GRAU INTERUNIVERSITARI D'ESTADÍSTICA I INVESTIGACIÓ OPERATIVA (UB- UPC)

CURS 2016-2017 Q1 – EXAMEN PARCIAL : MODEL LINEAL GENERALITZAT

(Data: 14 d'Octubre del 2016 a les 15:00 Aula S02-FME)

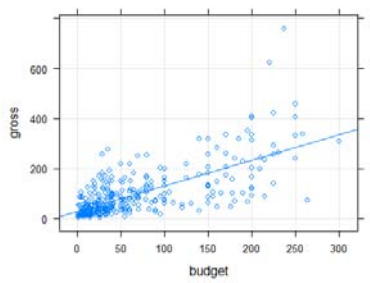
Professors:	Lídia Montero – Josep Anton Sanchez
Localització:	ETSEIB 6a Planta 6-67
Normativa de l'examen:	ÉS POT DUR APUNTS TEORIA *SENSE ANOTACIONS*, CALCULADORA I TAULES ESTADÍSTIQUES
Durada de l'examen:	2h 00 min
Sortida de notes:	Abans del 27 d'Octubre al Web Docent de MLGz
Revisió de l'examen:	27 d'octubre a les 15 h al despatx 6-67 (ETSEIB 6a. Planta)

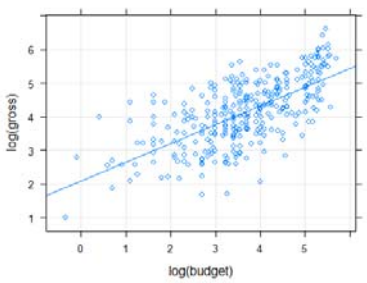
El conjunto de datos contiene información de 322 películas de cine de USA de la última década. Los datos se han recogido de la web www.imdb.com e incluyen la siguiente información:

```
movie_title:    Título de la película
gross:          Recaudación total (millones de $)
budget:         Presupuesto   (millones de $)
duration:       Duración (minutos)
title_year:     Año de estreno
actor1_fl:      Popularidad del primer actor (número de "Likes" en Facebook)
actor2_fl:      Popularidad del segundo actor (número de "Likes" en Facebook)
actor3_fl:      Popularidad del tercer actor (número de "Likes" en Facebook)
cast_fl:        Popularidad total del casting (número de "Likes" en Facebook)
faces_poster:   Número de caras que aparecen en el póster
Genre:          Género de la película (Comedy, Drama, Action, Horror)
```

Se pretende explorar las relaciones entre las variables recogidas y la recaudación de la película. Para las variables de tipo factor, el contraste activo es de tipo baseline con la primera categoría como referencia.

En primer lugar se trata de analizar la relación lineal que hay entre el Presupuesto y la Recaudación de la película. Se plantean dos regresiones lineales simples, una entre ambas variables y otra entre los logaritmos de las mismas:





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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

