Universidad De La Habana

MATCOM

Proyecto de Estadística

Fase 2

Autores:
Olivia González Peña C411
Juan Carlos Casteleiro Wong C411
Juan Carlos Vázquez García C412

Introducción

Nuestro trabajo se centra en el análisis e interpretación de datos a partir de la aplicación de técnicas de regresión, reducción de dimensión y análisis de varianza; para ello se nutre del archivo "IMBD.csv" (que deberá situarse en la misma carpeta que los códigos que brindamos como solución). El mismo hace referencia a una selección de las "mejor valoradas" 118 películas en inglés en el período comprendido entre los años 2010 y 2016, e incluye en su información datos sobre las mismas tales como nombre, presupuesto destinado, total de votos obtenidos para la valoración, puntuación alcanzada (rating), duración (en minutos) de la película, entre otros (55 en total). Para hallar los resultados tras aplicar cada técnica, dedicamos a cada una nuestra solución programada en el lenguaje de programación R, recurriendo para solucionar ciertos subproblemas (sobre todo en el pre procesamiento de los datos) a la fase anterior del proyecto. Los tres integrantes del equipo debatimos la forma en la que se debía afrontar cada problemática, dejando a cada uno la escritura del código de alguna de estas (regresión lineal simple y múltiple: Olivia, reducción de dimensión y clustering: Juan Carlos Casteleiro, ANOVA: Juan Carlos Vázquez). Asimismo, cada contratiempo o duda que se presentó la compartimos entre todos.

Regresión Lineal Simple

Para aplicar la técnica de la regresión lineal simple a nuestro set de datos, escogimos inicialmente las variables Rating y Presupuesto, debido al interés que despertaba en nosotros esta posible asociación; pero tras realizar el test correlación, el coeficiente resultó demostrar que, al menos para los datos recogidos en este estudio, las variables no están correlacionadas, obteniéndose que r = 0.26. Debido a esto, decidimos escoger las variables Rating y VotesUS (indica el rating obtenido por la película solo contando las votaciones dentro de Estados Unidos) y aplicarles el test de correlación, obteniendo un coeficiente de correlación r = 0.8377. Podemos afirmar que existe correlación lineal fuerte y positiva, por lo que pasaremos a generar un modelo de regresión lineal simple que permita predecir el Rating obtenido por una película en función de su rating en Estados Unidos.

A continuación, los detalles del modelo propuesto:

```
>
> # Modelo de regresión lineal simple
> linear_regression <- lm(rating ~ votesus, data = my_data)
> summary(linear_regression)

Call:
lm(formula = rating ~ votesus, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25721 -0.07591 -0.00843  0.09157  0.34279

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.22544    0.40250   3.045  0.00288 **
votesus      0.83741    0.05068  16.523  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1298 on 116 degrees of freedom
Multiple R-squared:  0.7018,    Adjusted R-squared:  0.6992
F-statistic:   273 on 1 and 116 DF,  p-value: < 2.2e-16
```

Residuals muestra el error entre la predicción del modelo y los resultados reales, obteniendo valores bastante pequeños y, por tanto, positivos.

Analizando Coefficients, podemos ver que los coeficientes del intercepto y de la variable VotesUS son significativos al 0%, lo cual es también un resultado bueno. Aunque tiene sentido pensar que en el Rating definitivo de una película influyen como para predecirlo tanto el rating dentro de Estados Unidos como fuera de este, estos valores nos indican que se está aportando bastante al modelo

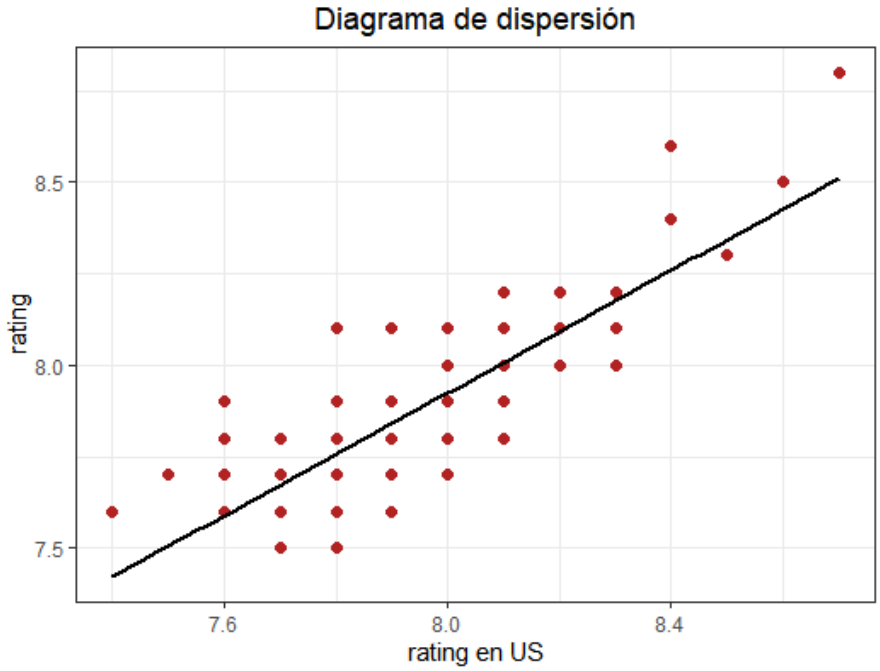
con esta selección por lo que parece ser más explicativo la variable escogida para este modelo.

El valor del Adjusted R-Square nos dice que este modelo explica el 70% de la variación. Aunque reconocemos que el valor para el indicador no es realmente bueno, decidimos quedarnos con este modelo ya que, tras realizar múltiples pruebas para distintos pares de variables de la tabla, el modelo que resultaba válido según los supuestos, y con mejores indicadores es el presentado en este trabajo.

Finalmente, al ser el p-value del F-Statistic mucho menor que 0.05, tiende a valorar que nuestro modelo funciona, y queda de la siguiente forma:

$$Rating = 1.225 + 0.837 * VotesUS \tag{1}$$

Correspondidos con el siguiente diagrama de dispersión:



Pasamos entonces a analizar los residuos para comprobar que se cumplen todos los supuestos del modelo, comenzando por comprobar que los errores son independientes, para lo cual realizamos la prueba de Durbin-Watson:

```
>
> dwtest(linear_regression)

        Durbin-Watson test

data:  linear_regression
DW = 2.015, p-value = 0.5375
alternative hypothesis: true autocorrelation is greater than 0
```

Como el valor del p-value es mucho mayor que el valor de significación con el que estamos trabajando que es 0.05, no se rechaza la hipótesis nula por lo que podemos afirmar que los errores son independientes.

Para revisar que se cumpla que la varianza del error aleatorio es constante, que corresponde al supuesto de homocedasticidad, nos apoyamos en la prueba de Breusch-Pagan:

```
>
> bptest(linear_regression)

        studentized Breusch-Pagan test

data:  linear_regression
BP = 0.0015452, df = 1, p-value = 0.9686
```

Sucede lo mismo con el p-value de esta prueba, es mucho mayor que 0.05 por lo que no podemos rechazar la hipótesis nula y podemos afirmar que se cumple la heterocedasticidad.

Para verificar que los errores, además, son idénticamente distribuidos y siguen distribución normal con media cero y varianza constante realizamos la prueba de Shapiro-Wilk:

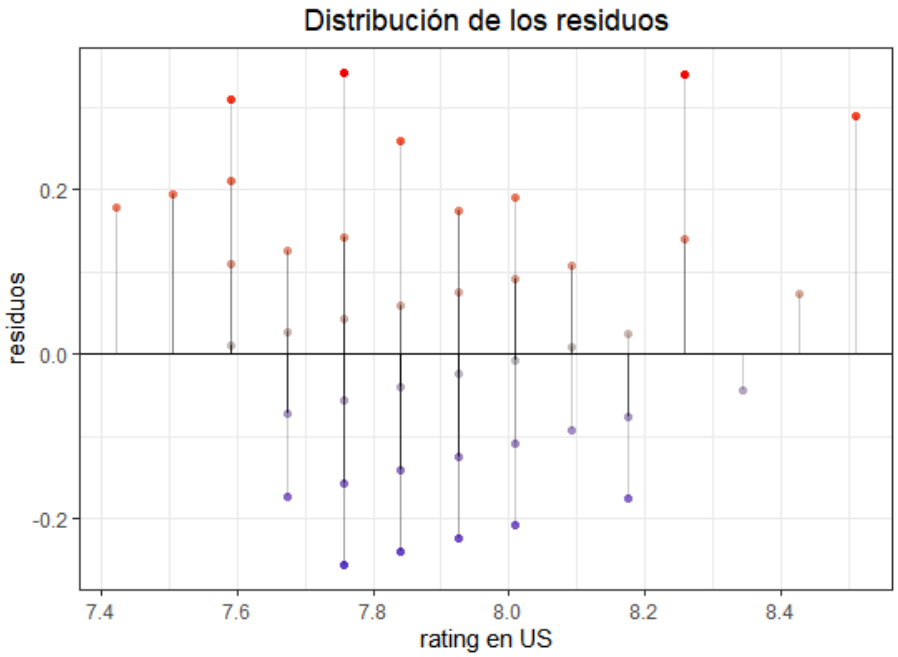
```
>
> shapiro.test(linear_regression$residuals)

        Shapiro-Wilk normality test

data:  linear_regression$residuals
W = 0.98245, p-value = 0.127
```

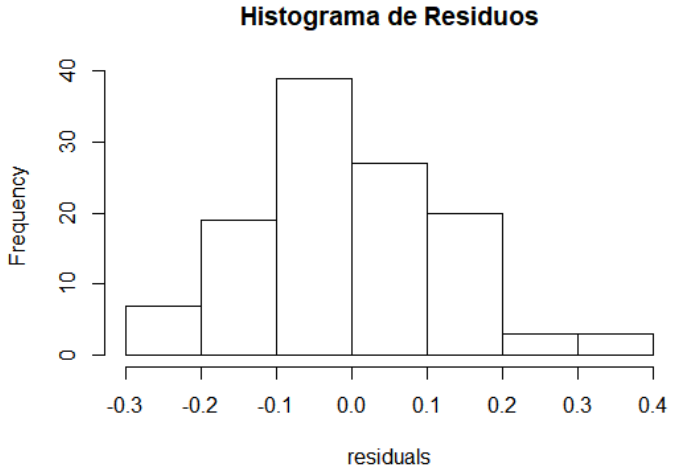
En este caso, el p-value también es mayor que el nivel de significación y, por tanto, no se puede rechazar la hipótesis nula por lo que los errores siguen una distribución normal.

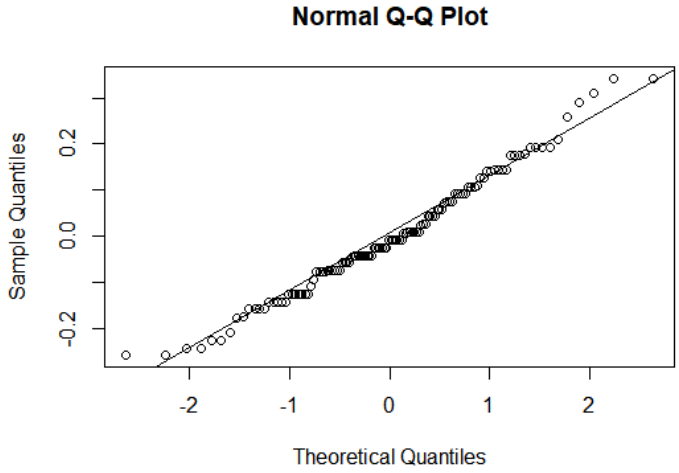
Además, la media y la suma de los errores es cero:

```
> mean(residuals)
[1] 2.585661e-18
> sum(residuals)
[1] 3.053113e-16
```

Por tanto, al cumplirse todos los supuestos, se valida el modelo propuesto. Se anexan los gráficos de los residuos que constatan de forma visual estos resultados.




