Informe del Segundo Mini Proyecto de Estadística. Curso 2020-2021

Daniel Alberto García Pérez D.GARCIA@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

Leonel Alejandro García López L.GARCIA3@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

Roberto Marti Cedeño R.MARTI@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

Tutor(es):
Msc. Dalia Diaz Sistachs, *Facultad de Matemática y Computación, Universidad de La Habana*

Tema: Estadística, Regresión Lineal Múltiple.

1. Ejercicios

Mostramos a continuación los enunciados de los ejercicios propuestos.

Los datos que se muestran a continuación son un año de las ventas y gastos de una compañía divididos en meses.

Mes	Gastos	Ventas
1	1000	9914
2	4000	40487
3	5000	54324
4	4500	50044
5	3000	34719
6	4000	42551
7	9000	94871
8	11000	118914
9	15000	158484
10	12000	131348
11	7000	78504
12	3000	36284

1. Defina quienes son las variables independientes y la variable dependiente.
2. Realice en R un modelo de regresión lineal simple con una de las variables independientes y analice los resultados.
3. Realice en R un modelo de regresión múltiple utilizando todas las variables independientes.
4. Compare los resultados de ambos modelos.

1.1 Ejercicio 2

Encuentre el modelo de regresión que mejor se ajuste a los datos que se encuentran en el archivo *Advertising.csv*. La variable dependientes es las ventas realizadas (*sales*) y las independientes es la cantidad de dinero invertido en publicidad en tres medios, televisión (*TV*), Radio (*radio*) y Periódicos (*newspaper*).

2. Soluciones

A continuación, se listan las soluciones a los ejercicios planteados.

2.1 Solución del Ejercicio 1

Como parte del análisis del modelo, primero analizamos las variables. Emplearemos para la solución un enfoque Backward.

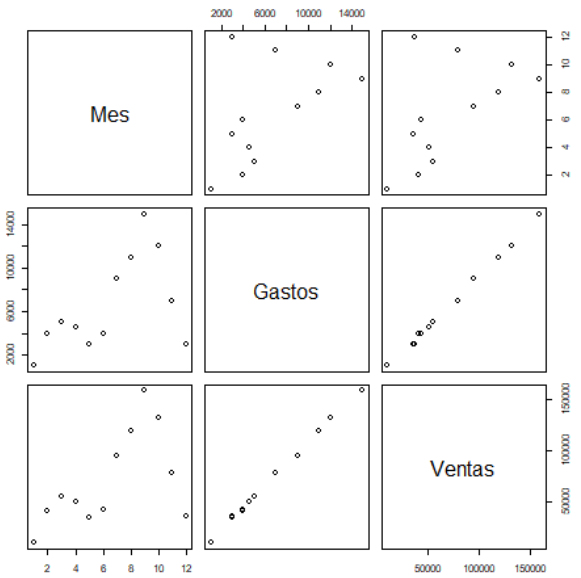

Figure 1: Gráfico de Dispersión.

Del gráfico de dispersión correspondiente (figura 1)

	Mes	Gastos	Ventas
Mes	1	0.5271186	0.5573783
Gastos	0.5271186	1	0.9988322
Ventas	0.5573783	0.9988322	1

Figure 2: Matriz de correlación.

podemos notar que, los gastos y las ventas están linealmente relacionados, proposición que corroboramos con la tabla de correlación (figura 2).

También podemos deducir la misma conclusión atendiendo al comportamiento del comercio. La ganancia obtenida de las ventas es como mínimo la cantidad de productos o servicios comercializados menos la inversión realizada.

Por la poca información obtenida respecto a los meses, y, que no esta relacionado fuertemente ni con las ventas ni con los gastos, se determinó que se asume no relacionada linealmente a estas dos últimas variables del modelo. Es posible que las ventas dependan de una estacíon específica del año pero no se dispone de información suficiente para corroborar este planteamiento.