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
```

1) **(1p) Comenta las ventajas e inconvenientes de continuar la modelización de cada planteamiento. ¿Cuál consideras más adecuado? ¿Se puede usar el coeficiente de determinación (R2) o el coeficiente de determinación ajustado (R2-Adj) para decidir qué modelo es mejor? Razona la respuesta.**

El modelo con las variables sin transformar permite una interpretación más sencilla. El coeficiente del predictor se puede interpretar directamente como la variación en términos absolutos de la respuesta por cada incremento en un millón de dólares del presupuesto de la película. Sin embargo, el gráfico pone de manifiesto en claro incremento de la varianza de los residuos, al aumentar la variable explicativa, y por consiguiente las predicciones. Esto invalidaría el modelo y, en caso de no aplicar las transformaciones, obligaría a ajustar un modelo con heterocedasticidad (GLS) en lugar del modelo OLS.

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

Por otro lado, el modelo ajustado con ambas variables transformadas presenta mejores características para validar las premisas, ya que la varianza condicional parece más constante. Sin embargo, la interpretación es más complicada. El coeficiente adquiere el papel de una potencia del presupuesto.

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

A falta de comprobar otras premisas para las que se necesitan otros gráficos/tests (como la normalidad de los residuos) el segundo modelo sería el que presenta una mejor validación. La comparación de dos modelos con diferente variable respuesta no se puede hacer ni con la R2 ni con la R2-adj, ya que no son comparables desde el momento en que la variabilidad de la respuesta se mide incluso en diferentes unidades (el primero en millones de dólares y el segundo en logaritmo de millones de dólares).

2) **(1p) ¿Qué incremento de recaudación se espera (predicción puntual) si en una película cuyo presupuesto era de 100 millones, éste se incrementa en un 20%? Haz la predicción para ambos modelos.**

En el primer caso:

E(Gross|Budget=100)=28.5+1.025*100=131
E(Gross|Budget=120)=28.5+1.025*120=151.5

Incremento=151.5-131=20.5 millones de dólares

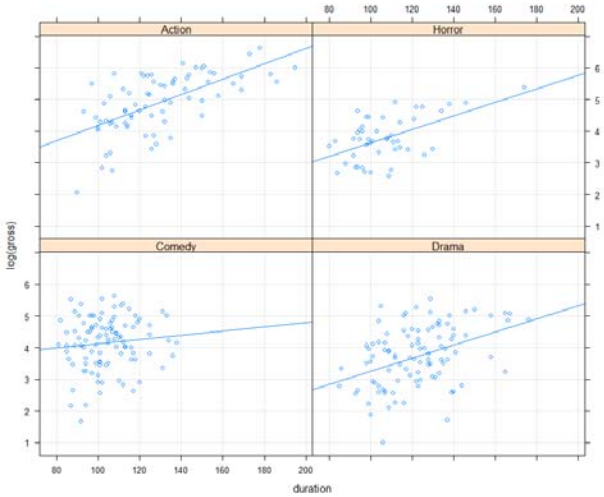
Para el segundo modelo:

E(Gross|Budget=100)=exp(2.1)*100^0.558=106.6
E(Gross|Budget=120)=exp(2.1)*120^0.558=118.1

Incremento=118.1-106.6=11.5 millones de dólares

(en este último cálculo, si se asume el modelo log-normal se debería incluir el efecto de la varianza residual en el cálculo del valor esperado)

Independientemente de la respuesta del primer apartado, se decide trabajar con la escala logarítmica de ambas variables. A continuación se desea ver si la relación entre la recaudación y la duración es la misma para todos los géneros.



Se estiman los siguientes modelos, ajustando también por el logaritmo del presupuesto:

Model_3
`lm(formula = log(gross) ~ log(budget) + duration + Genre, data = imdb)`
 	
`Residuals:`
	Model_4
	`lm(formula = log(gross) ~ log(budget) + duration * Genre, data = imdb)`
	
	`Residuals:`

```
     Min       1Q   Median       3Q      Max
-2.25155 -0.44033  0.04854  0.54310  1.63000

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
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

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
```

```
     Min       1Q   Median       3Q      Max
-2.18509 -0.45038  0.04113  0.50016  1.65504

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
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

Comparamos ambos modelos con el método anova.

```
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
```

3) (1.5p) Completa los valores --(1)-- a --(6)-- que aparecen en la tabla anterior, indicando como se ha realizado el cálculo para cada valor. ¿Es significativa la interacción entre la duración de la película y el género de la misma? Interpreta este resultado