Histograma de Residuos
Frequency
0
10
20
30
40
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
residuals


Normal Q-Q Plot
Sample Quantiles
-0.2
0.0
0.2
-2
-1
0
1
2
Theoretical Quantiles

Regresión Lineal Múltiple

Para la aplicación de la técnica de regresión lineal múltiple, intentamos establecer una relación entre variables cuantitativas que nos parecían interesantes para una predicción del rating de una película, siempre partiendo de la matriz de correlación para escoger variables que estuvieran correlacionadas con este, y que su a su vez fueran no correlacionadas entre sí.

De las distintas formas que estudiamos para llegar al modelo final más adecuado, realizamos una mezcla ayudándonos de la información que nos brinda la matriz de correlación para escoger las variables iniciales y finalmente, con estas, aplicamos el método Backward. Llegamos a un modelo donde podemos predecir el rating final de una película a partir de la información del rating que le da la población de votantes masculinos, el rating que le da la población de votantes femeninas y el rating que obtiene fuera de los Estados Unidos. A continuación, sus detalles:

```
> # Generamos el modelo de regresión múltiple
> multi_regression <- lm(rating ~ votesm + votesf
+               + votesnus, data = my_data)
> summary(multi_regression)

Call:
lm(formula = rating ~ votesm + votesf + votesnus, data = my_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.096962 -0.025226 -0.000346  0.027033  0.083830

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.29509    0.12238   2.411  0.01751 *
votesm       0.59026    0.05548  10.640  < 2e-16 ***
votesf       0.19207    0.01913  10.042  < 2e-16 ***
votesnus     0.18383    0.06203   2.964  0.00371 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04 on 113 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9721,    Adjusted R-squared:  0.9714
F-statistic:  1312 on 3 and 113 DF,  p-value: < 2.2e-16
```

Obteniendo el modelo siguiente:

$$Rating = 0.30 + VotesM * 0.59 + VotesF * 0.19 + VotesnUS * 0.18 \qquad (2)$$

Residuals muestra el error entre la predicción del modelo y los resultados reales, obteniendo valores bastante pequeños y, por tanto, positivos.

Analizando Coefficients, podemos ver que el coeficiente del intercepto es significativo al 1%, mientras que las tres variables escogidas mantienen una significación del 0%; se entienden estos resultados como bastante buenos, aunque no sean óptimos.

El valor del Adjusted R-Square nos dice que este modelo explica el 97% de la variación, lo cual nos parece un resultado muy bueno para la expresividad de este modelo que, unido a que el p-value del F-Statistic es mucho menor que el valor de significación con el que estamos trabajando que es 0.05 nos hace pensar que, de cumplirse los supuestos, nuestro modelo funciona. Pasemos a comprobar el cumplimiento de todos los supuestos para validarlo.

Comenzaremos por revisar que la media y la suma de los errores sea cero:

```
> residuals <- multi_regression$residuals
> mean(residuals)
[1] -1.270405e-19
> sum(residuals)
[1] -1.490778e-17
>
```

Al ver que esto se cumple, pasamos a verificar que los errores siguen una distribución normal con media cero y varianza constante y son idénticamente distribuidos, para lo cual realizamos la prueba de Shapiro-Wilk:

```
> shapiro.test(residuals)

        Shapiro-Wilk normality test

data:  residuals
W = 0.98946, p-value = 0.5066
```

Como el valor del p-value es mayor que 0.05 y, por tanto, no se puede rechazar la hipótesis nula los errores siguen una distribución normal.

Para comprobar que los errores son independientes realizamos la prueba de Durbin-Watson:

```
> dwtest(multi_regression)

        Durbin-Watson test

data:  multi_regression
DW = 2.0245, p-value = 0.5604
alternative hypothesis: true autocorrelation is greater than 0
```

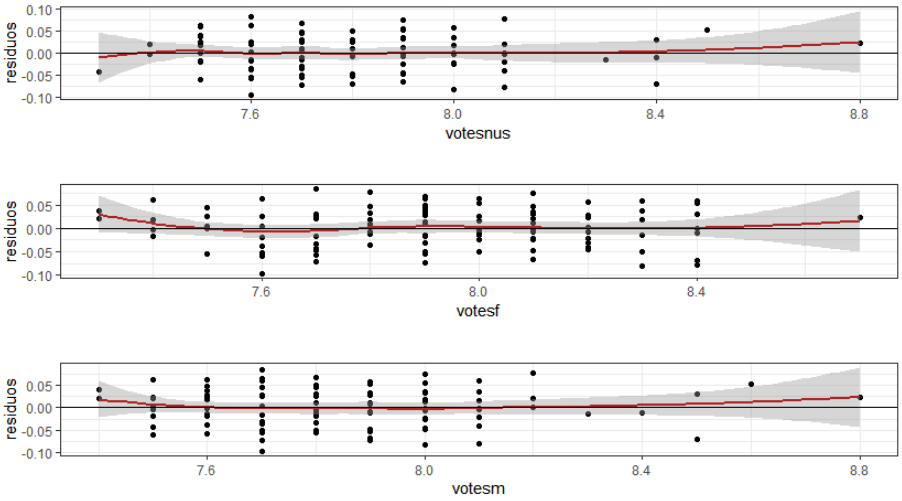
Al ser el p-value mayor que el valor de significación con el que estamos trabajando que es 0.05, no se rechaza la hipótesis nula por lo que podemos afirmar que los errores son independientes.

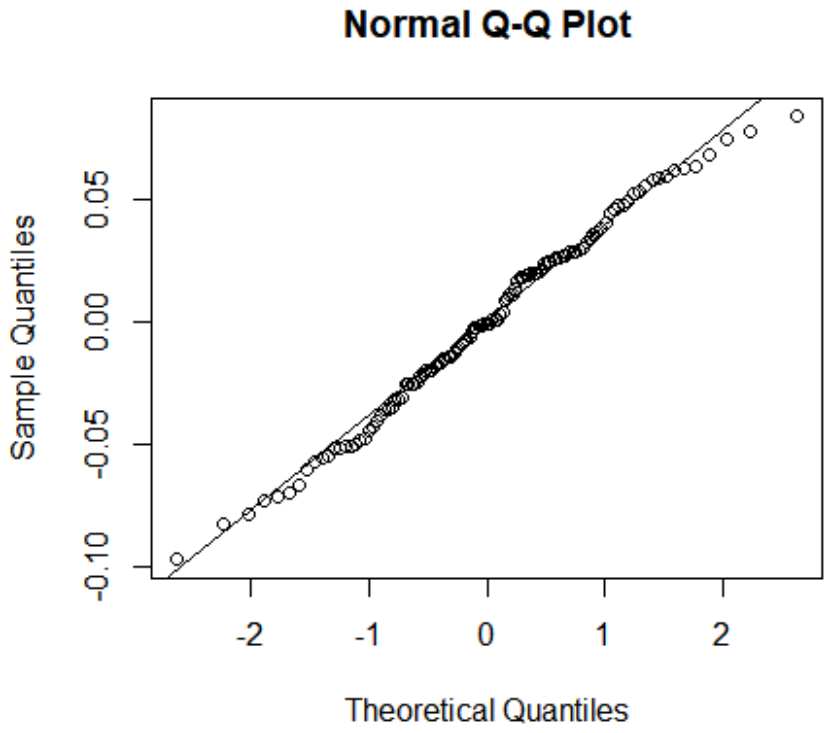
Finalmente, para revisar que se cumple el supuesto de homocedasticidad, nos apoyamos en la prueba de Breusch-Pagan:

```
> bptest(multi_regression)

        studentized Breusch-Pagan test

data:  multi_regression
BP = 0.13624, df = 3, p-value = 0.9872
```

Como el p-value de esta prueba es mucho mayor que 0.05 no podemos rechazar la hipótesis nula por lo que podemos afirmar que, también se mantiene el supuesto de homocedasticidad.

Por tanto, al cumplirse todos los supuestos, el modelo propuesto es válido. Se anexan los gráficos de los residuos que constatan de forma visual estos resultados.


residuos
0.10
0.05
0.00
-0.05
-0.10
7.6
8.0
8.4
8.8
votesnus
residuos
0.05
0.00
-0.05
-0.10
7.6
8.0
8.4
votesf
residuos
0.05
0.00
-0.05
-0.10
7.6
8.0
8.4
8.8
votesm


Normal Q-Q Plot
Sample Quantiles
0.05
0.00
-0.05
-0.10
-2
-1
0
1
2
Theoretical Quantiles

Reducción de dimensión

Una de las dificultades que hemos tenido hasta este punto del desarrollo del proyecto, es el trabajo con un alto volumen de información, que resulta tormentoso y sumamente incómodo. Para aliviar esta situación, podemos hacer un análisis de componentes principales.

Para ello, partiremos de la correlación entre las variables. Si bien la matriz de correlación es efectiva, suele ser incómodo tratar de ver si es una matriz altamente correlacionada o no a simple vista cuando tenemos tantos datos. Sin embargo, la forma gráfica de esta nos aclara un poco mejor la correlación entre las variables, pudiendo observarse mayoría de espacios en blanco y puntos, lo que nos lleva a pensar que no está altamente correlacionada. Podemos proseguir entonces al análisis de las componentes principales:

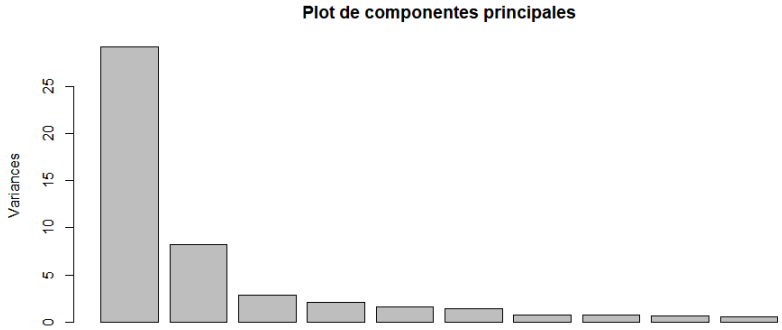
```
> # Análisis de componentes principales
> acp <- prcomp(my_data, scale = TRUE)
> summary(acp)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     5.4076 2.8630 1.68044 1.44997 1.26548 1.17435 0.87870 0.86574 0.78996 0.73590 0.63171 0.5832 0.47375 0.45719
Proportion of Variance 0.5848 0.1639 0.05648 0.04205 0.03203 0.02758 0.01544 0.01499 0.01248 0.01083 0.00798 0.0068 0.00449 0.00418
Cumulative Proportion  0.5848 0.7488 0.80525 0.84729 0.87932 0.90690 0.92235 0.93734 0.94982 0.96065 0.96863 0.9754 0.97992 0.98410
                          PC15    PC16    PC17    PC18    PC19    PC20    PC21    PC22    PC23    PC24    PC25    PC26    PC27    PC28
Standard deviation     0.37813 0.34353 0.31472 0.26155 0.24744 0.23547 0.20393 0.19095 0.1726 0.15079 0.13016 0.12661 0.11602 0.10423
Proportion of Variance 0.00286 0.00236 0.00198 0.00137 0.00122 0.00111 0.00083 0.00073 0.0006 0.00045 0.00034 0.00032 0.00027 0.00022
Cumulative Proportion  0.98696 0.98932 0.99130 0.99267 0.99390 0.99500 0.99584 0.99657 0.9972 0.99762 0.99796 0.99828 0.99854 0.99876
                          PC29    PC30    PC31    PC32    PC33    PC34    PC35    PC36    PC37    PC38    PC39    PC40    PC41    PC42
Standard deviation     0.09823 0.09518 0.09037 0.08317 0.07873 0.06957 0.06430 0.06164 0.06102 0.04729 0.03408 0.02711 0.02313 0.02013
Proportion of Variance 0.00019 0.00018 0.00016 0.00014 0.00012 0.00010 0.00008 0.00008 0.00007 0.00004 0.00002 0.00001 0.00001 0.00001
Cumulative Proportion  0.99896 0.99914 0.99930 0.99944 0.99956 0.99966 0.99974 0.99982 0.99989 0.99994 0.99996 0.99997 0.99999 0.99999
                          PC43    PC44    PC45    PC46    PC47    PC48    PC49    PC50
Standard deviation     0.01701 0.005787 0.002548 0.001163 0.0009531 0.0004552 0.000211 2.322e-05
Proportion of Variance 0.00001 0.000000 0.000000 0.000000 0.0000000 0.0000000 0.000000 0.000e+00
Cumulative Proportion  1.00000 1.000000 1.000000 1.000000 1.0000000 1.0000000 1.000000 1.000e+00
```

todas las componentes

```
> # Análisis de componentes principales
> acp <- prcomp(my_data, scale = TRUE)
> summary(acp)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     5.4076 2.8630 1.68044 1.44997 1.26548 1.17435 0.87870
Proportion of Variance 0.5848 0.1639 0.05648 0.04205 0.03203 0.02758 0.01544
Cumulative Proportion  0.5848 0.7488 0.80525 0.84729 0.87932 0.90690 0.92235
```

primeras 7 componentes

Para escoger las componentes principales tenemos dos criterios, el de Kaiser que propone quedarnos con aquellas cuyo valor propio sea mayor que uno; y el del porciento, que se remite a la proporción acumulativa y propone tomar las componentes que lleguen a explicar cierto porcentaje que nos parezca suficientemente expresivo, por ejemplo, un 80%. Según el primero de los criterios mencionados nos quedaríamos con las primeras seis componentes, mientas que, según el criterio del porciento, pudiéramos quedarnos con las primeras tres componentes. Esta última estrategia fue la que utilizamos, se muestra el histograma para visualizar de forma gráfica la significación de nuestra selección:



En consecuencia, la interpretación de los datos es la siguiente:

	PC1	PC2	PC3	PC4
Rating	-0.145229635	-0.2025275888	6.489877e-02	-0.072866723
TotalVotes	-0.179686698	0.0645377055	-1.922965e-02	0.005421849
MetaCritic	0.001131481	-0.0550137339	3.822235e-01	0.228869921
Budget	-0.095794271	0.0818184996	-2.023552e-01	-0.024604039
Runtime	-0.071514839	0.0015341806	-1.330675e-01	-0.136555176
i..CVotes10	-0.168833937	-0.0002455375	-8.962052e-03	-0.088488167
CVotes09	-0.179810750	-0.0007373688	1.203760e-02	-0.012677136
CVotes08	-0.166056558	0.0824098212	-5.336068e-02	0.104932197
CVotes07	-0.139638341	0.1819709311	-7.603922e-02	0.121940134
CVotes06	-0.133092201	0.2231345274	-2.923217e-02	0.031220309
CVotes05	-0.135740086	0.2153819819	3.265795e-02	-0.042008018
CVotes04	-0.135884419	0.1977794605	8.386257e-02	-0.091768567
CVotes03	-0.136319270	0.1758035196	1.219879e-01	-0.113848113
CVotes02	-0.137069345	0.1611899446	1.450681e-01	-0.126268153
CVotes01	-0.145851723	0.1242441332	7.918999e-02	-0.122439546
CVotesMale	-0.176244402	0.0767307677	-7.325704e-04	0.035784963
CVotesFemale	-0.164991533	0.0512796842	-7.709313e-02	-0.028183936
CVotesU18	-0.153418291	0.0150837442	-9.647493e-02	-0.163113100
CVotesU18M	-0.155121618	0.0140677623	-5.482569e-02	-0.119065471
CVotesU18F	-0.111547379	0.0136993531	-1.932384e-01	-0.251687816
CVotes1829	-0.178314068	0.0633590542	-1.474457e-02	-0.011863981
CVotes1829M	-0.176722152	0.0656389960	3.063692e-03	0.004266397
CVotes1829F	-0.160262340	0.0463988531	-7.746594e-02	-0.068720998
CVotes3044	-0.173513211	0.0916214774	7.122090e-04	0.075165422
CVotes3044M	-0.171630687	0.0951519025	9.101774e-03	0.078799875
CVotes3044F	-0.165749530	0.0638432559	-4.479208e-02	0.048790477
CVotes45A	-0.170374350	0.0845309959	-3.398249e-02	0.152151711
CVotes45AM	-0.169481147	0.0894204131	-2.036287e-02	0.148952902
CVotes45AF	-0.156888553	0.0523317703	-9.898593e-02	0.154218102
CVotes1000	-0.134152275	0.1619711648	7.276765e-05	0.198040378
CVotesUS	-0.168952034	0.0985084878	-2.574642e-02	0.072144271
CVotesnUS	-0.175898559	0.0786154667	6.689571e-03	0.048568524

PC1

Podemos ver como los valores de la meta crítica y el tiempo de duración de las películas no son representativos, mientras que sí lo son, y de forma negativa, tanto el rating final, como la cantidad final de votos, como las cantidades de votos recogidos en las categorías de edades, sexo, ubicación del votante (con respecto a Estados Unidos) y las combinaciones de estas categorías en una misma. De igual forma sucede con el rating alcanzado por una película teniendo en cuenta las anteriormente mencionadas categorías, excluyendo el rating asociado a los votantes de 45 años o más (Votes45A), las votantes de 45 años o más femeninas (Votes45AF), y al rating en IMDb (VotesIMDB); estas tres variables no son representativas.

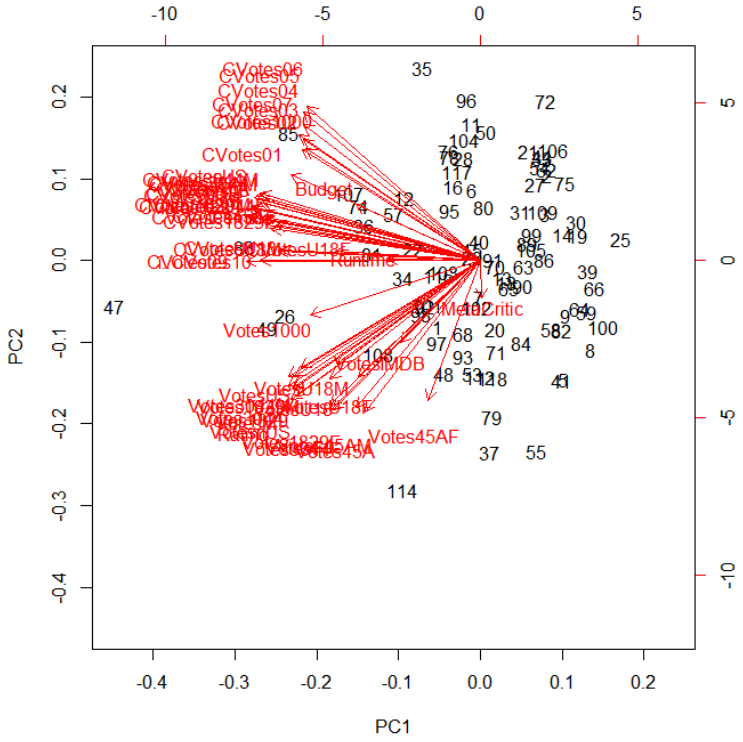
PC2

Resultan representativas de forma positiva la cantidad de votos recogidos en las categorías CVotes descendiendo desde CVotes07 hasta CVotes01, e igualmente se comporta CVotes1000. Por otra parte, son también significativos, pero de manera negativa, el rating final obtenido por una película, así como los ratings asignados según la categorización explicada anteriormente, exceptuando Votes1000 que no resulta representativo para la descripción de esta componente, al igual que el resto de las variables no mencionadas.

PC3

Resultan representativos de manera positiva tanto el rating final obtenido por las películas, como el obtenido por los votantes cuyas edades están comprendidas entre 30 y 44 años. Igualmente sucede con los valores de la cantidad de votos obtenidos en los Estados Unidos, los obtenidos por votantes cuyas edades están comprendidas entre 30 y 44 años, solo los votantes masculinos que cumplen esta condición y las categorías CVotes1000, CVotes01 y Cvotes04. Por otra parte, de forma negativa resultan significativos la cantidad de votos obtenidos por todos los votantes masculinos, todas las votantes femeninas, las votantes femeninas cuyas edades se encuentran entre 18 y 29 años, así como las que superan los 45; los votantes masculinos de menos de 18 años, todos los votantes menores de 18 años y, finalmente, las categorías CVotes10, CVotes08 y CVotes07.

Por último, podemos ver el biplot de las dos primeras componentes:

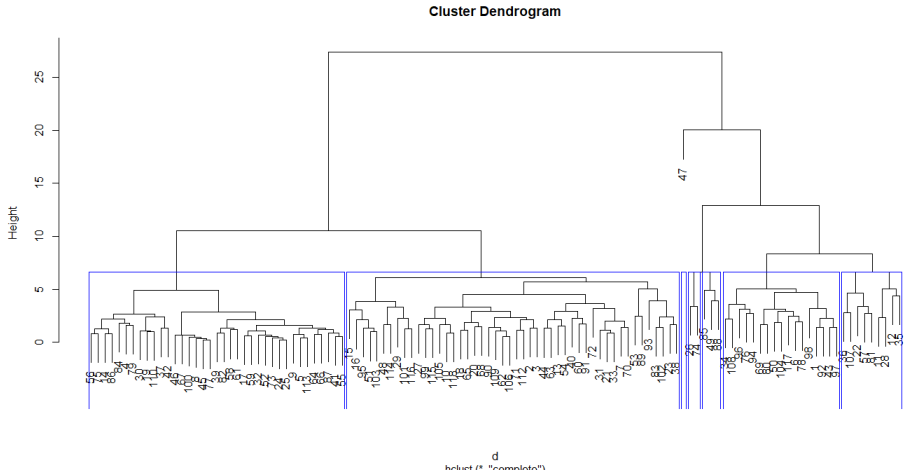


Clustering

La agrupación es una técnica para agrupar datos similares y separar las diferentes observaciones según sus resultados. Los clústers se crean de manera que tengan un orden predeterminado, es decir, una jerarquía.

Para realizar la agrupación, los datos deben seguir ciertas pautas: las columnas deben ser variables, los datos no pueden tener valores faltantes, los datos en las columnas deben estar estandarizados para que las variables sean comparables y, este análisis se realiza a variables cuantitativas solamente. Dado que, nuestros datos recogen varios tipos de información tangibles a diferentes escalas (por ejemplo, las columnas de rating tienen valores continuos entre 6 y 10, mientras que la columna del presupuesto contiene valores relativos a los millones de dólares que se emplean en la realización de una película) y la mayoría de ellos hablan de conteo de votos, decidimos quedarnos con estas columnas para la aplicación de la técnica de clustering.

