Trabajo 1

Estudiantes
Neila Sirley Perilla Perdomo
Julian David Vargas Bedoya
Ingrid Tatiana Yanangona Jojoa

3,8

Grupo 36

Docente
Mateo Ochoa Medina
Asignatura Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín
05 de Octubre de 2023

ÍNDICE

1. Pregunta

1.1. Modelo de regresión 3

1.2. Significancia de la regresión 3

1.3. Significancia de los parámetros 4

1.4. Interpretación de los parámetros 5

1.5. Coeficiente de determinación múltiple R2 6

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido 6

2.2. Estadístico de prueba y conclusión 7

3. Pregunta 3

3.1. Planteamiento de la pregunta 7

3.2. Prueba de hipótesis y prueba de hipótesis matricial 7

3.3. Estadístico de prueba 8

4. Pregunta 4

4.1. Supuestos del modelo 8

4.1.1. Normalidad de los residuales 8

4.1.2. Media 0 y Varianza constante 9

4.2. Observaciones extremas 11

4.2.1. Datos atípicos 11

4.2.2. Puntos de balanceo 12

4.2.3. Puntos influenciales 13

4.3. Conclusiones 15

Índice de figuras

1. Gráfico cuantil-cuantil y normalidad de residuales 9

2. Gráfico residuales estudentizados vs valores ajustados 10

3. Identificación de datos atípicos 11

4. Identificación de puntos de balanceo 12

5. Criterio distancias de Cook para puntos influenciales 13

6. Criterio Dffits para puntos influenciales 14

Índice de cuadros 1.

1. Tabla de valores de los parámetro ajustados 3

2. Tabla ANOVA para el modelo 4

3. Resumen de los coeficientes 5

4. Resumen tabla de todas las regresiones 6

5. Tabla de puntos de Balanceo 12

6. Criterio Dffits para puntos influenciales 14

1. **PREGUNTA 1:** 18 pt

Teniendo en cuenta la base de datos del grupo 36, en la cual hay 5 variables regresoras, denominadas por:

Variable

Y **:** Riesgo de infección
$\boldsymbol{X_1}$**:** Duración de la estadía
$\boldsymbol{X_2}$**:** Rutina de cultivos
$\boldsymbol{X_3}$**:** Número de camas
$\boldsymbol{X_4}$**:** Censo promedio diario
$\boldsymbol{X_5}$**:** Número de enfermeras

Se plantea que los datos pueden seguir un modelo de regresión lineal múltiple:

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i,\ \varepsilon_i\ iid \sim N(0, \sigma 2\ );\ 1 \leqslant i \leqslant 64$

1.1. Modelo de regresión

Al ajustar el modelo, se estiman los siguientes coeficientes:

Cuadro 1: Valor de los parámetros estimados

Valor del parámetro	
$\widehat{\beta}_0$	0.4562
$\widehat{\beta}_1$	0.2308
$\widehat{\beta}_2$	-0.0068
$\widehat{\beta}_3$	0.0479
$\widehat{\beta}_4$	0.0092
$\widehat{\beta}_5$	0.0018

3 pt

Por lo tanto, el modelo de regresión ajustada sería:

$\widehat{Y}_i = 0.4562 + 0.2308X_{1i} - 0.0068X_{2i} + 0.0479X_{3i} + 0.0092X_{4i} + 0.0018X_{5i}$

1.2. Significancia de la regresión

Para analizar la significancia de la regresión, se plantea el siguiente juego de hipótesis:

$$\begin{cases} H_0: \beta_1=\beta_2=\beta_3=\beta_4=\beta_5=0 \\ H_a: \text{Algún } \beta_j \text{ distinto de 0 para } j = 1,2,3,4,5 \end{cases}$$

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \sim f_{5,58} \qquad (1)$$

Teniendo la información anterior, se presenta la tabla ANOVA:

4 e t

Cuadro 2: Tabla ANOVA del modelo

	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	Valor-p
Regresión	61.9252	5	12.3850	12.977	1.75617e-08
Error	55.3542	58	0.9543		

Al observar la tabla ANOVA, observamos como el valor-p < 0.05 = α, rechaza H_0, se concluye que el modelo de RLM es significativo,es decir que la probabilidad promedio estimada para adquirir el riesgo de infección en el hospital está ~~afectando~~ significativamente al menos unas de las predictoras

siendo afectada

1.3 significancia de los parámetros

Ahora analizemos nuestros parametros teniendo en cuenta el siguiente juego de hipótesis:

$$\begin{cases} H_0: \beta_j = 0 \\ H_a: \text{Algún } \beta_j \neq 0 \text{ para } i= 1, 2, 3, 4, 5 \end{cases}$$

¿ β_0 ?