(1) Grados de libertad del modelo 1: (N-p)=322-6=316
(2) Suma de cuadrados residual del modelo 1: RSS1=(N-p)S_1^2=316*0.7015^2=155.5043
(3) Suma de cuadrados residual del modelo 2: RSS2=(N-p)S_2^2=313*0.6936^2=150.5783
(4) Grados de libertad del test de devianza: df=df1-df2=316-313=3
(5) Suma de cuadrados del test de devianza: RS=RSS1-RSS2=155.5043-10.5783=4.9260
(6) Estadístico F del test de devianza: F=((RSS1-RSS2)/q)/(RSS2/df2)=(4.926/3)/(150.5783/313)=3.41

El p-valor del test es 0.018 que está por debajo del nivel de significación del 5%, lo cual indica que hay evidencias estadísticas significativas que establecen que el segundo modelo es diferente del primero y que por lo tanto la interacción en el modelo 2 es significativa. Quiere decir que el efecto en la recaudación en función de la duración de la película depende del género de la misma. La representación de los modelos segmentado por género apuntan a que la pendiente de la recta ajustada en cada gráfico se puede considerar que no es la misma en los diferentes casos.

4) (2p) Para el modelo Model_4 escribe los modelos estimados que relacionan la duración con la recaudación para cada género, incluyendo el ajuste por el logaritmo del presupuesto. Para estos modelos, indica si hay diferencias significativas de esta relación comparando las Comedias con cada uno del resto de géneros. ¿De qué tipo son estas diferencias?

Para Genre=Comedy
Gross= 2.148 + 0.00260 *duration +0.4878*log(Budget)
Para Genre=Drama
Gross=(2.148-0.640) + (0.00260+0.00313)*duration +0.4878*log(Budget)
Gross=1.508 + 0.00573*duration +0.4878*log(Budget)
Para Genre=Action
Gross=(2.148-2.119) + (0.00260+0.01602)*duration +0.4878*log(Budget)
Gross=0.029 + 0.01861*duration +0.4878*log(Budget)
Para Genre=Horror
Gross=(2.148+0.210)+ (0.00260+0.00184)*duration +0.4878*log(Budget)
Gross=2.358+ 0.00444*duration +0.4878*log(Budget)

Teniendo en cuenta que el contraste activo es de tipo Baseline con la primera categoría como referencia (“contr.treatment”) los coeficientes de la variable categórica Genre se interpretan como el cambio en el coeficiente de la ordenada en el origen entre una

películo del género comedia y cada uno de los géneros restantes. De la misma manera, la interacción se interpreta en término de cambio en la pendiente de los diferentes géneros respecto a las comedias. De todos los p-valores que aparecen asociados a estos dos términos, sólo son significativos el del intercept y de la pendiente asociados a las películas de acción. Así pues, en el caso de las películas de acción, la diferencia en el intercept respecto a una comedia es de -2.119 significativa, y en la pendiente hay un incremento significativo de 0.016. Para el resto de géneros no se ha establecido la significación de la diferencia en ningún parámetro del modelo respecto al de las comedias.

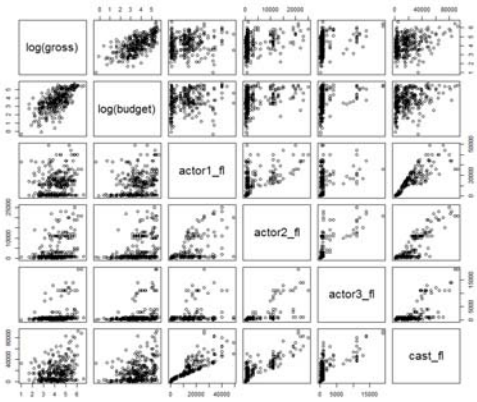
Para determinar si en las películas del género Drama la duración influye en la recaudación de forma significativa, usando el Model_4, se realizan los siguientes contrastes:

```
Linear hypothesis test
Hypothesis:
GenreDrama = 0
Model 1: restricted model
Model 2: log(gross) ~ log(budget) + duration * Genre
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    314 150.91
2    313 150.57  1   0.34071 0.7082 0.4007
-------------------------------------------------------
Linear hypothesis test
Hypothesis:
duration:GenreDrama = 0
Model 1: restricted model
Model 2: log(gross) ~ log(budget) + duration * Genre
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    314 150.68
2    313 150.57  1   0.10366 0.2155 0.6428
-------------------------------------------------------
Linear hypothesis test
Hypothesis:
duration  + duration:GenreDrama = 0
Model 1: restricted model
Model 2: log(gross) ~ log(budget) + duration * Genre
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    314 151.51
2    313 150.57  1   0.94044 1.9549  0.163
```