Para dar cumplimiento al supuesto de que, las variables independientes no pueden estar relacionadas, sólo podemos tomar como variable dependiente a los gastos o las ventas. Como por norma general las ganancias se obtienen después de una inversión, debido a la naturaleza que podemos deducir del problema se tomó las ventas como variable dependiente.

```
Call:
lm(formula = Ventas ~ Gastos, data = ejercicio1)

Residuals:
   Min     1Q Median     3Q    Max
 -3385  -2097    258   1726   3034

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1383.4714  1255.2404   1.102    0.296
Gastos        10.6222     0.1625  65.378 1.71e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2313 on 10 degrees of freedom
Multiple R-squared:  0.9977,    Adjusted R-squared:  0.9974
F-statistic:  4274 on 1 and 10 DF,  p-value: 1.707e-14
```

Figure 3: Resumen del modelo linar simple.

```
Call:
lm(formula = Ventas ~ Gastos + Mes, data = ejercicio1)

Residuals:
     Min       1Q   Median       3Q      Max
-1793.73 -1558.33    -1.73  1374.19  1911.58

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -567.6098  1041.8836  -0.545  0.59913
Gastos        10.3825     0.1328  78.159 4.65e-14 ***
Mes          541.3736   158.1660   3.423  0.00759 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1607 on 9 degrees of freedom
Multiple R-squared:  0.999,     Adjusted R-squared:  0.9988
F-statistic:  4433 on 2 and 9 DF,  p-value: 3.368e-14
```

Figure 4: Resumen del modelo linar múltiple.

El modelo lineal simple (figura 3) es representado por la siguiente ecuación.

$$Ventas = B_{10} + Gastos * B_{11}$$

El modelo lineal múltiple (figura 3) es representado por la siguiente ecuación.

$$Ventas = B_{20} + Gastos * B_{21} + Mes * B_{22}$$

2.1.1 Residuals

El modelo lineal simple muestra perturbaciones considerables con respecto a los parámetros presentes en los datos. Esto implica que podemos obtener valores de ventas distantes de los presentes en los datos.

El modelo lineal múltiple por otro lado, brinda una excelente aproximación respecto a la mediana, y mejora también el resto de los residuos con respecto al modelo lineal simple.

2.1.2 Coefficients

Tanto en el modelo lineal simple como en el múltiple, podemos observar que el valor del intercepto no aporta nada al modelo en cuestión ($Pr(> |t|) > 0.05$), ya sea por falta de datos o por la necesidad de incluir una nueva variable en el modelo, el intercepto no aporta a la estimación de las ventas sobre las variables independientes. Una vez mas la falta de información sobre el tema afecta la capacidad de decisión sobre los resultados. Es posible que en el caso del modelo lineal, se ingrese algún tipo de interés por un préstamo o servicio y en el modelo múltiple se pague la deuda de un proveedor mensualmente. Al no disponer de datos suficientes y al no significar para los modelos, se toman $B_{10} = 0$ y $B_{20} = 0$.

En el modelo simple, por cada unidad invertida en los gastos podemos esperar un incremento en las ventas de $B_{11} = 10.62$ unidades aproximadamente. En el modelo múltiple, podemos esperar un incremento de las ventas en $B_{21} = 10.38$ y $B_{22} = 541.37$ unidades monetarias por cada unidad invertida en los gastos y por cada mes que transcurre en el año respectivamente.

2.1.3 Performance Measures

Como reflejo de la conclusión que se llegó con respecto a los residuos, el error estándar de los residuos es menor en el modelo múltiple que el simple.

Mientras que en el modelo simple, el valor de $Rcuadradoajustado$ es 0.9974, en el modelo múltiple, se aumenta hasta llegar a 0.9988. En ambos casos podemos decir que mas del 99% de las ventas puede ser explicadas a partir de las variables tomadas en el modelo.

En ambos modelos el p-valor del Estadígrafo F es menor que 0.05 por lo que podemos afirmar que existe una variable significativamente distinta de 0.

2.1.4 Supuestos del modelo

La media de los errores de los residuos es $-2.368846e-13$ y $-6.635896e-14$ para el modelo simple y el múltiple respectivamente, mientras la suma de los errores de los residuos es $-2.842171e-12$ y $-7.958079e-$

13 respectivamente. Podemos afirmar entonces que, tanto la media como la suma de los errores son 0 para ambos modelos.

Mientras que en el modelo lineal simple, la prueba de Durbin-Watson indica que no se puede rechazar la hipótesis nula ($p - valor = 0.03062 < 0.05$), mientras que en el modelo múltiple podemos decir que los residuos son independientes ($p - valor = 0.3133 >> 0.05$).