Donde el estadístico de prueba corresponde a:

$$Tj, 0 = \frac{\widehat{\beta j}}{SE\,(\widehat{\beta j})} H_0 \sim f_{58} \qquad \textbf{(2)}$$

A continuación, se presenta el siguiente cuadro de información de los parámetros,la cual permitirá determinar cuáles de ellos son significativos.

Cuadro 3: tabla de parámetros estimados

	Estimación $\hat{\beta}_j$	$se(\hat{\beta}_j)$	T_{0j}	Valor-P
β_0	0.456211816	1.5180845889	0.3005180	0.7648566133
β_1	0.230891380	0.0786215235	2.9367452	0.0047499431
β_2	-0.006830656	0.0276662693	-0.2468947	0.8058615079
β_3	0.047959706	0.0126402687	3.7941999	0.0003558754
β_4	0.009267275	0.0076233308	1.2156464	0.2290446102
β_5	0.001816767	0.0006373561	2.8504743	0.0060355734

5 pt

De la tabla de parámetros estimados, a un nivel de significancia $\alpha = 0.05$, se concluye que los parámetros individuales β_1, β_3 y β_5 son significativos ya que sus P-valores son menores α. Además, se encuentra que β_0, β_2, y β_4 son individualmente no significativos en presencia de los demás parámetros.

1.4 Interpretación de los parámetros

3 pt

Ahora podemos hacer el siguiente análisis de cada variable:

- $\hat{\beta}_0$**= 0.456211816** *(Intercepto)*: como $X_i = 0 \notin [X_{i,min}, X_{i,max}] \forall i$ entonces este valor no es interpretable.
- $\hat{\beta}_1$= **0.230891380** **(***Duración de la estadía***)**: indica que por cada unidad de aumento en la duración de la estadía, el promedio del riesgo de infección aumenta en 0.230891380 unidades, cuando las demás variables se mantienen fijas
- $\hat{\beta}_2$= **-0.006830656** *(Rutina de cultivos)*: El parámetro $\hat{\beta}_2$, no podemos interpretar nada, ya que no es significativo.
- $\hat{\beta}_3$=**0.047959706** *(Número de camas)*: El parámetro $\hat{\beta}_3$, por cada unidad que aumente el número de camas, la probabilidad promedio de adquirir una infección aumenta 0.047959706 unidades mientras las demás variables regresoras permanecen constantes.
- $\hat{\beta}_4$=**0.009267275** *(Censo promedio diario)*: El parámetro $\hat{\beta}_4$, no podemos interpretar nada, ya que no es significativo
- $\hat{\beta}_5$=**0.001816767** *(Número de enfermeras)*: Por cada unidad que aumente el número de enfermeras , la probabilidad promedio de adquirir una infección aumenta

0.001816767 unidades mientras las demás variables regresoras permanecen constantes.

1.5. Coeficiente de determinación múltiple R^2

3pt

con la tabla ANOVA podemos calcular usando la siguiente fórmula:

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE}$$

$$R^2 = \frac{61.9252}{61.9252 + 55.3542} = 0.5280143$$

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5280143$ aproximadamente el 52.801% de la variabilidad total en el riesgo de infección es explicada por el modelo de regresión propuesto en el trabajo.

2. PREGUNTA 2

1pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con el P-valor más alto en el modelo fueron X2, X4, X5. Así, planteamos las siguientes prueba de hipótesis:

será el más bajo

$$\begin{cases} H_0: \beta_2 = \beta_4 = \beta_5 = 0 \\ H_a: \beta_j \neq 0, \quad \text{para } j = 2,4,5 \end{cases}$$

A partir de la tabla de todas las regresiones, se construye la siguiente tabla donde se evidencia la suma de cuadrados del error del modelo completo y reducido lo cual nos permitirá hacer cálculos posteriores.