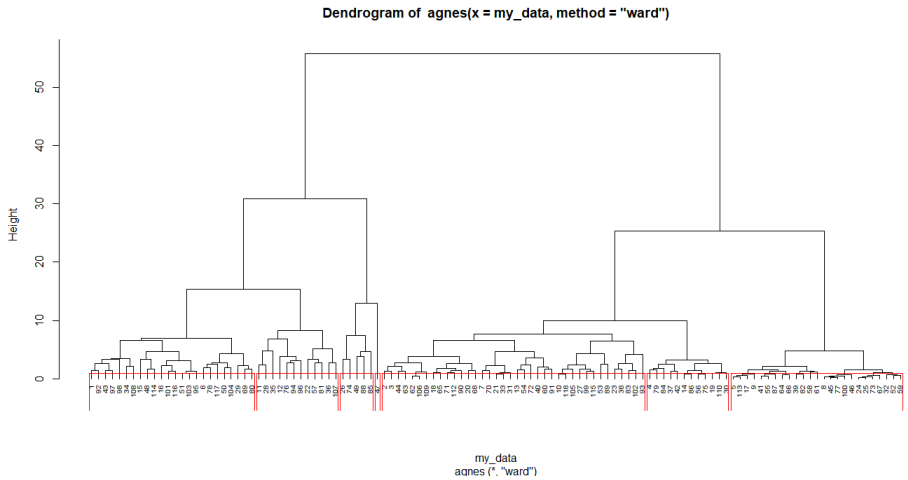
El agrupamiento jerárquico aglomerativo también es conocido como agrupación aglomerativa jerárquica (HAC) o AGNES (acrónimo de aglomeración de anidación). En este método, cada observación se asigna a su propio clúster. Luego, se calcula la similitud (o distancia) entre cada uno de los clústers y los dos clústers más similares se fusionan en uno. Finalmente, se repite este paso hasta que solo quede un grupo. Para la función hclust, se requieren los valores de distancia, tendremos en cuenta la distancia euclideana; así como el método de vinculación (completo, promedio, singular), utilizaremos el completo. Los resultados se muestran a continuación:



Otra alternativa es la función agnes, bastante similiar a la anterior, pero que además arroja el coeficiente de aglomeración (mientras más cercano a uno sea el valor de este, la estructura de agrupación se considera más sólida). De esta forma pudimos analizar la solidez de los resultados utilizando los diferentes métodos posibles, obteniendo como resultado que el mejor método de vinculación es ward ya que es el valor más cercano a la unidad:

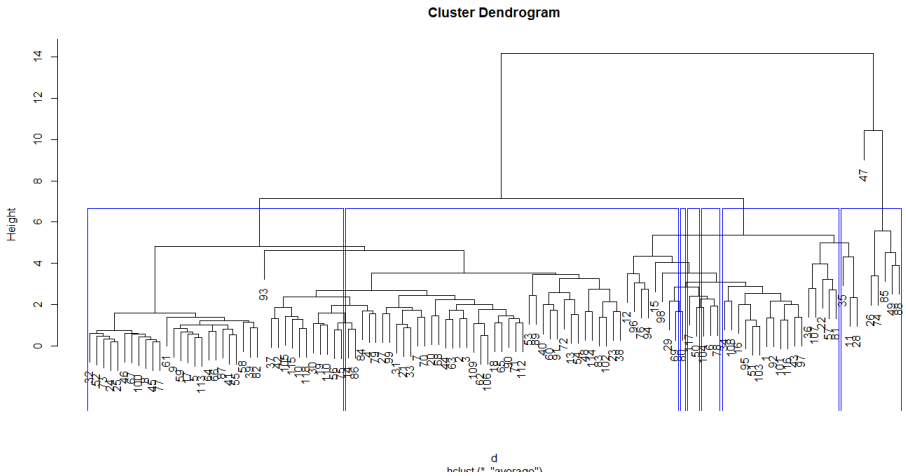
```
> metodos <- c("average", "single", "complete", "ward")
> names(metodos) <- c("average", "single", "complete", "ward")
> map_dbl(metodos, ac)
  average    single  complete      ward
0.8830269 0.8063843 0.9361561 0.9696653
```

En consecuencia, rehicimos el agrupamiento jerárquico aglomerativo con este nuevo método como se muestra en la siguiente imagen:

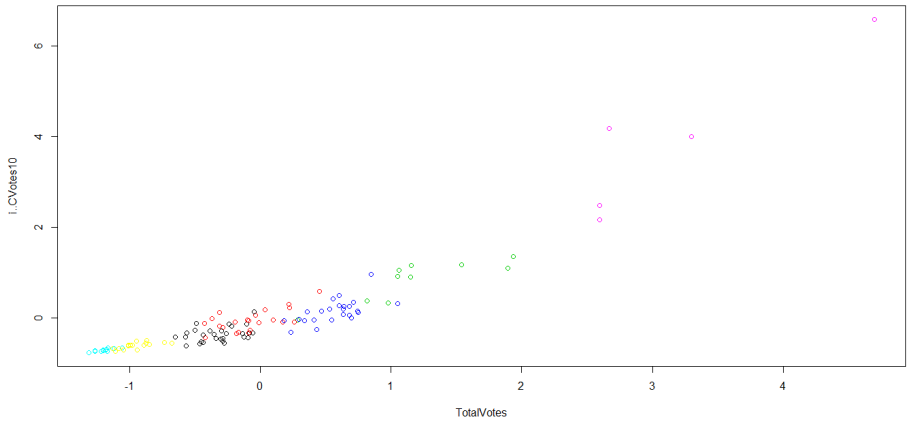


El método de ward apunta a minimizar la varianza total dentro del grupo. En cada paso, se fusionan el par de clústers con una distancia mínima entre ellos. En otras palabras, forma grupos de una manera que minimiza la pérdida asociada con cada grupo. En cada paso, se considera la unión de cada par de clústers posible y se combinan los dos clústers cuya fusión da como resultado un aumento mínimo en la pérdida de información.

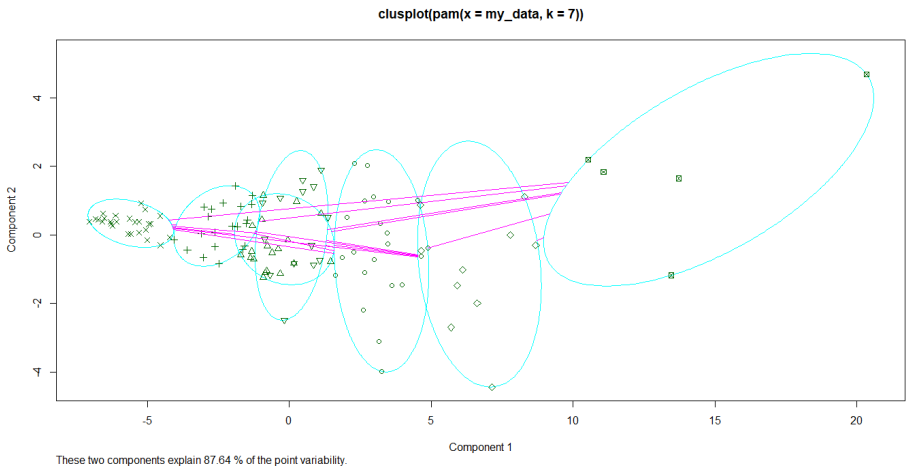
Se anexa también el resultado utilizando el método average:



Consideramos además, la aplicación del método divisivo, donde suponemos que todas las observaciones pertenecen a un único grupo y luego lo dividimos en dos grupos menos similares. Esto se repite recursivamente en cada grupo hasta que haya un grupo para cada observación. Esta técnica también se llama DIANA, llamada así por ser el acrónimo de análisis divisivo. Los algoritmos de partición son enfoques de agrupamiento que dividen los conjuntos de datos, que contienen n observaciones, en un conjunto de k grupos. Para esto, utilizamos el algoritmo de k-means, donde se particiona un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Si graficamos el resultado de k-means, debido al alto volumen de datos con el que trabajamos, obtendremos una matriz de gráficos muy grande, por lo que si queremos analizar las relaciones de forma visual entre variables debemos escoger pares de variables. A continuación, se muestra el resultado para TotalVotes y CVotes10:



Dentro de este mismo grupo de métodos, se encuentra el algoritmo PAM (partición alrededor de mediods), en el que, cada grupo está representado por uno de los objetos en el grupo (similar a la estructura de datos Disjoint-Set) e igualmente intenta minimizar la distancia entre los elementos de un mismo grupo.



Ambos métodos requieren la especificación previa de la cantidad de grupos en que se desea particionar el conjunto de datos. Utilizamos k = 7 debido a que en el algoritmo de clustering previamente realizado habíamos decidido que esta cantidad de clústers era la que mejor se ajustaba al problema (disminuye significativamente el volumen de unidades con las que se trabaja y mantienen una estrecha relación los elementos pertenecientes al mismo grupo).

ANOVA

Para aplicar la técnica de análisis de varianzas modificamos los datos de IMDB.csv para poder agrupar los datos. Primeramente, se realizó el experimento con 5 géneros en grupos de 10, con 7 géneros en grupos de 6, nuevamente con 5 géneros en grupos de 10, modificando esta vez los valores de Rating y Budget para crear 5 grupos de cada uno, ordenándolos de menor a mayor, y asociándoles el valor del promedio de su grupo de 10 respectivamente, y luego con los datos originales solamente tomando la columna Genre1 para definir el género de la película.

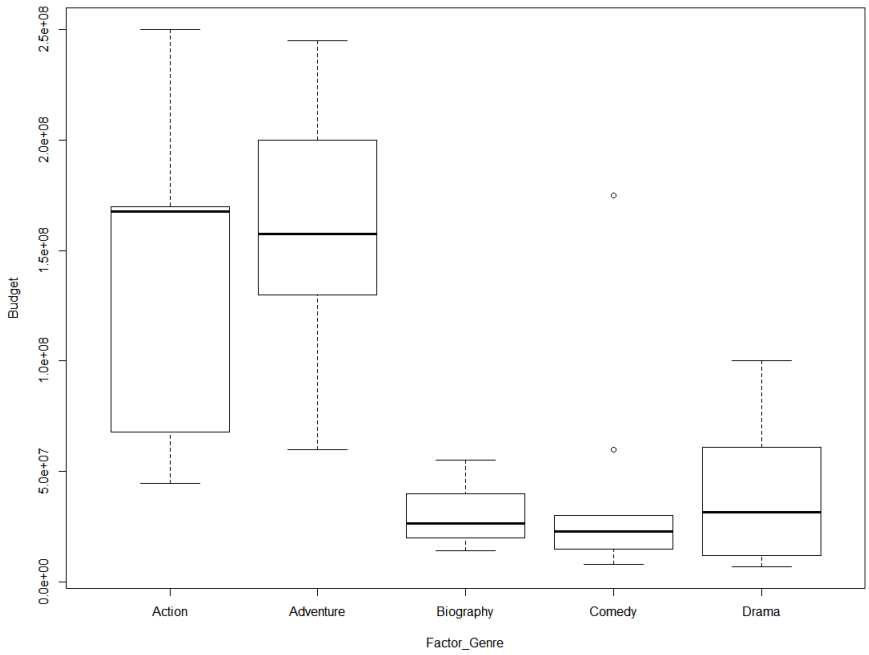
En cada uno de los experimentos se realizaron las siguientes combinaciones:

- 1 - Budget $\sim$ Genre1
- 2 - Rating $\sim$ Genre1
- 3 - MetaCritic $\sim$ Genre1
- 4 - MetaCritic $\sim$ Genre1 + Budget
- 5 - VotesM $\sim$ Genre1
- 6 - CVotesU18 $\sim$ Genre1
- 7 - VotesU18 $\sim$ Genre1
- 8 - Votes45A $\sim$ Genre1
- 9 - Runtime $\sim$ Genre1
- 10 - Runtime $\sim$ Budget
- 11 - MetaCritic $\sim$ Budget
- 12 - Rating $\sim$ VotesUS
- 13 - VotesUS $\sim$ VotesU18

Experimento con 5 géneros en grupos de 10

1 - Budget $\sim$ Genre1

Comprobaremos si existe diferencia entre la cantidad del presupuesto otorgado dependiendo del género de la película.



Al realizar un gráfico con las medias de cada nivel y comparar las medias de los niveles del factor, podemos deducir que probablemente se rechace H_0.

Se muestran los resultados de realizar el análisis de varianzas:

```
> summary(genre_budget.anova)
             Df    Sum Sq   Mean Sq F value   Pr(>F)
Factor_Genre  4 1.581e+17 3.953e+16   17.93 7.19e-09 ***
Residuals    45 9.919e+16 2.204e+15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como el p-value es menor que la significación prefijada $\alpha = 0.05$, se rechaza H_0, y se acepta que al menos existe un par de tipos de género de películas con un promedio de presupuesto otorgado diferente.

Procedemos a comprobar el cumplimiento de los supuestos:

Supuesto 1: Los e_{ij} siguen una distribución normal con media cero.