Analizando los histogramas derivados de los modelos (Figuras 9, y 10), no nos podemos percatar que sigan una forma de campana, pero al valorar los QQ-plots (Figuras 11 y 12) podemos asumir que los errores siguen una distribución normal en ambos modelos, dado que presentan pequeñas desviaciones.

Finalmente podemos comprobar que el supuesto de Homocedasticidad se cumple al analizar los gráficos de los residuos estandarizados. (Figura 13 y 14)

2.1.5 Conclusiones

Tras un análisis de ambos modelos lineales, podemos llegar a la conclusión de que el modelo lineal no funciona, dado que incumple el supuesto de independencia de los residuos, mientras que el modelo múltiple presenta una mejora considerable respecto a la estimación y cumple con todos los supuestos. Para la estimación de las ventas emplearíamos entonces el modelo múltiple.

2.2 Solución del Ejercicio2

Primero analizamos la relación entre las variables.

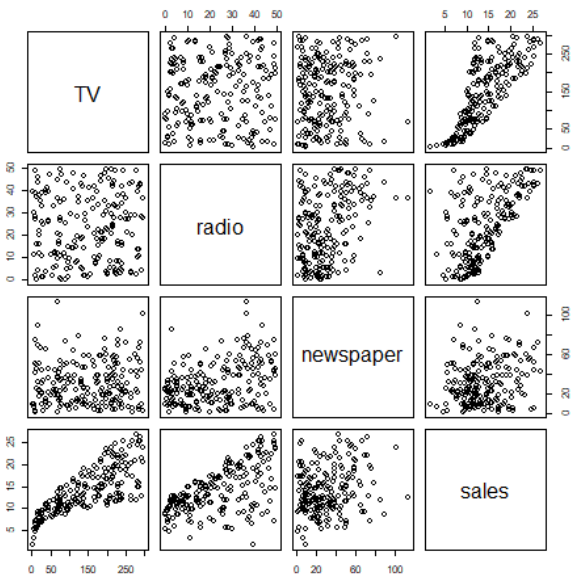

Figure 5: Gráfico de Dispersión.

Como podemos determinar del gráfico de dispersión y la tabla de correlación (Figuras 5 y 6 respectivamente), existe una relación lineal entre las inversiones realizadas en TV y las ventas obtenidas, por lo que tiene sentido emplear un modelo lineal. Como mismo se delimita en los requerimientos del ejercicio, tomamos pues a las ventas como variable dependiente y a las inversiones en radio, periódico y tv como independientes.

	TV	radio	newspaper	sales
TV	1	0.06	0.06	0.78
radio	0.06	1	0.35	0.58
newspaper	0.06	0.35	1	0.23
sales	0.78	0.58	0.23	1

Figure 6: Matriz de correlación.

2.3 Enfoque Backward

Primero comenzamos con el modelo teniendo en cuenta todas las variables independientes.

```
Call:
lm(formula = sales ~ TV + radio + newspaper, data = ejercicio2)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
radio        0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Figure 7: Descripción del modelo.

El modelo descrito anteriormente (Figura 7) define la siguiente ecuación:

$$Sales = B_{30} + TV * B_{31} + Radio * B_{32} + Newspaper * B_{33}$$

2.3.1 Análisis con todas las variables independientes

En el modelo anteriormente planteado los residuos se encuentran cerca de los valores reales planteados por los datos disponibles.

Con respecto a los coeficientes, podemos decir que tanto el intercepto como la TV y la radio son de gran significación para la estimación de las ventas. No es el caso con el periódico el cual no aporta valor al cálculo lineal de las ventas.

Como reflejo de la buena aproximación de los residuos, el valor del error estándar residual es pequeño. Por otro lado el valor de $Rcuadradoajustado$ es 0.8956 lo cual nos indica que el modelo es bueno, pero no óptimo. Como el valor del $p-valor$ es considerablemente menor a 0.05 podemos decir que existe al menos una variable significativamente distinta de 0.

Como parte del análisis de los supuestos del modelo podemos decir que, la media de los errores es 3.009962_e-17 y la suma de los errores de los residuos es $6.036838_e - 15$ ambos son bien cercanos a 0 por lo que el primer supuesto se cumple. La prueba de Durbin-Watson resulta en un $p - valor = 0.7236 >> 0.05$ por

lo que podemos concluir que los residuos son independientes. Analizando el histograma (Figura 15) y comprobando con la gráfica QQ-plot (Figura 16), podemos llegar a la conclusión de que los errores siguen una distribución normal. Finalmente al valorar el gráfico de los residuo estandarizados (Figura 17) podemos decir que el supuesto de homocedasticidad se cumple.

El desempeño del modelo planteado anteriormente es válido, pero podemos percatarnos que los periódicos no son de valor significativo para las ventas, posiblemente debido a la radio y la televisión como medios de difusión masiva. Entonces procedamos a remover la variable periódico de nuestro modelo y comprobar si mejora el mismo.