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	***Covariables en el modelo***
Modelo completo	55.3542	*X1 X2 X3 X4 X5*
Modelo reducido	63.820	*X1 X3*

Luego, un modelo reducido para la prueba de significancia del subconjunto es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon_i, \; \varepsilon_i \; iid \sim N(0, \sigma^2); \quad 1 \leq i \leq 64$$

2.2. Estadístico de prueba y conclusión

Al menos fueron consecuentes con el error

Se construye el estadístico de prueba como:

$$F0 = \frac{((SSE(\beta0,\beta1,\beta3)-SSE(\beta0,\beta1,\beta2,\beta3,\beta4,\beta5))/3}{MSE(\beta0,\beta1,\beta2,\beta3,\beta4,\beta5)} \quad H_0 \sim f_{3,58}$$

$$Fo = \frac{(63.820-55.3542/3}{0.9543} \quad H_0 \sim f_{3,58} \qquad \textbf{(3)}$$

1pt

$$Fo = 2.957072$$

→ 2,95 > 2,7

El f_0 obtenido con el cuantil $f_{0.95,3,58}$ = 2.763552, se puede observar que $F_0 < f_{0.95,3,58}$, teniendo en cuenta los datos la prueba de hipótesis obtenemos la información se rechaza la hipótesis nula, por que se acepta la hipotesis alternativa, se llega a una conclusion que al menos uno de los parámetros X2, X4, X5 no son significativos por lo tanto, puede ser descartado y el modelo logra explicar el riesgo en el hospital. ✗

6pt

Dijeron Fo < f, por lo que no rechazaban

↳ Falla teórica gigante

4pt

3. PREGUNTA

3.1 Planteamiento del problema.

El encargado del estudio desea realizar una investigación adicional, con el fin de seguir estudiando la eficacia en el control de infecciones en los hospitales desea comparar si las variables (X1, X5) son similares en sus efectos sobre la respuesta, y si (X3, X4) presentan similitudes ¿Será que estos conjuntos de variables si tienen el mismo efecto sobre la respuesta como piensa el encargado?

3.2 Prueba de hipótesis lineal general.

Con el fin de responder esta pregunta se planteó una prueba de hipótesis lineal general de la forma H0 : Lβ = 0

$$H_0 : \beta_1 = \beta_5,\ \beta_3 = \beta_4 \quad vs. \quad H_1 : \beta_1 \neq \beta_5 \text{ ó } \beta_3 \neq \beta_4$$

Podemos Plantear La Hipótesis Nula De Forma Matricial Como:

$$H_0 : \begin{cases} \beta_1 - \beta_5 = 0 \\ \beta_3 - \beta_4 = 0 \end{cases}$$

Asís es más claro verla laforma Ho :Lβ=0 donde:

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix}$$

2pt

$$\bar{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \end{bmatrix}^T$$

Modelo reducido.

Con la información de las matrices tenemos que el modelo reducido es:

$$Y = \beta_0 + \beta_1(X_1 + X_5) + \beta_3(X_3 + X_4) + \varepsilon$$

supuestos

0pt

Estadístico de prueba.

Con esta información y los temas vistos en clase se tiene que el estadístico de prueba F0 es:

$$F_0 = \frac{\frac{SSE(MR) - SSE(MC)}{Gl_{(MR)} - Gl_{(MC)}}}{MSE_{(MC)}}$$

Donde MC representa el modelo completo y MR al modelo reducido, ahora si tomamos la información de

la tabla ANOVA:

	Sum_of_Squares	DF	Mean_Square	F_Value	P_value
Model	61.9252	5	12.385043	12.977	1.75617e-08
Error	55.3542	58	0.954382		

2pt

El estadístico de prueba quedaría como:

$$F_0 = \frac{\frac{SSE(MR) - 55.3542}{61 - 58}}{0.954382} = \frac{SSE(MR) - 55.3542}{2.863146}$$

PREGUNTA 4

14,5pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Para validar este supuesto de normalidad , se debe hacer una prueba de hipótesis en donde se logra determinar el conjunto de datos que proviene de una distribución normal y la otra es la prueba gráfica cuantil-cuantil con el fin de afirmar y negar la suposición.

$$\begin{cases} H_0 : \varepsilon_i \sim \text{Normal} \\ H_1 : \varepsilon_i \nsim \text{Normal} \end{cases}$$