5) **(1p) ¿Es significativa la relación entre la duración y la recaudación si la película es un Drama? Justifica la respuesta indicando el p-valor del test en que basas la decisión.**

No es significativa (p-valor=0.163). Se debe utilizar el tercer test, ya que el test que se pretende decidir es el que pretende establecer la significación de la pendiente del modelo para un Drama. El coeficiente de esta pendiente a partir del modelo ajustado se obtiene sumando el coeficiente de la variable duration (pendiente del modelo para una comedia) al coeficiente de la interacción entre la duración y el género correspondiente, en este caso Drama.

Analizamos a continuación la influencia de la popularidad de los actores en la recaudación.



Se ajusta los siguientes modelos:

Model_5	**Model_6**
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	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)  2.174e+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 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
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(m1)
log(budget)   actor1_fl   actor2_fl   actor3_fl     cast_fl
   1.151489   47.399810   14.493582   10.336643  122.549023
```

```
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
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

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(m2)
log(budget)   actor1_fl   actor2_fl   actor3_fl
   1.139089    1.569572    2.849449    2.061186
```

6) **(1p) ¿Qué representan los estadísticos VIF que aparecen en el listado? (no hace falta dar la expresión matemática). Si se desea establecer la relación entre las variables explicativas y la respuesta a partir de los coeficientes del modelo, ¿Qué modelo usarías para hacer la interpretación? Justifica la respuesta.**

Los VIFs (Variance Inflation Factor) o factores de inflamiento de la varianza son estadísticos que permiten diagnosticar la presencia de multicolinealidad en el modelo. Este fenómeno implica que las variables predictoras puedan presentar correlación alta que impida una interpretación simple de cada coeficiente del modelo, manteniendo el resto de predictores fijo (ceteris paribus). Así mismo, la estimación en presencia de multicolinealidad deja de ser eficiente. Para cada coeficiente, el VIF calculado es una medida de cuan correlacionada está ese predictor con el resto de predictores (Si R^2 es el coeficiente de determinación en tanto por 1 del modelo que se ajusta con cada variable como si fuera la respuesta en base al resto de predictores, entonces $VIF=1/(1-R^2)$). En el caso en que los predictores sean independientes (y no haya multicolinealidad) el valor de los VIFs será 1. Cuanto más se aleje de 1 indica una mayor presencia de multicolinealidad. Un VIF superior aproximadamente a 8 indica que el problema de multicolinealidad puede ser importante.