```
Call:
lm(formula = sales ~ TV + radio, data = ejercicio2)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7977 -0.8752  0.2422  1.1708  2.8328

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.92110    0.29449   9.919   <2e-16 ***
TV           0.04575    0.00139  32.909   <2e-16 ***
radio        0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

Figure 8: Resumen del modelo sin la variable newspaper.

A primera vista, el modelo mejora de forma pequeña el valor de $Rcuadradoajustado$ a 0.8962, el resto de los parámetros se comporta de forma similar al modelo anteriormente planteado, este mejora, poco, y depende de una variable menos, lo que simplifica el proceso de cálculo.

Tal como se cumple en el modelo anterior podemos comprobar los supuestos con las gráficas (18, 19, 20), sabiendo además que la media de los errores es $-6.464447_e - 17$ y su suma $-1.301736_e - 14$. Para tratar de mejorar la aproximación se intentó también emplear los datos estandarizados, sin cambio alguno.

Uno de los principales factores que no se mejore la aproximación puede ser que tanto los anuncios, publicados tanto en tv como en la radio describen una curva mas que una recta y pueden tener alguna relación que no sea lineal entre ellos.

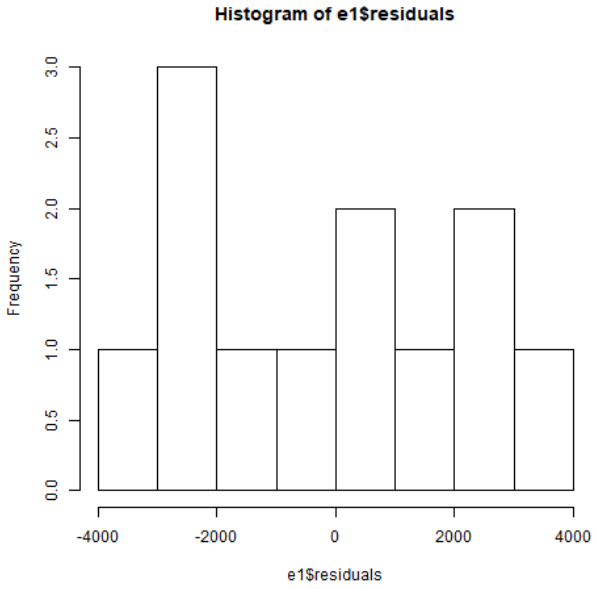

Figure 9: Histograma de los residuos del modelo simple.

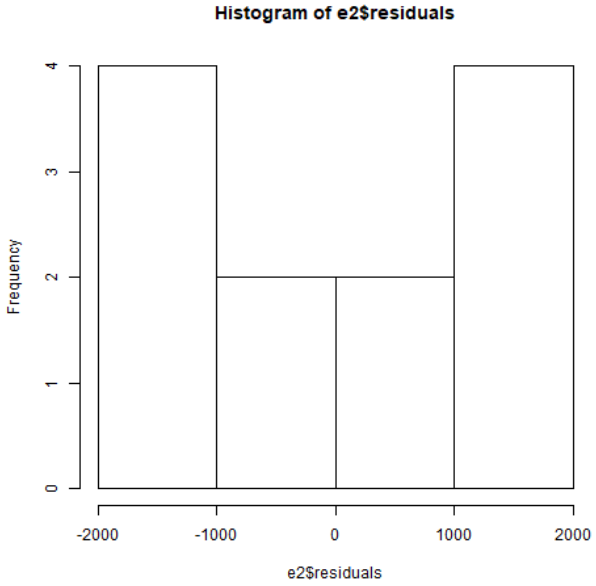

Figure 10: Histograma de los residuos del modelo múltiple.