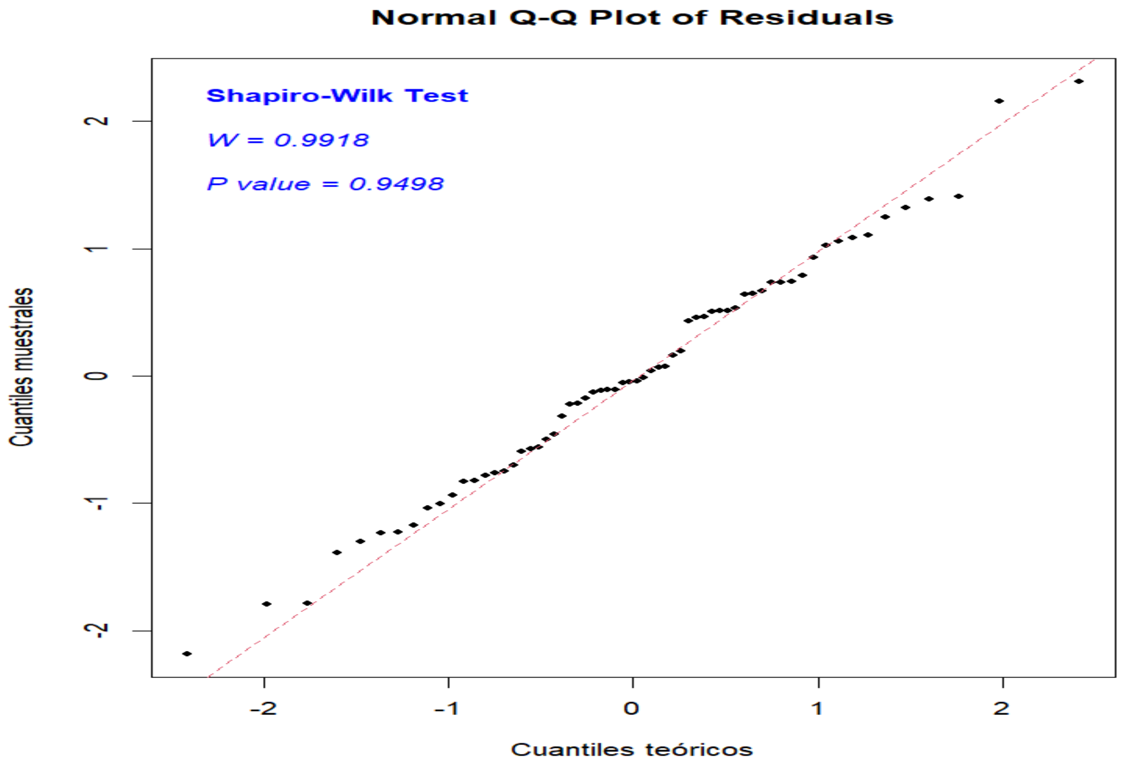


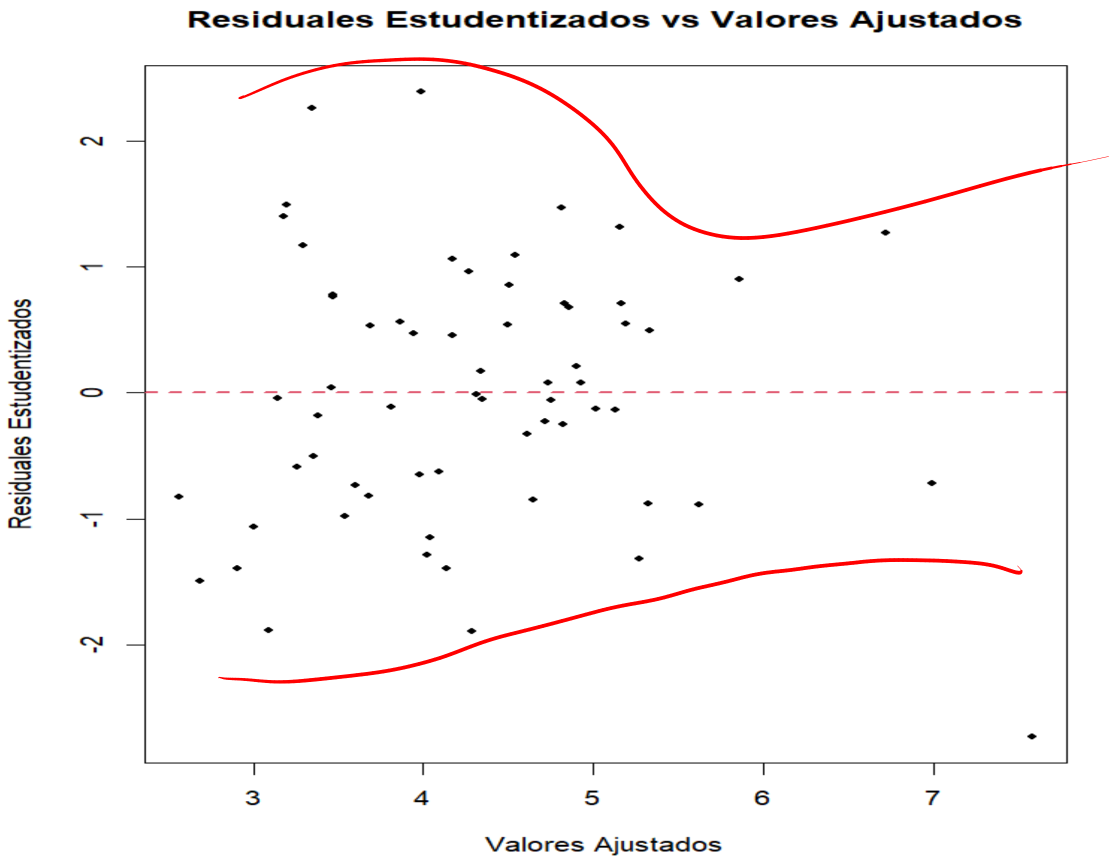
Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