```
> res <- genre_budget.anova$residuals
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.94159, p-value = 0.01558
```

Al ser la prueba de Shapiro-Wilk significativa podemos rechazar H_0, con lo que rechazaríamos la hipótesis de normalidad en los residuos, por tanto el supuesto no se cumple y el experimento pierde validez.

2 - Rating $\sim$ Genre1

Comprobaremos si existe diferencia entre el valor del rating dependiendo del género de la película. Para esto, primeramente, realizaremos el análisis de varianzas:

```
> summary(genre_rating.anova)
             Df Sum Sq Mean Sq F value Pr(>F)
Factor_Genre  4  0.061 0.01530   0.213   0.93
Residuals    45  3.230 0.07178
```

Como el p-value es mayor que la significación prefijada $\alpha = 0.05$, podemos concluir que no es válido realizar ANOVA en esta combinación.

3 - MetaCritic ~ Genre1

Comprobaremos si existe diferencia entre el valor del rating de MetaCritic dependiendo del género de la película. Para esto, primeramente, realizaremos el análisis de varianzas:

```
> summary(genre_rMeta.anova)
             Df Sum Sq Mean Sq F value Pr(>F)
Factor_Genre  4  586.3  146.57    2.25 0.0786 .
Residuals    45 2931.8   65.15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como el p-value es mayor que la significación prefijada $\alpha = 0.05$, podemos concluir que no es válido realizar ANOVA en esta combinación.

4 - MetaCritic ~ Genre1 + Budget

Comprobaremos si existe diferencia entre el valor del rating de MetaCritic dependiendo del presupuesto que se destinó a la película y del género de la misma. Para esto, primeramente, realizaremos el análisis de varianzas:

```
> summary(genre_budget_rMeta.anova)
              Df Sum Sq Mean Sq F value Pr(>F)
Factor_Genre   4  586.3  146.57   3.441 0.0572 .
Factor_Budget 36 2548.5   70.79   1.662 0.2130
Residuals      9  383.3   42.59
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como el p-value es mayor que la significación prefijada $\alpha = 0.05$, podemos concluir que no es válido realizar ANOVA en esta combinación.

5 - VotesM ~ Genre1

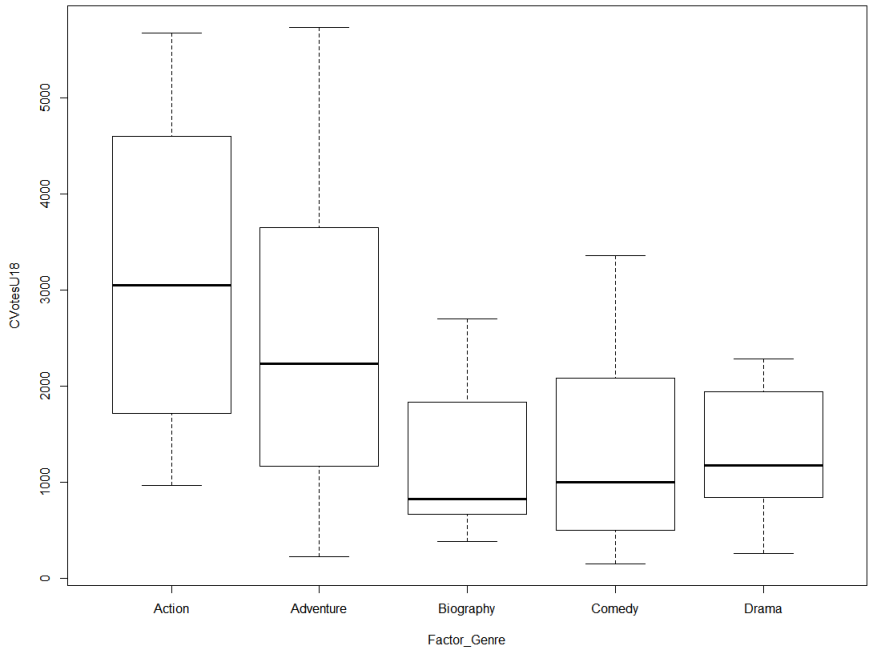
Comprobaremos si existe diferencia entre el valor rating dado por los votos de personas del género masculino dependiendo del género de la película. Para esto, primeramente, realizaremos el análisis de varianzas:

```
> summary(genre_votesm.anova)
             Df Sum Sq Mean Sq F value Pr(>F)
Factor_Genre  4  0.053 0.01330   0.171  0.952
Residuals    45  3.491 0.07758
```

Como el p-value es mayor que la significación prefijada $\alpha = 0.05$, podemos concluir que no es válido realizar ANOVA en esta combinación.

6 - CVotesU18 ∼ Genre1

Comprobaremos si existe diferencia entre la cantidad de votos de personas menores de 18 años, dependiendo del género de la película.



Al realizar un gráfico con las medias de cada nivel y comparar las medias de los niveles del factor, podemos deducir que probablemente se rechace H_0.

Se muestran los resultados de realizar el análisis de varianzas:

```
> summary(genre_cvotesu18.anova)
             Df   Sum Sq Mean Sq F value  Pr(>F)
Factor_Genre  4 32178103 8044526   4.987 0.00205 **
Residuals    45 72587980 1613066
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como el p-value es menor que la significación prefijada $\alpha = 0.05$, se rechaza H_0 y se acepta que al menos existe un par de tipos de género de películas con un promedio de cantidad de votos de personas menores de 18 años diferente.

Procedemos a comprobar el cumplimiento de los supuestos:

Supuesto 1: Los e_{ij} siguen una distribución normal con media cero.

```
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.97565, p-value = 0.3863
```

La prueba de Shapiro-Wilk no es significativa por tanto no podemos rechazar H_0, y podemos afirmar que se cumple la hipótesis de normalidad en los residuos.

Supuesto 2: Los residuos de cada tratamiento tienen la misma varianza σ^2.

```
> bartlett.test(res, my_data5$Factor_Genre)

        Bartlett test of homogeneity of variances

data:  res and my_data5$Factor_Genre
Bartlett's K-squared = 12.411, df = 4, p-value = 0.01454
```

La prueba de Bartlett es significativa (p-value = $0.01 < 0.05$) por lo que podemos rechazar la homogeneidad de las varianzas, por tanto, el supuesto no se cumple y el experimento pierde validez.

7 - VotesU18 ~ Genre1

Comprobaremos si existe diferencia entre el valor del rating de los votos de personas menores de 18 años dependiendo del género de la película. Para esto, primeramente, realizaremos el análisis de varianzas:

```
> summary(genre_votesu18.anova)
             Df Sum Sq Mean Sq F value Pr(>F)
Factor_Genre  4  0.113  0.0282    0.26  0.902
Residuals    45  4.884  0.1085
```

Como el p-value es mayor que la significación prefijada $\alpha = 0.05$, podemos concluir que no es válido realizar ANOVA en esta combinación.

8 - Votes45A ~ Genre1

Comprobaremos si existe diferencia entre el valor del rating de los votos de personas mayores de 45 años dependiendo del género de la película. Para esto, primeramente, realizaremos el análisis de varianzas:

```
> summary(genre_votes45a.anova)
             Df Sum Sq Mean Sq F value Pr(>F)
Factor_Genre  4 0.1532 0.03830   0.809  0.526
Residuals    45 2.1310 0.04736
```

Como el p-value es mayor que la significación prefijada $\alpha = 0.05$, podemos concluir que no es válido realizar ANOVA en esta combinación.

9 - Runtime ~ Genre1

Comprobaremos si existe diferencia entre el valor tiempo de duración de la película dependiendo del género de la misma. Para esto, primeramente, realizaremos el análisis de varianzas:

```
> summary(genre_runt.anova)
             Df Sum Sq Mean Sq F value Pr(>F)
Factor_Genre  4   2794   698.5   2.377  0.066 .
Residuals    45  13224   293.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como el p-value es mayor que la significación prefijada $\alpha = 0.05$, podemos concluir que no es válido realizar ANOVA en esta combinación.

10 - Runtime ~ Budget

Comprobaremos si existe diferencia entre el valor del tiempo de duración de la película dependiendo del presupuesto otorgado a la misma. Para esto, primeramente, realizaremos el análisis de varianzas:

```
> summary(budget_runt.anova)
              Df Sum Sq Mean Sq F value Pr(>F)
Factor_Budget 36  13602   377.8   2.033 0.0852 .
Residuals     13   2416   185.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como el p-value es mayor que la significación prefijada $\alpha = 0.05$, podemos concluir que no es válido realizar ANOVA en esta combinación.

11 - MetaCritic ~ Budget

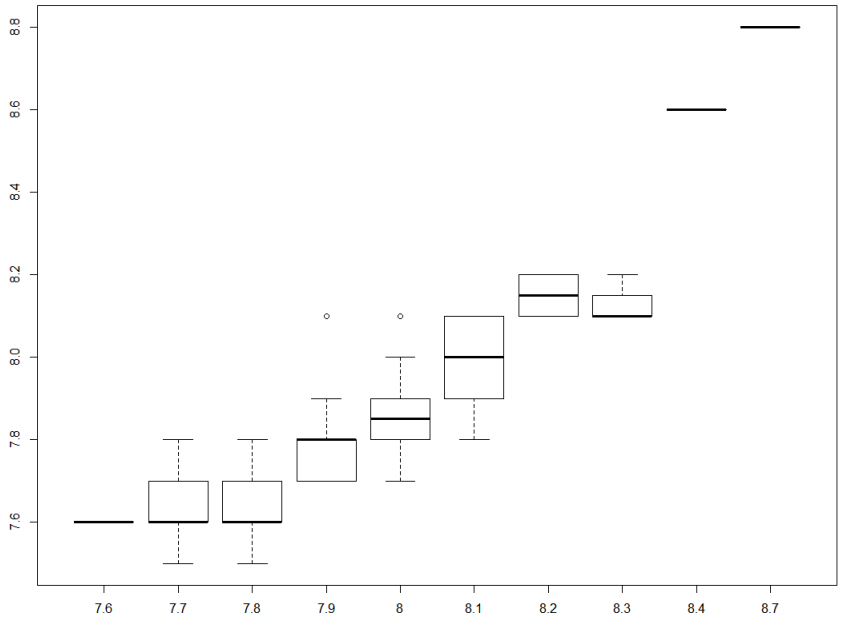
Comprobaremos si existe diferencia entre el valor del rating de MetaCritic dependiendo del presupuesto otorgado a la película. Para esto, primeramente, realizaremos el análisis de varianzas:

```
> summary(budget_rMeta.anova)
              Df Sum Sq Mean Sq F value Pr(>F)
Factor_Budget 36 2841.9   78.94   1.518  0.212
Residuals     13  676.2   52.01
```

Como el p-value es mayor que la significación prefijada $\alpha = 0.05$, podemos concluir que no es válido realizar ANOVA en esta combinación.

12 - Rating ~ VotesUS

Comprobaremos si existe diferencia entre el valor del rating general de la película dependiendo de su rating en Estados Unidos. Para esto, primeramente, realizaremos el análisis de varianzas:



Al realizar un gráfico con las medias de cada nivel y comparar las medias de los niveles del factor, podemos deducir que probablemente se rechace H_0.