En el primer modelo, las variables correspondientes a las medidas de popularidad de los 3 primeros actores y el casting completo, poseen VIFs muy altos, indicando una alta correlación entre estas variables. Esto también se aprecia en el matrix-plot, puesto que considerar el número de likes en Facebook del total del casting es lógico que se halle altamente correlacionado con los likes de cada uno de los tres primeros componentes.

En el segundo modelo, se ha suprimido la variable del casting, haciendo que la multicolinealidad producida por la presencia de estas variables haya desaparecido. Si el objetivo de la construcción del modelo es relacionar la recaudación con la popularidad de los actores, con intención explicativa, es importante que el modelo no tenga multicolinealidad, y por lo tanto, se utilizaría el segundo modelo.

Si interpretamos directamente el primer modelo, diríamos que al aumentar la popularidad del primer actor, disminuiría de forma significativa la recaudación (!) ya que el coeficiente es negativo y significativo (-6.606e-05). Lo que no se tiene en cuenta es que al aumentar esta variable y debido a su correlación con la del casting, también aumentaría ésta compensando el descenso en la recaudación. El segundo modelo establece de forma clara que en realidad la popularidad del primer actor, pese a tener signo negativo, no es significativa.

7) **(1p) Según el modelo Model_6 ¿Hay relación significativa entre la popularidad de los actores y la recaudación? ¿Es correcto interpretar simultáneamente los p-valores de las 3 variables de popularidad?**

Las variables predictoras incluidas en el modelo 6 aparecen como no significativas. Esto indicaría que la popularidad de los actores no parece tener una relación significativa con la recaudación de la película. Aun así, afirmar que ninguna de ellas sea significativa presupone que la eliminación de cada una no hará cambiar la significación de las que queden en el modelo. Si se realiza un procedimiento de eliminación secuencial, al cambiar la desviación estándar residual irán cambiando los errores estándar de los coeficientes, que en algún caso es próximo al nivel de significación.

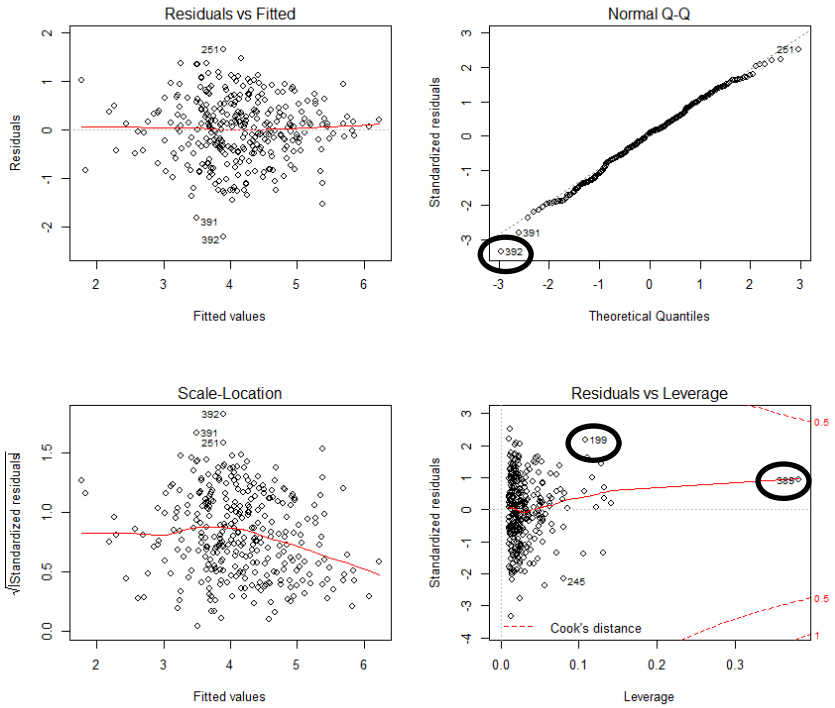
Finalmente, se utiliza el procedimiento stepwise con criterio BIC para seleccionar las variables significativas:

```
Model_7
lm(formula = log(gross) ~ log(budget) + duration + Genre + actor3_fl +
    log(budget):Genre, data = imdb)
Residuals:
     Min       1Q   Median       3Q      Max
-2.19765 -0.40614  0.05207  0.42752  1.65329
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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
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

> anova(m4)
Analysis of Variance Table
Response: log(gross)
                   Df  Sum Sq Mean Sq  F value    Pr(>F)
log(budget)         1 135.704 135.704 309.7586 < 2.2e-16 ***
duration            1   4.371   4.371   9.9762  0.001741 **
Genre               3   7.493   2.498   5.7013  0.000823 ***
actor3_fl           1   4.587   4.587  10.4693  0.001344 **
log(budget):Genre   3  14.211   4.737  10.8127 8.883e-07 ***
Residuals         312 136.686   0.438
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Los gráficos para hacer la validación del modelo son los siguientes:



8) **(1.5p) Realiza la validación del modelo, indicando en cada gráfico las premisas que se analizan. Caracteriza en base a si son datos atípicos y/o influyentes las observaciones 199, 389 y 392 que aparecen señaladas en los gráfico.**

El primer plot es el de los residuos frente las predicciones, permite ver si la disposición de los residuos es aleatoria alrededor del cero, sin que se observe ningún patrón que indicas desviaciones de la relación lineal. El ajuste local (línea roja) es prácticamente horizontal, confirmando en este caso no parece haber patrones de no linealidad. En este plot también se puede verificar descriptivamente si la varianza puede considerarse constante, frente a las predicciones. En este caso, no se observa incremento de la variabilidad de los residuos a medida que aumenta la predicción, indicando que se puede asumir homocedasticidad. También en este plot, aparecen etiquetadas las observaciones con residuos estandarizados superior a 2 (aprox) en valor absoluto (valores atípicos).

El segundo plot es el plot de normalidad, que permite determinar si podemos considerar que la distribución Normal es adecuada para los residuos. Si los puntos están alineados podemos asumir Normalidad de los residuos. Este plot permitiría ver patrones de asimetría o colas pesadas en los residuos que irían en contra de la hipótesis de normalidad. También se etiquetan los atípicos. En este caso, la disposición de los puntos está claramente alineada lo que permite asumir normalidad en los residuos.

El tercer plot representa la raíz cuadrada de los valores absolutos de los residuos frente a las predicciones. Es un plot que permite determinar de forma más clara la presencia de heteroscedasticidad. El ajuste local mediante la recta no indica un claro descenso de los valores que constituyen una estimación de la varianza de los residuos. No es concluyente para confirmar la presencia de varianza no constante y además se puede ver influido por la poca presencia de observaciones que estén relacionados con valores altos de las predicciones, lo cual puede suponer una peor estimación de la variabilidad.

El cuarto gráfico permite identificar y caracterizar los datos influyentes. Representa los residuos estandarizados frente al factor de anclaje/apalancamiento (leverage). Además incluye curvas de nivel para indicar la distancia de Cook de las observaciones. Valores con una distancia de Cook alta pueden ser valores influyentes y se debe analizar su efecto en el ajuste del modelo. La distancia de Cook es una función creciente de los residuos al cuadrado y del leverage. Las observaciones que tienen un valor alto de la distancia de Cook aparecen etiquetadas (pueden ser por tener muy leverage, o tener un residuos alto en valor absoluto o una combinación de ambas situaciones no tan extremas). Las observaciones etiquetadas como influyentes parece que tienen un leverage alto ya la vez tienen un residuo de magnitud elevada. Habría que analizar qué efecto tienen en la estimación del modelo.

La observación 199 tiene un residuo estandarizado ligeramente superior a 2 que no permite caracterizarlo como atípico. Tampoco presenta un leverage muy alto (aunque es de los que lo tiene mayor). La combinación de estos dos factores hace que se sitúe más cerca de las curvas de nivel de la distancia de Cook, haciéndola una de las observaciones que puedan influir más en el modelo.
La observación 389 es la observación con el leverage más alto, muy por encima del resto, haciendo que sea una observación influyente a priori. Sin embargo, su residuo estandarizado es próximo a 1, por lo que la observación está explicada de forma adecuada por el modelo. Aun así, sería necesario comprobar su distancia de Cook para decidir si es también influyente a posteriori.
La observación 392 tiene un residuo estandarizado inferior a -3, siendo el valor más extremo de los residuos. Sería un dato atípico. Su leverage es muy pequeño, indicando que se encuentra próximo al centro de gravedad de los individuos que describen la matriz de diseño. Debido a ello, muy probablemente no se trata de un dato influyente.