3pt

Analizando la prueba de hipótesis de shapiro-wilk, observamos que el valor P es aproximadamente 0.9918, por lo que a un nivel de significancia de α = 0.05, no se rechaza la hipótesis nula dado que el valor P es mucho mayor, es decir que según el test de Shapiro-Wilk, los resultados se distribuyen normal. Por otro lado, al ver la *gráfica: cuantil-cuantil y normal de residuales* el patrón de los residuales no sigue la recta de ajuste de la distribución de los residuales a una distribución normal, además de que se presentan patrones irregulares en los datos. para finalizar el supuesto de normalidad NO se cumple ya que no hay buen ajuste.

Ojo que uso como tal no es un supuesto.

4.1.2. Varianza constante



2pt

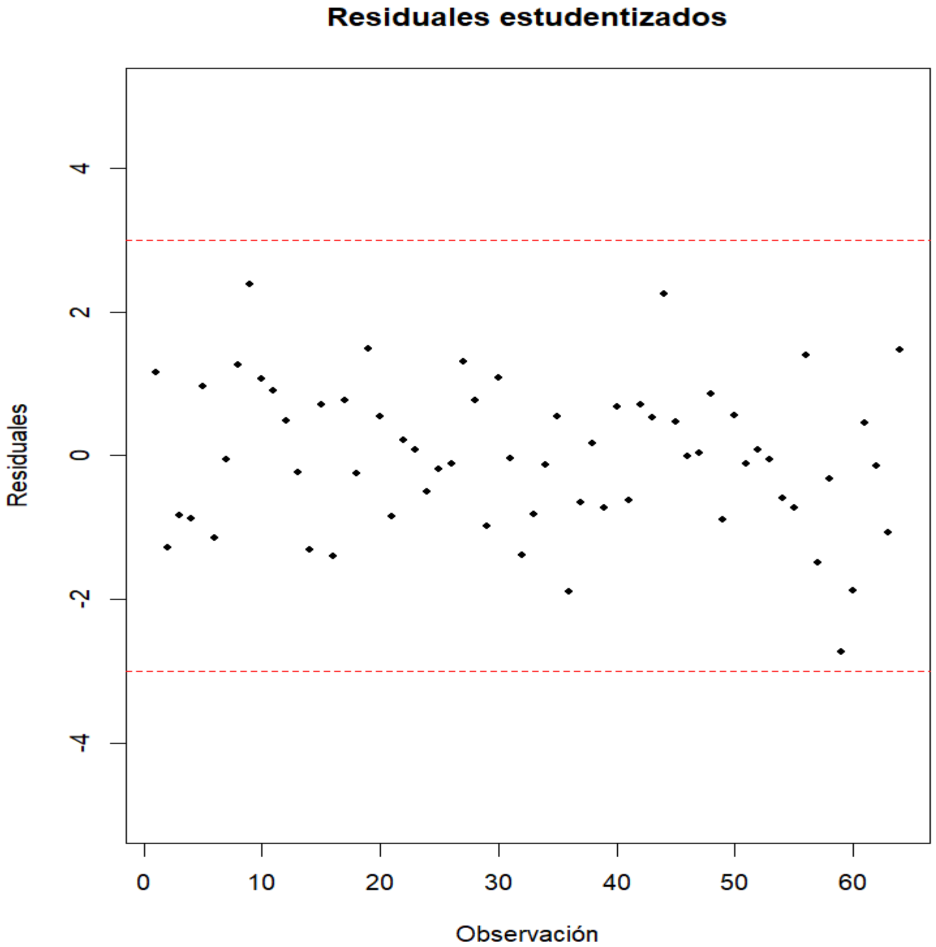
Figura 2: Gráfico residuales estudentizados vs valores ajustados

De la anterior gráfica concluimos que la media no se aproxima a 0 y la varianza no es constante, esto se evidencia en la falta de una dispersión uniforme de los datos, ya que la varianza muestra un patrón decreciente.