Se muestran los resultados de realizar el análisis de varianzas:

```
> summary(votesus_rating.anova)
              Df Sum Sq Mean Sq F value   Pr(>F)
Factor_VotesUS  9 2.8453 0.31615   28.36 1.19e-14 ***
Residuals      40 0.4459 0.01115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como el p-value es menor que la significación prefijada $\alpha = 0.05$, se rechaza H_0 y se acepta que al menos existe un par de rating en Estados Unidos con un promedio de rating general diferente.

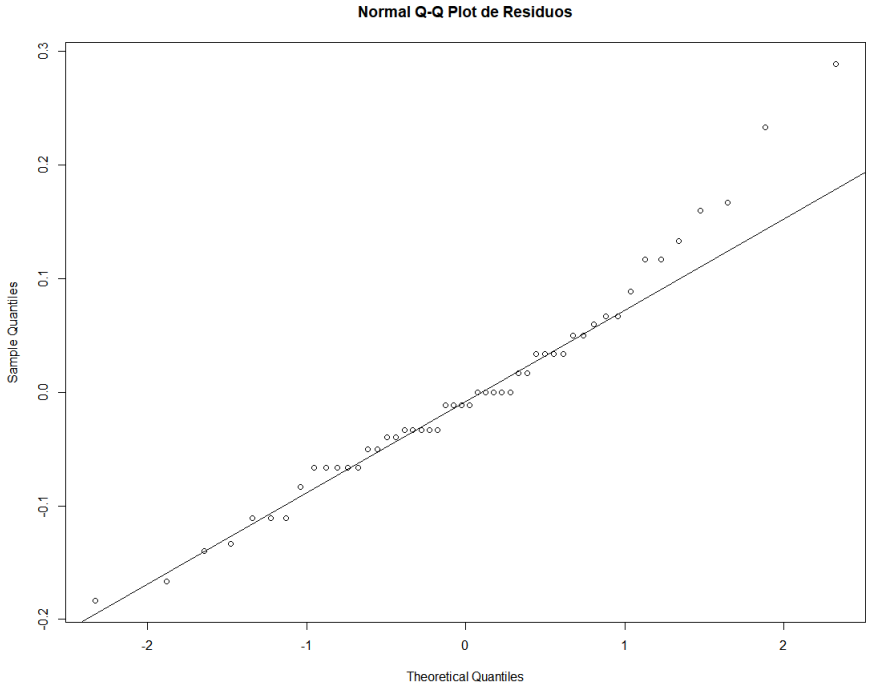
Procedemos a analizar el cumplimiento de los supuestos:

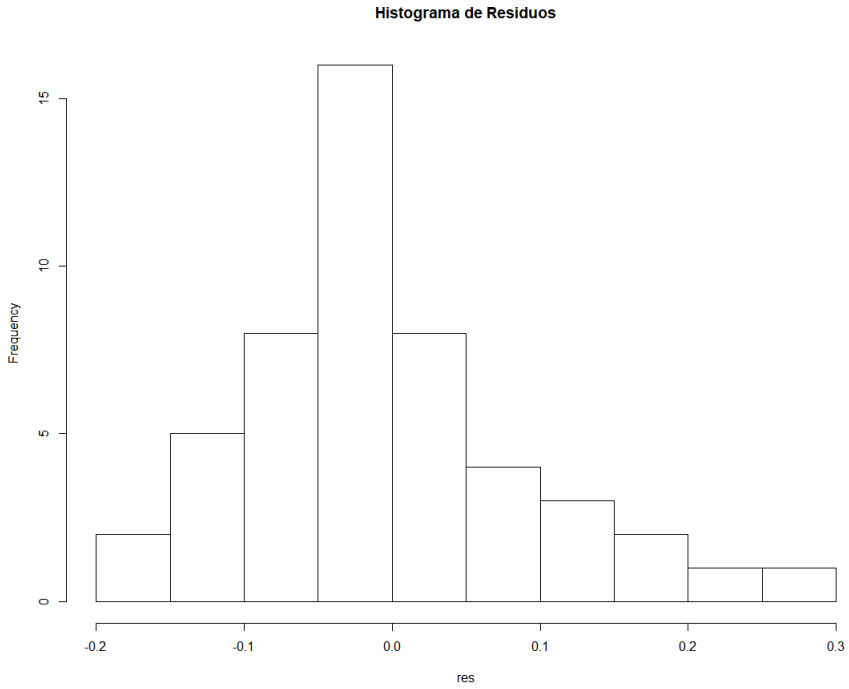
Supuesto 1: Los e_{ij} siguen una distribución normal con media cero.

```
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.9618, p-value = 0.1057
```

La prueba de Shapiro-Wilk no es significativa por tanto no podemos rechazar H_0, y podemos afirmar que se cumple la hipótesis de normalidad en los residuos.

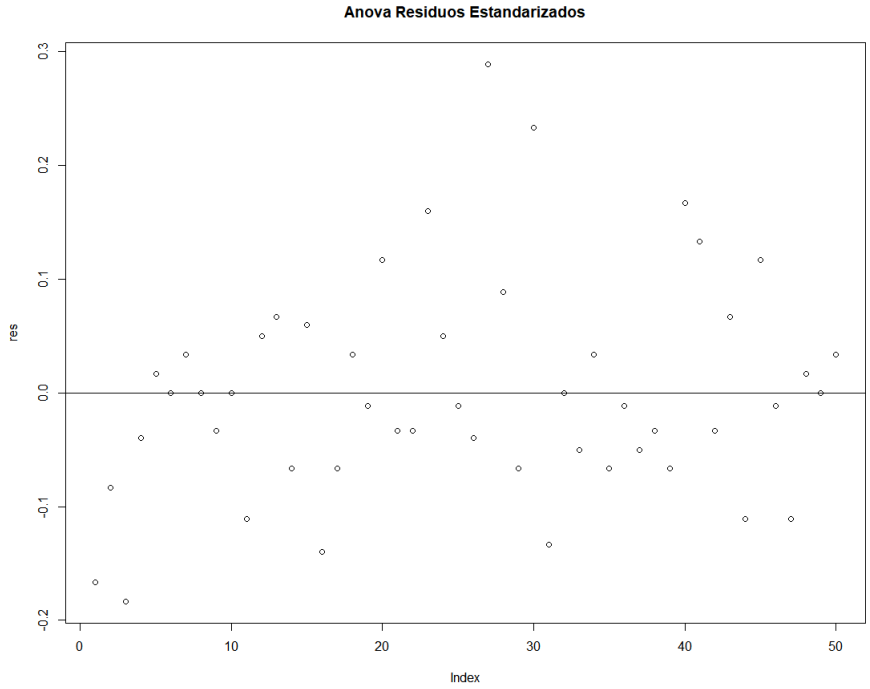
Se puede comprobar en el QQ Plot de Residuos que se encuentra acomodado sobre la recta, y el Histograma de Residuos sigue una forma de campana:





Supuesto 2: Los residuos de cada tratamiento tienen la misma varianza σ^2.

La prueba de Bartlett no es aplicable en este caso, pues se necesita que existan réplicas para aplicarla, pero podemos confirmar la homogeneidad de las varianzas de modo gráfico utilizando el gráfico de residuos estandarizados:



Al analizar la gráfica parece que no sigue un patrón de tubo, embudo o circular, por lo que podemos garantizar que se cumple la homogeneidad de las varianzas.

Supuesto 3: Los e_{ij} son independientes entre sí.

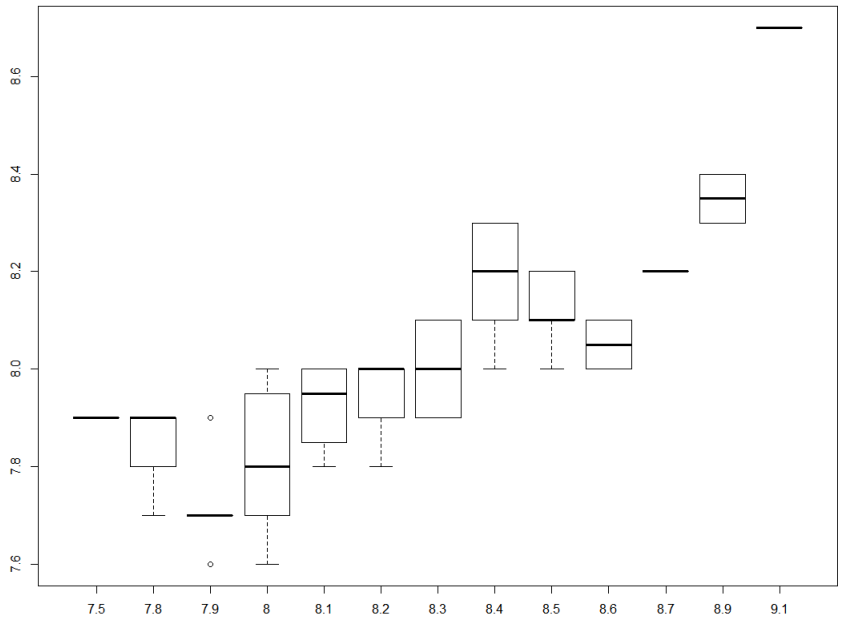
```
> dwtest(votesus_rating.anova)

        Durbin-Watson test

data:  votesus_rating.anova
DW = 2.0833, p-value = 0.6137
alternative hypothesis: true autocorrelation is greater than 0
```

El test de Durbin-Watson no es significativo, por lo que no podemos rechazar H_0 y en este caso los errores sí son independientes.

Al cumplirse los 3 supuestos queda validado el Modelo.

13 - VotesUS ~ VotesU18

Comprobaremos si existe diferencia entre el valor del rating en Estados Unidos dependiendo del rating de los votos de personas menores de 18 años.



Al realizar un gráfico con las medias de cada nivel y comparar las medias de los niveles del factor, podemos deducir que probablemente se rechace H_0.

Se muestran los resultados de realizar el análisis de varianzas:

```
> summary(votesu18_vetesus.anova)
                Df Sum Sq Mean Sq F value   Pr(>F)
Factor_VotesU18 12 1.8902 0.15751   11.32 5.03e-09 ***
Residuals       37 0.5148 0.01391
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como el p-value es menor que la significación prefijada $\alpha = 0.05$, se rechaza H_0 y se acepta que al menos existe un par de rating de personas menores de 18 años con un promedio de rating en Estados Unidos diferente.

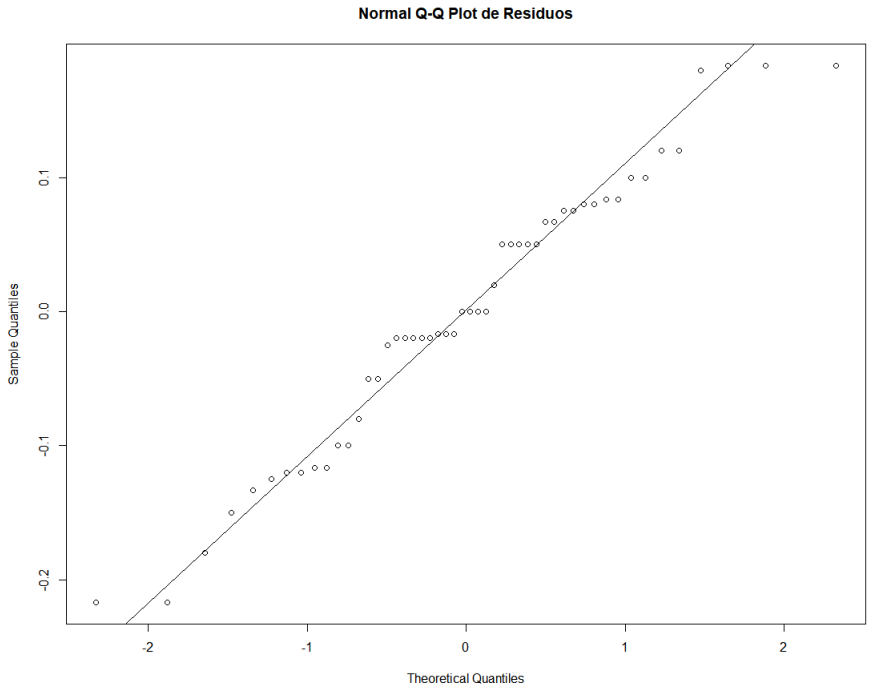
Procedemos a analizar el cumplimiento de los supuestos:

Supuesto 1: Los e_{ij} siguen una distribución normal con media cero.

```
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.96931, p-value = 0.2169
```

