

In the first model, the variables corresponding to the measures of popularity of the first 3 actors and the complete casting, have very high VIFs, indicating a high correlation between these variables.

In the second model, the casting variable has been suppressed, causing the multicollinearity produced by the presence of these variables to have disappeared. If the purpose of the model construction is to relate the collection to the popularity of the actors, with explanatory intention, it is important that the model does not have multicollinearity, and therefore, the second model would be used.

If we interpret the first model directly, we would say that by increasing the popularity of the first actor, the collection would be significantly reduced (!) Since the coefficient is negative and significant (-6.606e-05). What is not taken into account is that when increasing this variable and due to its correlation with the one of the casting, also it would increase this compensating the decrease in the collection. The second model clearly states that in reality the popularity of the first actor, despite having a negative sign, is not significant.

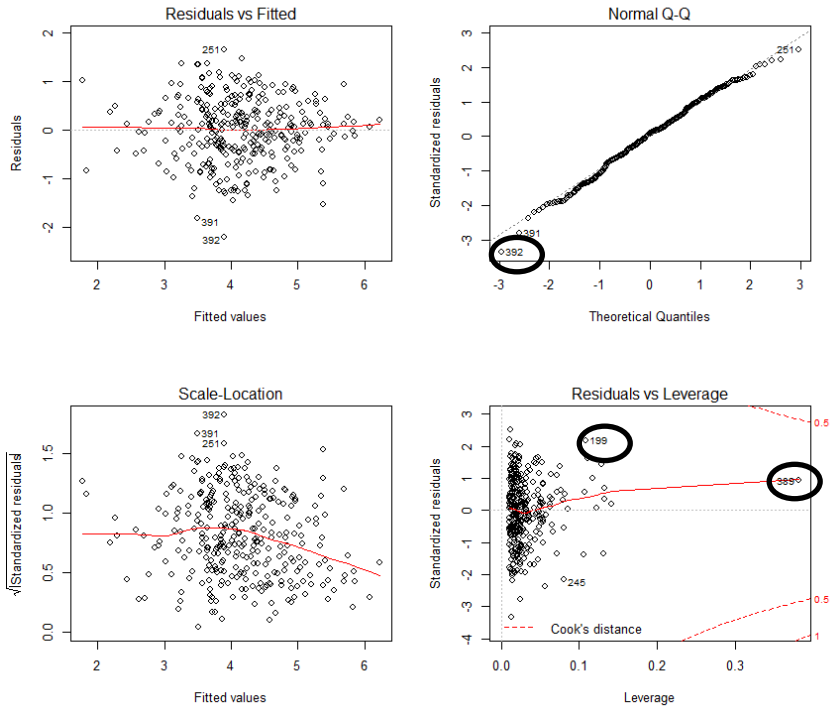
The predictor variables included in Model_6 are not significant. The popularity of the actors does not seem to have a significant relationship with the film's revenue.

The popularity of the actors does not seem to have a significant relationship with the film's revenue. Even so, to assert that none of them is significant presupposes that the elimination of each will not change the significance of those that remain in the model. If a sequential elimination procedure is performed, changing the residual standard deviation will change the standard errors of the coefficients, which in some cases is close to the significance level.

Results for a stepwise procedure monitored by BIC are included. The resulting model is Model_7:

```
Model_7
lm(formula = log(gross) ~ log(budget) + duration + Genre + actor3_fl +
    log(budget):Genre, data = imdb)
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              9.944e-01  4.014e-01   2.478 0.013758 *
log(budget)              6.209e-01  9.140e-02   6.793 5.58e-11 ***
duration                 8.717e-03  2.404e-03   3.626 0.000336 ***
GenreDrama               8.574e-02  3.926e-01   0.218 0.827267
GenreAction             -2.413e+00  9.414e-01  -2.563 0.010833 *
GenreHorror              1.625e+00  4.277e-01   3.799 0.000175 ***
actor3_fl                3.600e-05  1.312e-05   2.744 0.006428 **
log(budget):GenreDrama  -1.265e-01  1.133e-01  -1.117 0.264719
log(budget):GenreAction  3.898e-01  1.996e-01   1.954 0.051640 .
log(budget):GenreHorror -5.486e-01  1.292e-01  -4.247 2.87e-05 ***
```

```
Residual standard error: 0.6619 on 312 degrees of freedom
Multiple R-squared:  0.549,     Adjusted R-squared:  0.536
F-statistic: 42.19 on 9 and 312 DF,  p-value: < 2.2e-16
```



9. Perform the validation of the model, indicating in each graph the premises that are analyzed.

The first plot is that of the residuals against the predictions, it allows to see if the disposition of the residuals is random around the zero, without observing any pattern that indicate deviations of the linear relation. The local adjustment (red line) is practically horizontal, confirming in this case there do not seem to be any nonlinearity patterns. In this plot can also be verified descriptively if the variance can be considered constant, against the predictions. In this case, there is no increase in the variability of the residues as the prediction increases, indicating that homoscedasticity can be assumed. Also in this plot, the observations are labeled with standardized residuals greater than 2 (approx) in absolute value (atypical values).