→ sí lo hace, de hecho residuales estudentizados siempre tienen media 0

4.2. Verificación de las observaciones

4.2.1. Datos atípicos



3 pt

Figura 3: Identificación de datos atípicos

Se observa en la figura 3 que no existen datos atípicos bajo el criterio $|r_{estud}| > 3$

4.2.2. Puntos de balanceo

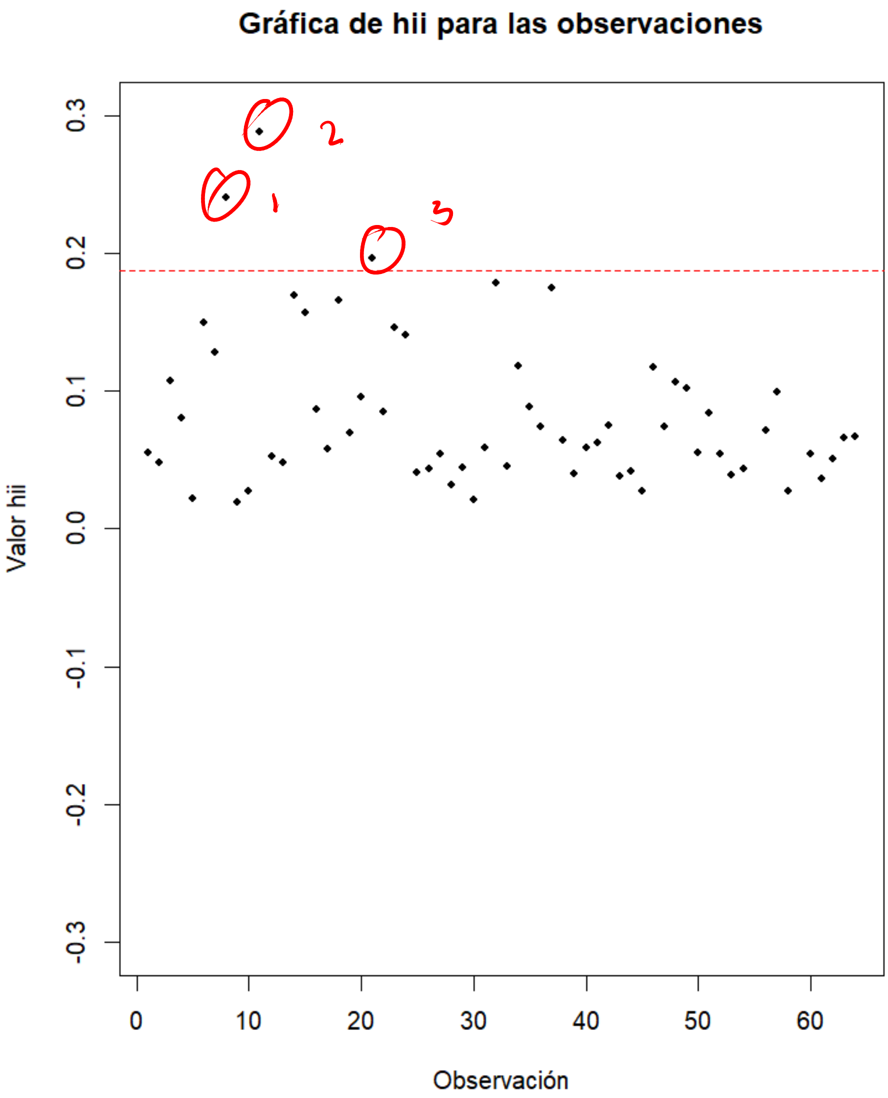


Figura 4: Identificación de puntos de balanceo

Cuadro 5: Tabla de puntos de balanceo

	Errores Estudentizados	*D.Cook*	*Valor hii*	*DFFITS*
8	1.2699	0.0851	0.2404	0.7183
11	0.8987	0.0544	0.2879	0.5704
21	-0.8511	0.0295	0.1962	0.4195
55	-0.7212	0.0870	0.5008	-0.7193
59	-2.7324	0.6131	0.3301	-2.0370