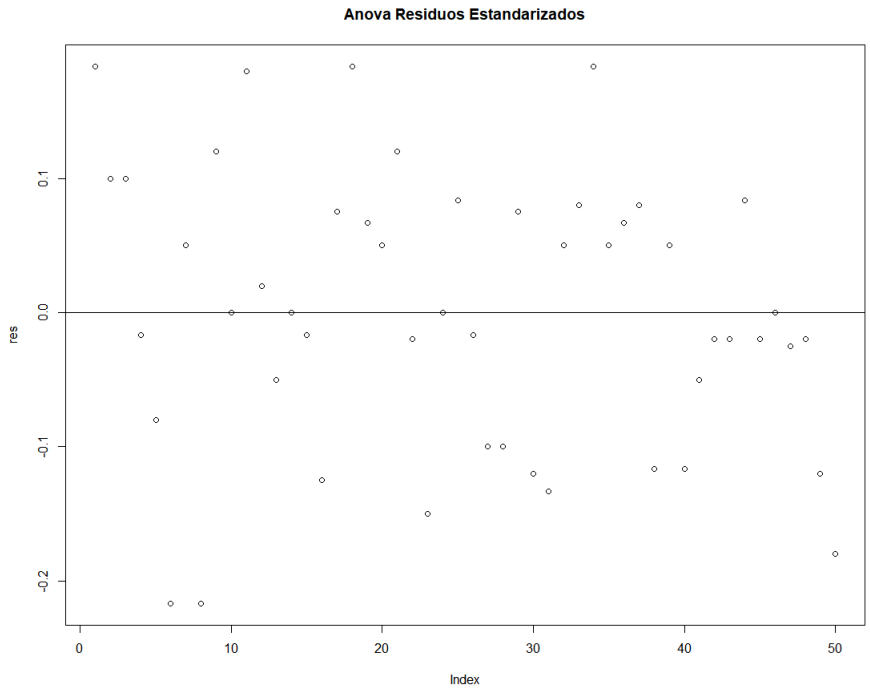
La prueba de Shapiro-Wilk no es significativa por tanto no podemos rechazar H_0, y podemos afirmar que se cumple la hipótesis de normalidad en los residuos.

Se puede comprobar en el QQ Plot de Residuos que se encuentra acomodado sobre la recta:



Supuesto 2: Los residuos de cada tratamiento tienen la misma varianza σ^2.

La prueba de Bartlett no es aplicable en este caso, pues se necesita que existan réplicas para aplicarla, pero podemos confirmar la homogeneidad de las varianzas de modo gráfico utilizando el gráfico de residuos estandarizados:



Al analizar la gráfica parece que no sigue un patrón de tubo, embudo o circular, por lo que podemos garantizar que se cumple la homogeneidad de las varianzas.

Supuesto 3: Los e_{ij} son independientes entre sí.

```
> dwtest(votesu18_vetesus.anova)

        Durbin-Watson test

data:  votesu18_vetesus.anova
DW = 1.577, p-value = 0.06467
alternative hypothesis: true autocorrelation is greater than 0
```

El test de Durbin-Watson no es significativo, por lo que no podemos rechazar H_0 y en este caso los errores sí son independientes.

Al cumplirse los 3 supuestos queda validado el Modelo.

Experimento con 7 géneros en grupos de 6

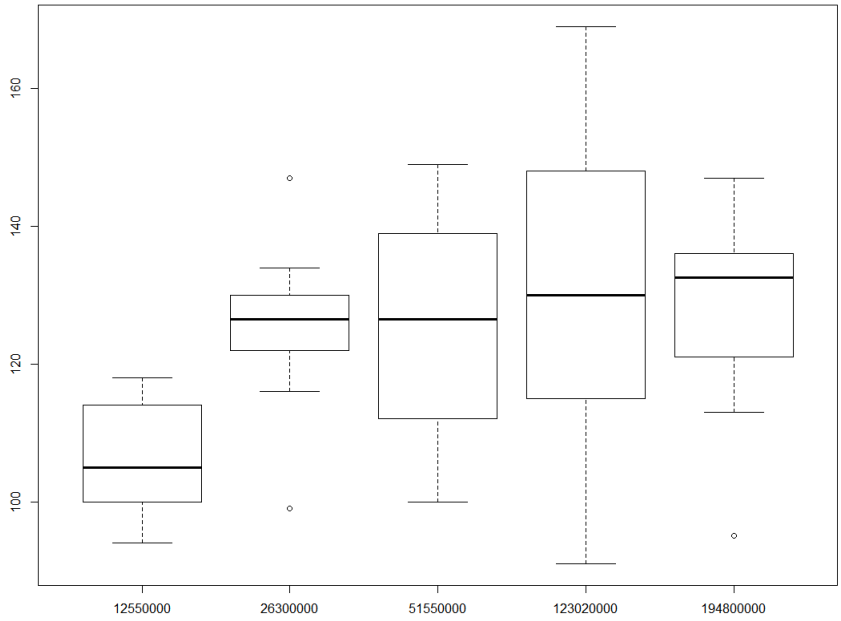
Al realizar el experimento con 7 géneros en grupos de 6, se aprecian los mismos resultados en las combinaciones, es decir, aquellas combinaciones válidas mantuvieron su validez, cumpliéndose esto también para las combinaciones inválidas.

Experimento con 5 géneros, presupuestos y ratings en grupos de 10

Al realizar el experimento con 5 géneros, presupuestos y ratings en grupos de 10, se aprecian cambios en los resultados de las siguientes combinaciones:

10 - Runtime $\sim$ Budget

Comprobaremos si existe diferencia entre el valor del tiempo de duración de la película dependiendo del presupuesto otorgado a la misma.



Al realizar un gráfico con las medias de cada nivel y comparar las medias de los niveles del factor, podemos deducir que probablemente se rechace H_0.

Se muestran los resultados de realizar el análisis de varianzas:

```
> summary(budget_runt.anova)
             Df Sum Sq Mean Sq F value  Pr(>F)
Factor_Budget  4   4233  1058.2   4.041 0.00697 **
Residuals     45  11785   261.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como el p-value es menor que la significación prefijada $\alpha = 0.05$, se rechaza H_0 y se acepta que al menos existe un par de presupuesto otorgado con un promedio de tiempo de duración de la película diferente.

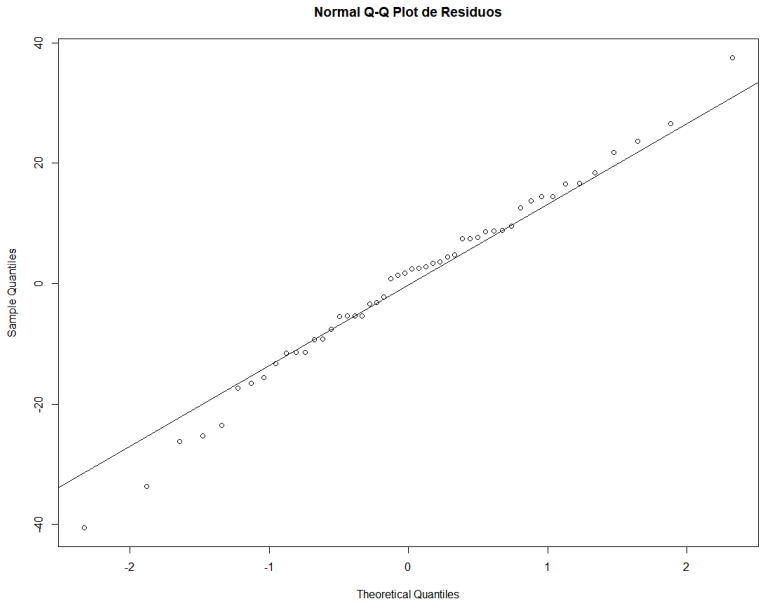
Procedemos a analizar el cumplimiento de los supuestos:

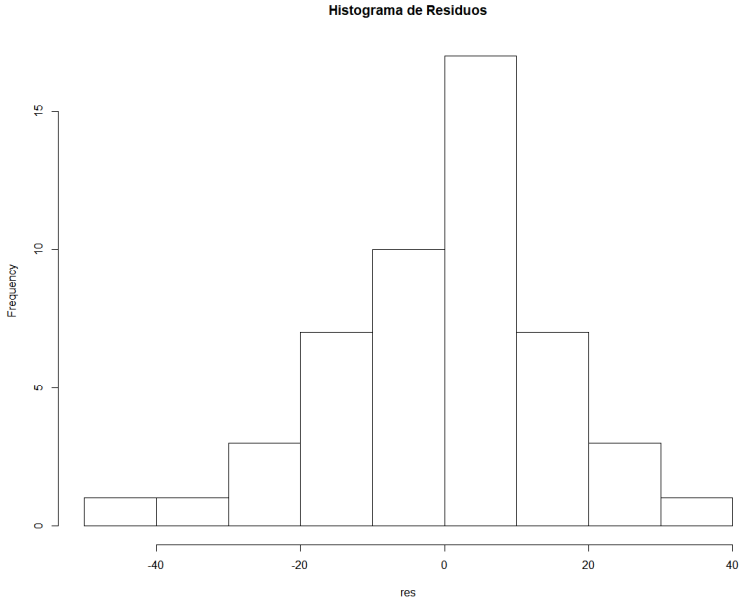
Supuesto 1: Los e_{ij} siguen una distribución normal con media cero.

```
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.98985, p-value = 0.9426
```

La prueba de Shapiro-Wilk no es significativa por tanto no podemos rechazar H_0, y podemos afirmar que se cumple la hipótesis de normalidad en los residuos.

Se puede comprobar en el QQ Plot de Residuos que se encuentra acomodado sobre la recta y el histograma de residuos sigue una forma de campana:





Supuesto 2: Los residuos de cada tratamiento tienen la misma varianza σ^2.

```
> bartlett.test(res, fixed_data$Factor_Budget)

        Bartlett test of homogeneity of variances

data:  res and fixed_data$Factor_Budget
Bartlett's K-squared = 9.3617, df = 4, p-value = 0.05267
```

La prueba de Bartlett no es significativa por tanto no podemos rechazar H_0, se cumple la homogeneidad en los residuos.

Supuesto 3: Los e_{ij} son independientes entre sí.

```
> dwtest(budget_runt.anova)

        Durbin-Watson test

data:  budget_runt.anova
DW = 2.1765, p-value = 0.7047
alternative hypothesis: true autocorrelation is greater than 0
```

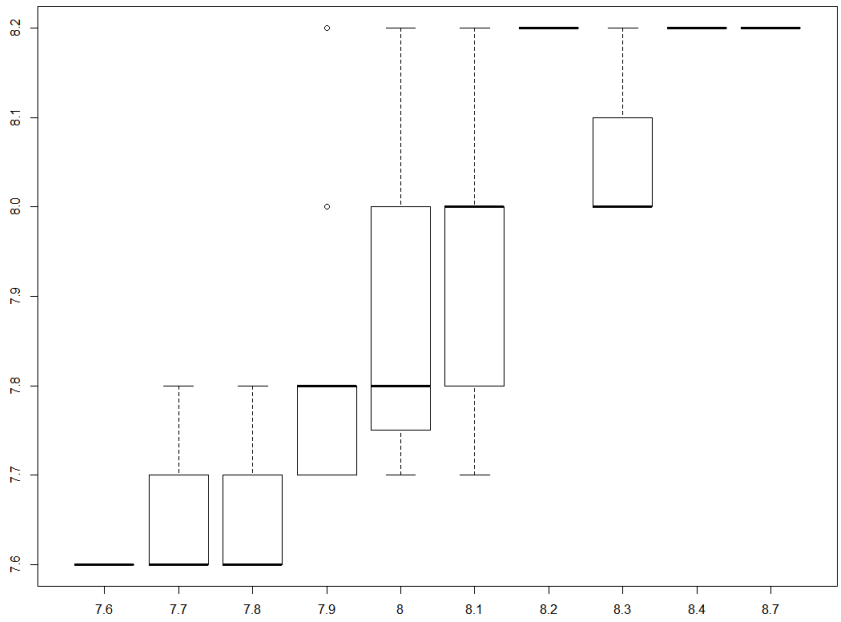
El test de Durbin-Watson no es significativo, por lo que no podemos rechazar H_0 y en este caso los errores sí son independientes.

Al cumplirse los 3 supuestos queda validado el Modelo.

12 - Rating $\sim$ VoteUS

Comprobaremos si existe diferencia entre el valor del rating general de la película dependiendo de su rating en Estados Unidos.

Al realizar un gráfico con las medias de cada nivel y comparar las medias de los niveles del factor, podemos deducir que probablemente se rechace H_0.



Se muestran los resultados de realizar el análisis de varianzas:

```
> summary(votesus_rating.anova)
              Df Sum Sq Mean Sq F value   Pr(>F)
Factor_VotesUS  9 1.5533 0.17259   9.004 2.87e-07 ***
Residuals      40 0.7667 0.01917
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como el p-value es menor que la significación prefijada $\alpha = 0.05$, se rechaza H_0 y se acepta que al menos existe un par de rating en Estados Unidos con un promedio de rating general diferente.