The second plot is the plot of normality, which allows us to determine if we can consider that the Normal distribution is adequate for the residues. If the points are aligned we can assume Normality of the residues. This plot would allow to see patterns of asymmetry or heavy queues in the residues that would go against the normality hypothesis. Also atypical are labeled. In this case, the arrangement of the points is clearly aligned allowing normality to be assumed in the residuals.

The third plot represents the square root of the absolute values of the residuals versus the predictions. It is a plot that allows to determine more clearly the presence of heteroscedasticity. The local adjustment by the straight line does not indicate a clear descent of the values that constitute an estimate of the variance of the residues. It is not conclusive to confirm the presence of non-constant variance and can also be influenced by the low presence of observations that are related to high values of the predictions, which may imply a worse estimation of the variability.

The fourth graph allows identifying and characterizing influential data.
Represents the standardized residuals versus the anchor / leverage factor. It also includes contours to indicate Cook's distance from observations. Values with a high Cook distance can be influential values and their effect on model setting must be analyzed. Cook's distance is a growing function of square and leverage residuals. Observations that have a high value of Cook's distance are labeled (they may be very leverage, or have a high residue in absolute value or a combination of both not so extreme situations). Observations labeled as influential seem to have a high leverage and at the same time have a residuals of high magnitude. One should analyze what effect they have on model estimation.

10. Specify for observations 199, 389 and 392 that are indicated in the graphs whether atypical and / or influential data are.

Observation 199 has a standardized residual slightly higher than 2 that does not allow to characterize it as atypical. Nor does it present a very high leverage (although it is one of those that has it higher). The combination of these two factors makes it closer to Cook's distance curves, making it one of the most influential observations on the model.
Observation 389 is the observation with the highest leverage, well above the rest, making it an a priori influential observation. However, its standardized residual is close to 1, so the observation is adequately explained by the model. Even so, it would be necessary to check his distance from Cook to decide whether he is also influential a posteriori.
Observation 392 has a standardized residue of less than -3 being the most extreme value of the residual. It would be an atypical data. Its leverage is very small, indicating that it is close to the center of gravity of individuals who describe the design matrix. Because of this, it is most probably not an influential piece of information.

11. **Those films in the list included in 100 films with large revenues in the history are marked in a binary factor** A table relating the super-revenue binary target and the Genre is included below. Determine a binary logit model for super-revenue films that states that the probability of super-revenue is the same for all genre-categories.

```
> table(imdb$Genre,imdb$superevenue)
         Regular SuperRevenue
  Comedy     152            5
  Drama      143            1
  Action      66           30
  Horror      73            0
```

Globally the probability of being a super-revenue film is 36/470=0.077, 36 into the total number of observations in the sample, 470.

```
> 36/470
[1] 0.07659574
> 36/434
[1] 0.08294931
> log(36/434)
[1] -2.489526
> m0<-glm(superevenue~1,family=binomial,data=imdb)
> summary(m0)

Call: glm(formula = superevenue ~ 1, family = binomial, data = imdb)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3992  -0.3992  -0.3992  -0.3992   2.2668

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.4895     0.1734  -14.35   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 254.15  on 469  degrees of freedom
Residual deviance: 254.15  on 469  degrees of freedom
AIC: 256.15
```

12. Calculate the logit model for a positive response being a super-revenue film according to Gemre.

You have to calculate the model that states that the probability of being a large revenue film depends on the Genre level. Pay attention to the horror genre class that has 0 super-revenue films and instead of a 0 you have to use any small positive number to calculate de log-odds otherwise you are not allowed to get any finite number.

```
> log(5/152)
[1] -3.414443
> log(1/143)-log(5/152)
[1] -1.548402
> log(30/66)-log(5/152)
[1] 2.625985
> log(0.0000001/73)-log(5/152)
[1] -16.99411
```

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=\eta+\alpha_i \qquad \alpha_{1\equiv Comedy}=0 \qquad i=1\div 4$$

Student Name:
DNI or PASSPORT:


```
> m1<-glm(superevenue~Genre,family=binomial,data=imdb)
> summary(m1)

Call: glm(formula = superevenue ~ Genre, family = binomial, data = imdb)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8657  -0.2544  -0.1181  -0.1181   3.1527

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.4144     0.4545  -7.512 5.81e-14 ***
GenreDrama    -1.5484     1.1016  -1.406     0.16
GenreAction    2.6260     0.5050   5.200 2.00e-07 ***
GenreHorror  -16.1516  1258.6621  -0.013     0.99
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 254.15  on 469  degrees of freedom
Residual deviance: 175.49  on 466  degrees of freedom
AIC: 183.49
```

13. Which is the deviance of Model calculated in point 12.

The deviance of the saturated model, because this is the model calculated in Point 12 is 0.