5, y solo se ven 3 en gráfica

→ Falso, se ven 3

Analizando la gráfica de observaciones vs valores hii, se pueden identificar cinco observaciones que cumplen con el criterio definido en los puntos de balanceo, el cual es $h_{ii} > 2\frac{p}{n}$, donde $h_{ii} = 2\,\frac{6}{64} = 0.19$, siendo "p" el número de parámetros y "n" el número de datos. De acuerdo con los datos presentados en la tabla, estamos hablando de las siguientes observaciones: la primera es la observación número 8 con un valor h_{ii}= 0.2404, la segunda es la observación número 11 con un valor h_{ii}=0.2879, la tercera es la observación número 21 con un valor h_{ii}=0.1962, la cuarta es la observación número 55 con un valor h_{ii}=0.5008 y por último la observación número 59 con un valor h_{ii}=0.3301, ya que estos cumplen con la desigualdad $h_{ii} > 2\frac{p}{n}$. Estos puntos de balanceo, a pesar de que posiblemente no afecte los coeficientes de regresión, si puede afectar las estadísticas de resumen como el R^2 y los errores estándar de los coeficientes estimados. 1,5 pt

4.2.3. Puntos influenciales

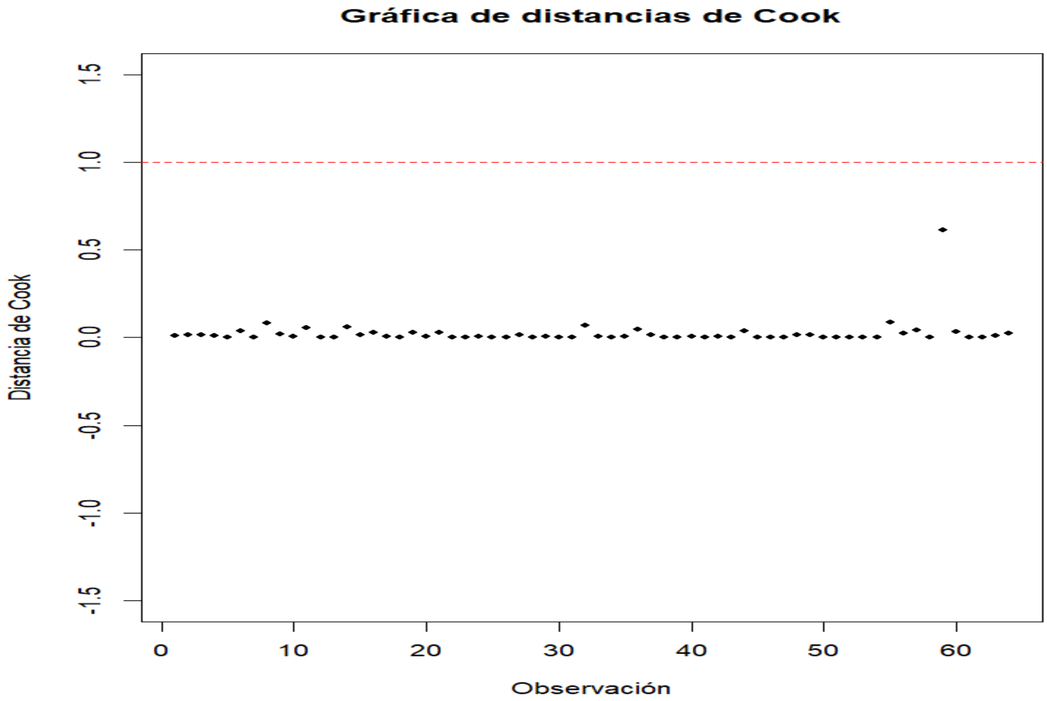


Figura 5: Criterio distancias de Cook para puntos influenciales

Al hacer el análisis de la evaluación del criterio de distancias de Cook, en el cual para cualquier punto que cumpla la condición $D_i > 1$, se considera una observación influencial, observamos que ningún valor cumple con esta condición. Es decir, que la influencia de cada una de las observaciones sobre el vector de parámetros no es lo bastante significativa como para clasificarlo como un punto influencial.