Procedemos a analizar el cumplimiento de los supuestos:

Supuesto 1: Los e_{ij} siguen una distribución normal con media cero.

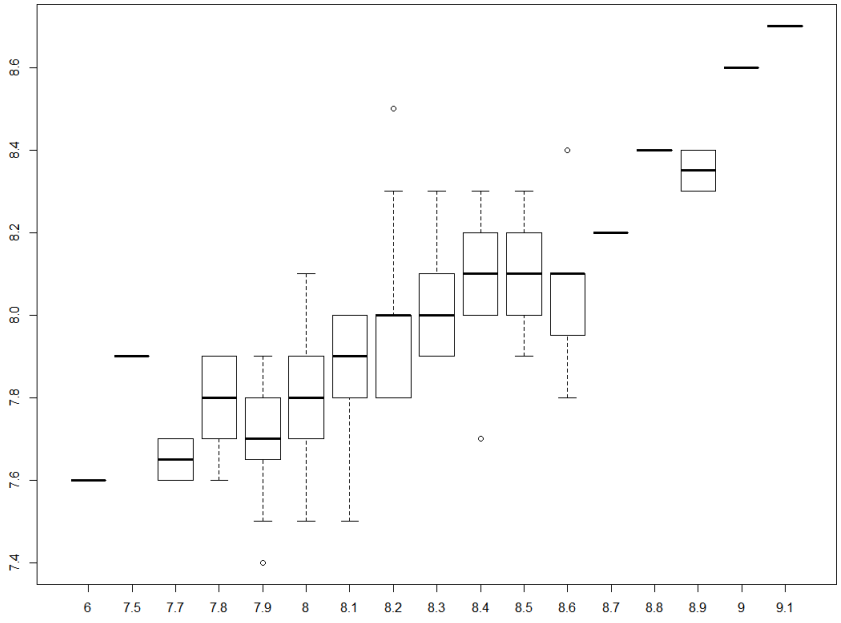
```
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.94279, p-value = 0.0174
```

Al ser la prueba de Shapiro-Wilk significativa podemos rechazar H_0, con lo que rechazaríamos la hipótesis de normalidad en los residuos, por tanto el supuesto no se cumple y el experimento pierde validez.

Experimento con datos originales

Al realizar el experimento con los datos originales, fallan todas las combinaciones realizadas excepto la combinación 13: VotesUS $\sim$ VotesU18 , la cual se comporta de la siguiente forma:

Comprobaremos si existe diferencia entre el valor del rating en Estados Unidos dependiendo del rating de los votos de personas menores de 18 años.



Al realizar un gráfico con las medias de cada nivel y comparar las medias de los niveles del factor, podemos deducir que probablemente se rechace H_0.

Se muestran los resultados de realizar el análisis de varianzas:

```
> summary(votesu18_vetesus.anova)
                Df Sum Sq Mean Sq F value   Pr(>F)
Factor_VotesU18  16  4.019 0.25120   9.992 1.55e-14 ***
Residuals       101  2.539 0.02514
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como el p-value es menor que la significación prefijada $\alpha = 0.05$, se rechaza H_0 y se acepta que al menos existe un par de rating de personas menores de 18 años con un promedio de rating en Estados Unidos diferente.

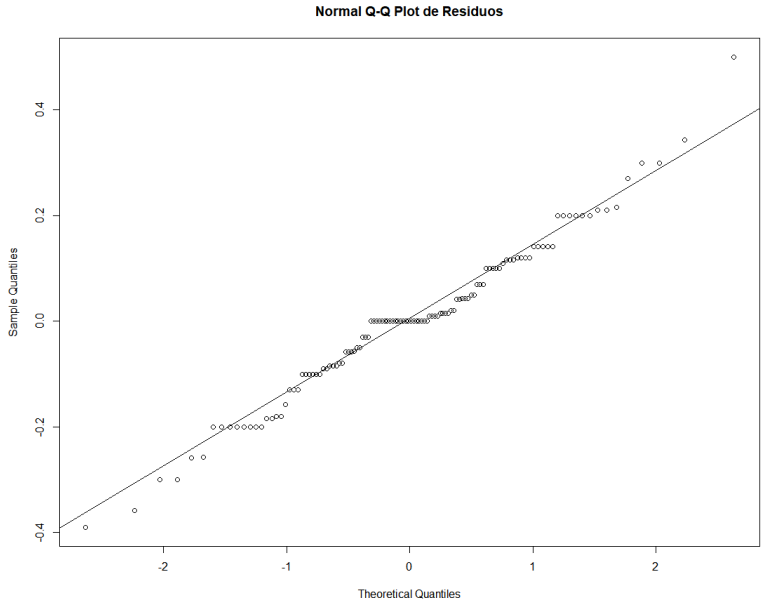
Procedemos a analizar el cumplimiento de los supuestos:

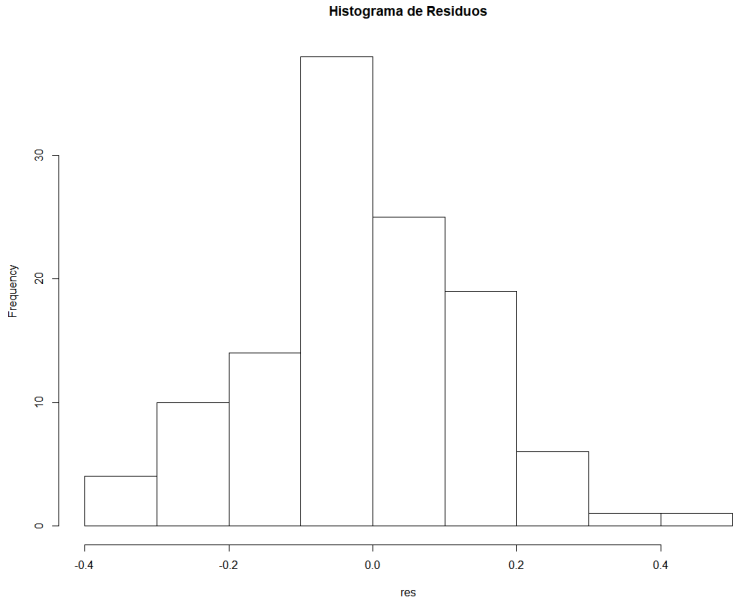
Supuesto 1: Los e_{ij} siguen una distribución normal con media cero.

```
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.98058, p-value = 0.08571
```

La prueba de Shapiro-Wilk no es significativa por tanto no podemos rechazar H_0, y podemos afirmar que se cumple la hipótesis de normalidad en los residuos.

Se puede comprobar en el QQ Plot de Residuos que se encuentra acomodado sobre la recta y el histograma de residuos sigue una forma de campana:

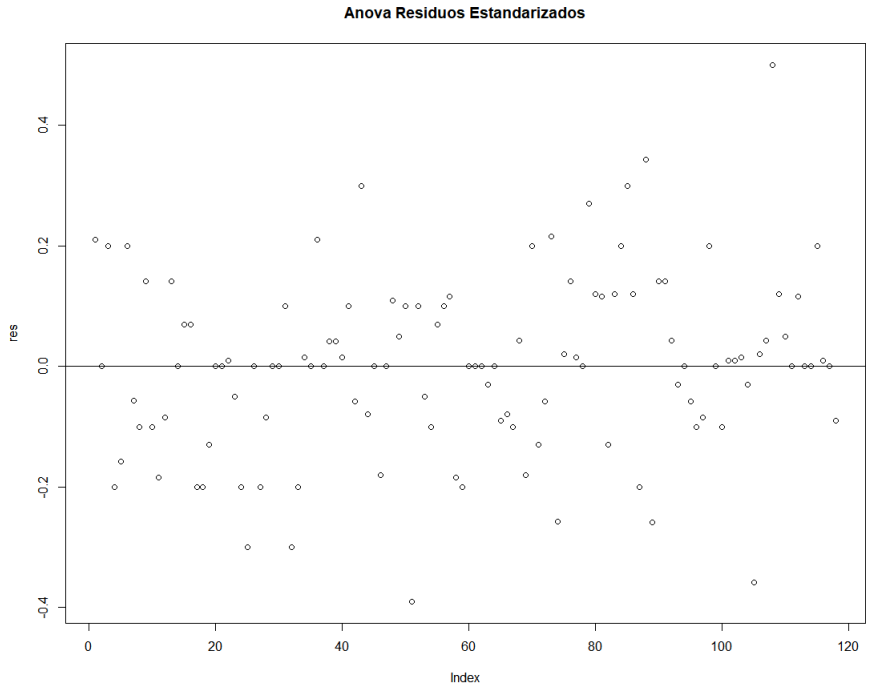




Supuesto 2: Los residuos de cada tratamiento tienen la misma varianza σ^2.

La prueba de Bartlett no es aplicable en este caso, pues se necesita que existan réplicas para aplicarla, pero podemos confirmar la homogeneidad de las

varianzas de modo gráfico utilizando el gráfico de residuos estandarizados:



Al analizar la gráfica parece que no sigue un patrón de tubo, embudo o circular, por lo que podemos garantizar que se cumple la homogeneidad de las varianzas.

Supuesto 3: Los e_{ij} son independientes entre sí.

```
> dwtest(votesu18_vetesus.anova)

        Durbin-Watson test

data:  votesu18_vetesus.anova
DW = 2.0204, p-value = 0.5351
alternative hypothesis: true autocorrelation is greater than 0
```

El test de Durbin-Watson no es significativo, por lo que no podemos rechazar H_0 y en este caso los errores sí son independientes.

Al cumplirse los 3 supuestos queda validado el Modelo.

Aclaraciones

Al no tener una leyenda de la notación utilizada para los nombres de las variables del data frame, nuestro equipo presenta confusión en el significado de las variables que se especifican a continuación:

- CVotes10 - CVotes01
- Votes10 - Votes01
- CVotes1000 y Votes 1000
- CVotesIMDb y VotesIMDb

Por tanto, se hace referencia a estas a través de su nombre.

Especificaciones del código

Para la aplicación de todas las técnicas, se nos hizo necesario el pre procesamiento de los datos, para lo cual recurrimos a las funciones utilizadas en la primera fase de este proyecto que se apoyan en ConvertCurrency [1].

El procedimiento que se utiliza para esto provoca que en los espacios en blanco de la tabla de los datos se inserten valores NA, por lo cual, para funciones como cor, que se utiliza para hallar la matriz de correlación se toman solo los pares donde ambos elementos sean distintos de NA (use = "pairwise.complete.obs"). De igual manera, al generar un modelo de regresión que involucra una columna que contine un valor NA, los residuos del mismo no tienen en cuenta los valores asociados a la fila donde se encuentra este (fila 111); esto conlleva que para graficar la relación entre los residuos y las columnas que contituyen los regresores, debe eliminarse la fila de estos últimos para que coicidan los tamaños.

Se requiere la instalación de los paquetes siguiente: lmtest, ggplot2, psych, purrr, cluster, factoextra, gridExtra.

Referencias

[1] - https://stackoverflow.com/questions/7337824/read-csv-file-in-r-with-currency-column-as-numeric