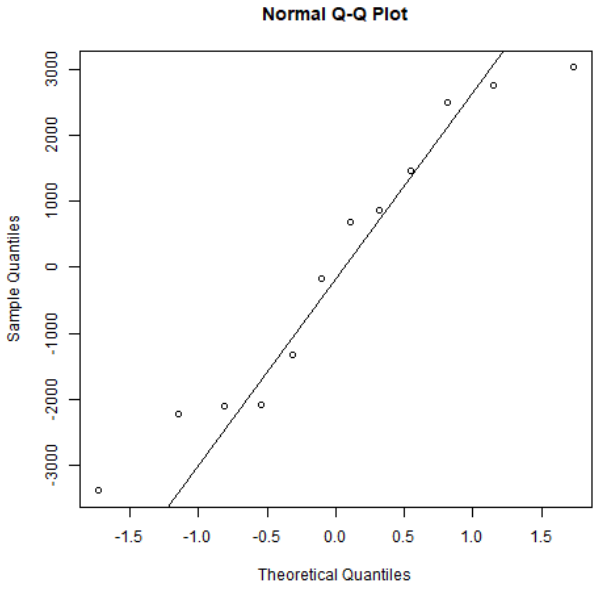

Figure 11: Gráfico QQ-plot del modelo simple.

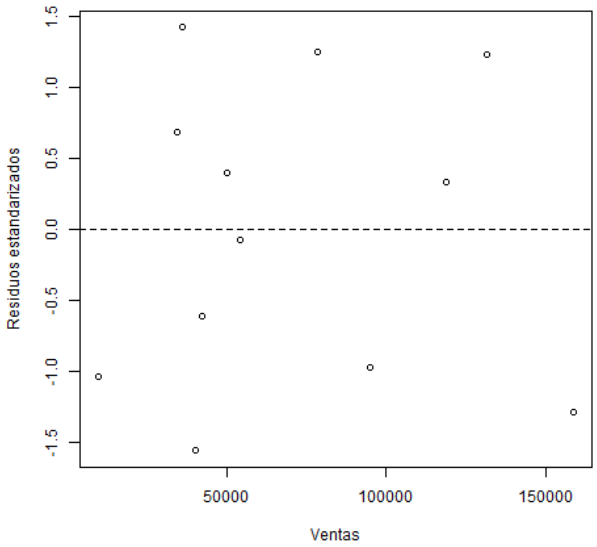

Figure 13: Gráfico Residuos estandarizados del modelo simple.

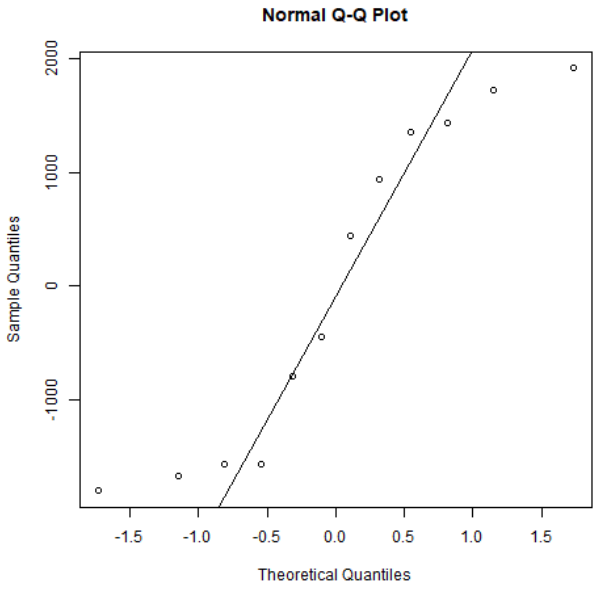

Figure 12: Gráfico QQ-plot del modelo múltiple.