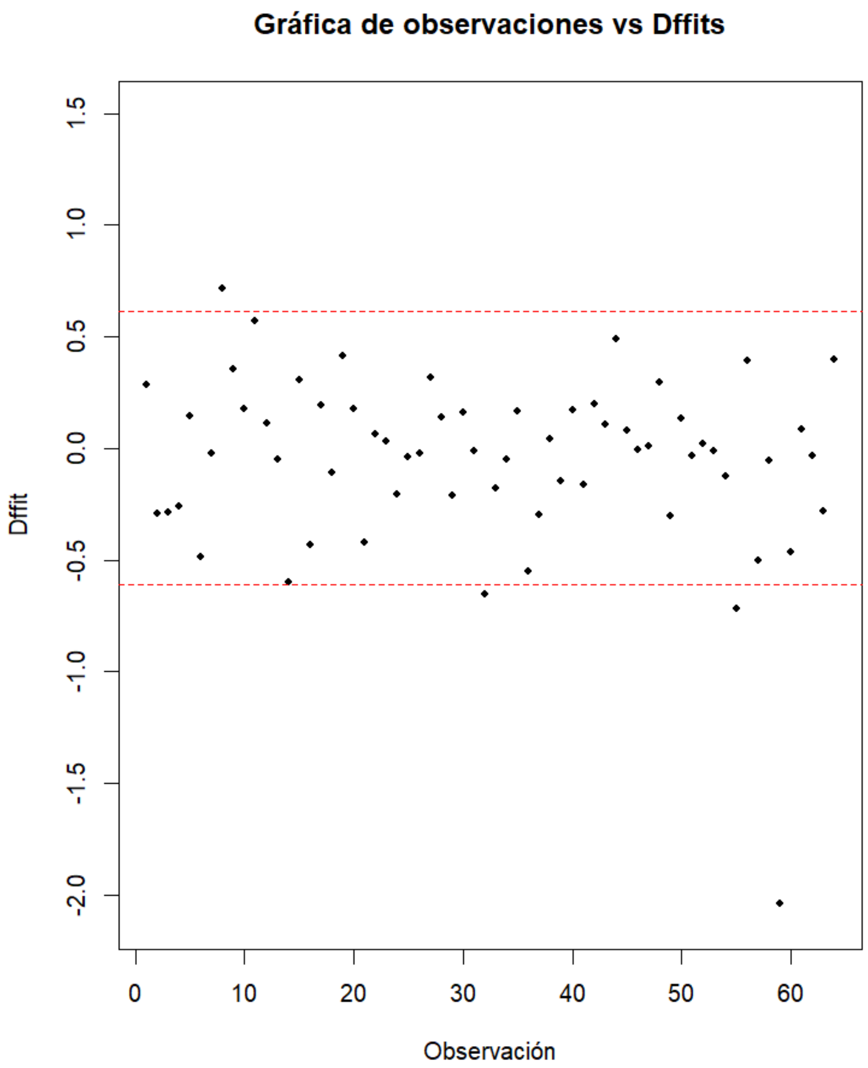


Figura 6: Criterio Dffits para puntos influenciales

Cuadro 6: Tabla de puntos influenciales

	Errores Estudentizados	***D.Cook***	***Valor hii***	***DFFITS***
8	1.2699	0.0851	0.2404	0.7183
32	-1.3918	0.0702	0.1785	-0.6542
55	-0.7212	0.0870	0.5008	-0.7193
59	-2.7324	0.6131	0.3301	-2.0370

4pt

Bajo el criterio de la prueba Dffits, se obtuvo la gráfica previamente mostrada, y a partir de esta gráfica, podemos concluir que las observaciones 8, 32,55 y 59 cumplen con el criterio definido por la prueba Dffits, la cual establece que para cualquier observación cuyo $|Dffit|$

> $2\sqrt{\frac{p}{n}}$ es un punto influential, donde $2\sqrt{\frac{6}{64}} = 0.6124$, por lo cual dichas observaciones son influenciales. Estos puntos influenciales tienen un gran efecto en el modelo de regresión, ya que ejercen un fuerte impacto sobre los coeficientes de regresión ajustados, lo que puede hacer que este no sea el más adecuado para ajustarse a los datos proporcionados.

1pt

4.3. Conclusión

Para concluir, el modelo de regresión lineal múltiple no cumple la validez, de manera más óptima, debido a que los errores del modelo no tienden a una distribución normal (supuestos de error) , el cual fue comprobado mediante el criterio gráfico. Además, se comprobó la presencia de puntos influenciales, los cuales tienen un gran impacto sobre los coeficientes de regresión ajustados, lo que a su vez, provoca estimaciones de la variable respuesta que se alejan de los valores reales o esperados. Debido a lo anterior se plantea y considera que los resultados de este modelo no deben tomarse como válidos, ya que es necesario reconstruir el modelo sin las observaciones atípicas, de balanceo e influenciales que modifican los resultados y a su vez los supuestos del modelo.

no necesariamente son lo que lo hace no válido

validez tampoco se cumple por varianza no cte