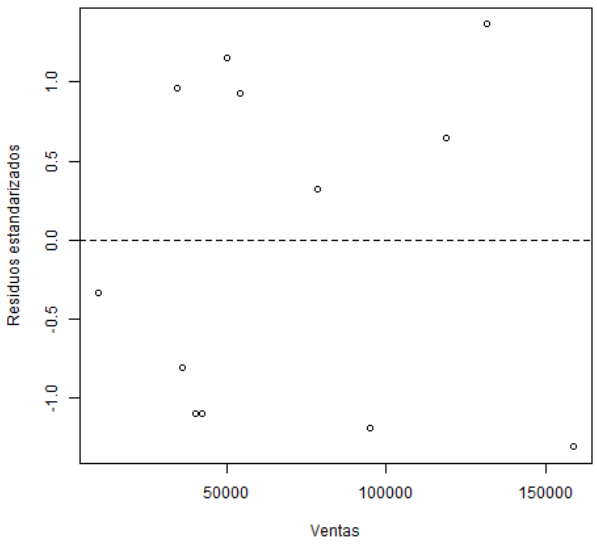

Figure 14: Gráfico Residuos estandarizados del modelo múltiple.

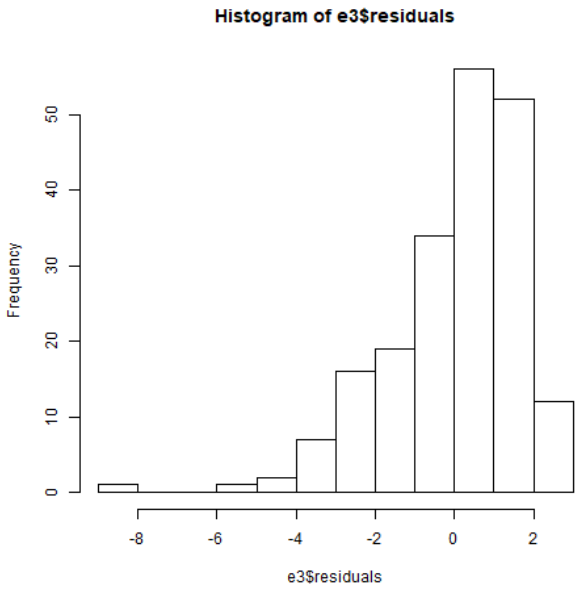

Figure 15: Histograma de los residuos.

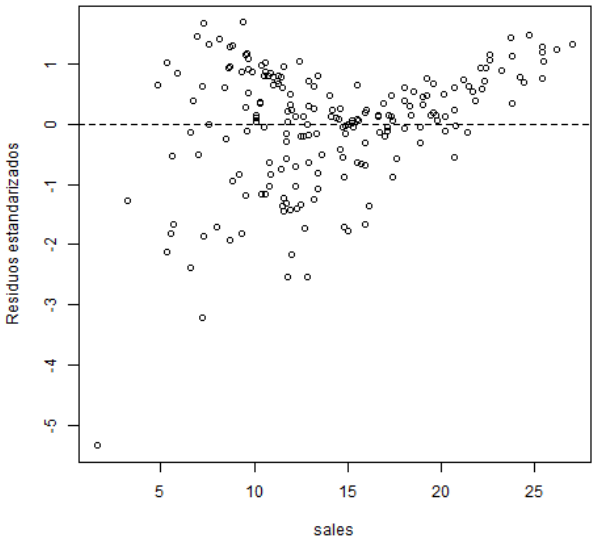

Figure 17: Gráfico Residuos estandarizados.

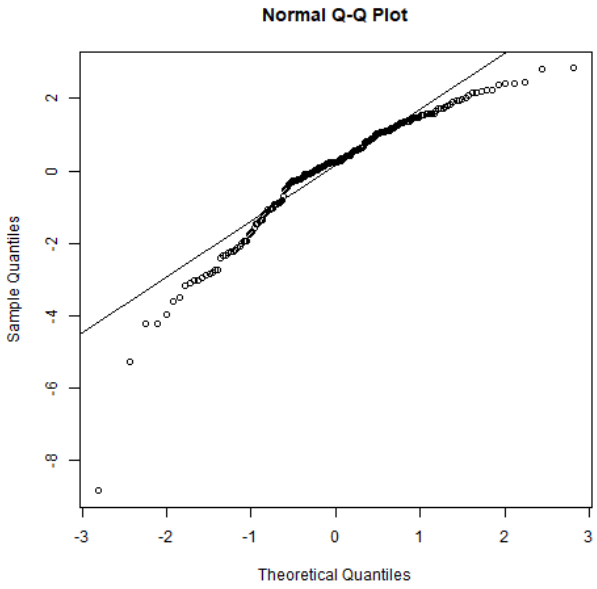

Figure 16: Gráfico QQ-plot.

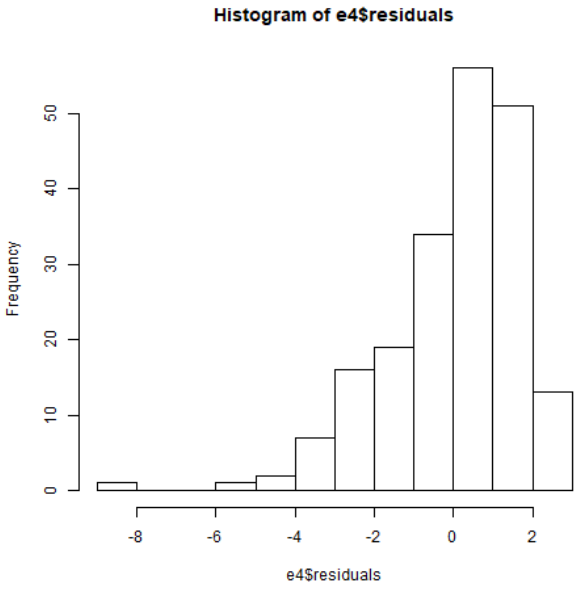

Figure 18: Histograma de los residuos.

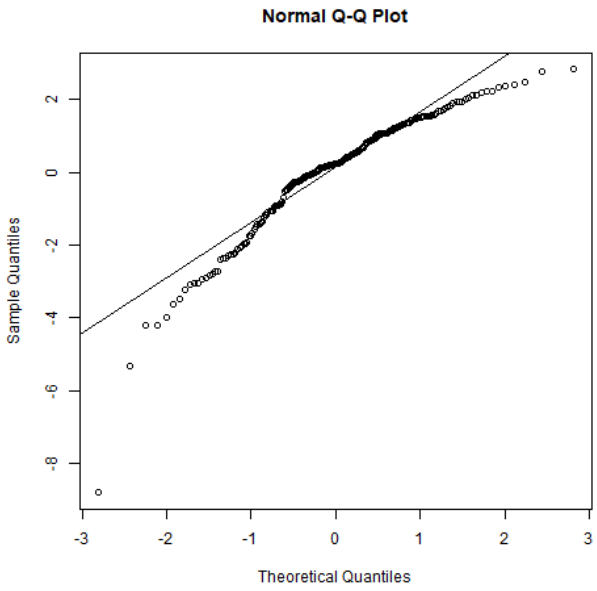

Figure 19: Gráfico QQ-plot.

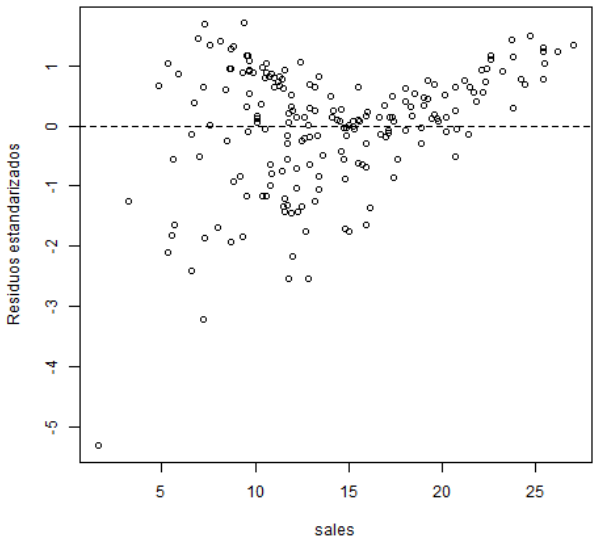

Figure 20: Gráfico Residuos estandarizados.