9,3

Trabajo 1

Estudiantes

Anderson Bedoya Ciro
Alejandro Giraldo Quiceno
Juan Felipe Calle Becerra

Equipo #46

Docente

Francisco Javier Rodríguez Cortés

Asignatura

Estadística II

Sede Medellín
30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

20 pt

La base de datos 46 incluye cinco variables regresoras que se denominan:

Y: Riesgo de infección

X_1: Duración de la estadía

X_2: Rutina de cultivos

X_3: Número de camas

X_4: Censo promedio diario

X_5: Número de enfermeras

Se formula el modelo de regresión múltiple:

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 65$$

1.1. Modelo de regresión

Al ajustar el modelo, se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-1.9215
β_1	0.2191
β_2	0.0331
β_3	0.0401
β_4	0.0190
β_5	0.0008

3pt

Por lo tanto, se construye el modelo de regresión ajustado:

$$\hat{Y}_i = -1.9215 + 0.2191 X_{1i} + 0.0331 X_{2i} + 0.0401 X_{3i} + 0.019 X_{4i} + 8 \times 10^{-4} X_{5i}$$

1.2. Significancia de la regresión

5pt

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

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

El estadístico de prueba es el siguiente:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,59} \tag{1}$$

Ahora, se presenta la tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	Suma Cuadrática	Grados de libertad.	Media Cuadrática	F_0	P-valor
Regresión	68.1201	5	13.62401	12.5248	2.67716e-08
Error	64.1778	59	1.08776		

En este caso, el valor P es aproximadamente igual a 0, lo que indica que es muy poco probable que los datos observados se hayan producido si la hipótesis nula fuera cierta. La hipótesis nula en este caso es que todos los coeficientes de regresión β_j son iguales a 0, lo que significa que ninguna de las variables independientes tiene un efecto significativo sobre la variable dependiente. Dado que el valor P es muy pequeño, se rechaza esta hipótesis y se acepta la hipótesis alternativa de que al menos uno de los coeficientes β_j es diferente de 0 y por lo cual hay algún parámetro significativo.

1.3. Significancia de los parámetros

La siguiente tabla contiene información de los parámetros que permitirán determinar cuáles de ellos son significativos.

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-1.9215	1.5957	-1.2042	0.2333
β_1	0.2191	0.1005	2.1815	0.0331
β_2	0.0331	0.0301	1.0998	0.2759
β_3	0.0401	0.0134	2.9961	0.0040
β_4	0.0190	0.0071	2.6658	0.0099
β_5	0.0008	0.0009	0.9134	0.3647

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_1 β_3 y β_4 son significativos, pues sus P-valores son menores a α.

1.4. Interpretación de los parámetros

$\hat{\beta}_1$**:** El valor estimado es 0.2191, lo que significa que un aumento de un día en la duración promedio de la estadía se asocia con un aumento del 21.91 % en la probabilidad promedio del riesgo de infección.

$\hat{\beta}_3$**:** El valor estimado es 0.0401, lo que significa que un aumento en el número promedio de camas en el hospital se asocia con un aumento del 4.01 % en la probabilidad promedio del riesgo de infección.

$\hat{\beta}_4$**:** El valor estimado es 0.0190, lo que significa que un aumento en el número promedio de pacientes en el hospital por día se asocia con un aumento del 1.9 % en la probabilidad promedio del riesgo de infección.

1.5. Coeficiente de determinación múltiple R^2

Se calcula de la siguiente manera:

$$R^2 = \frac{SSR}{SST}; SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5149$, lo que significa que aproximadamente el 51.49 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con el P-valor más alto en el modelo fueron X_1, X_2, X_5. Por lo tanto, se pretende realizar la siguiente prueba de hipótesis a través de la tabla de todas las regresiones posibles:

$$\begin{cases} \text{H}_0 : \beta_1 = \beta_2 = \beta_5 = 0 \\ \text{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 2, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	Suma Cuadrática del Error	Covariables en el modelo
Modelo completo	64.178	X1 X2 X3 X4 X5
Modelo reducido	80.273	X3 X4

Se crea un modelo reducido para la prueba de significancia del subconjunto:

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i \; ; \; \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, 2 \ldots$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$F_0 = \frac{(SSE(\beta_0, \beta_3, \beta_4) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4 \beta_5))/3}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4 \; \beta_5)} \overset{H_0}{\sim} F_{3,59}$$

$$= \frac{80.273 - 64.178/3}{64.178/59} \tag{2}$$

$$= 4.932$$

Ahora, comparando el F_0 con $f_{0.95,3,59} = 2.7608$, se puede ver que $F_0 > f_{0.95,1,45}$ entonces, el subconjunto es significativo, no es posible descartar las variables del subconjunto, al menos una es distinta de 0.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si ... por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_2 = \beta_3; \; \beta_4 = \beta_5 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

reescribiendo matricialmente:

$$\begin{cases} \mathrm{H}_0 : \mathbf{L}\underline{\beta} = \underline{\mathbf{0}} \\ \mathrm{H}_1 : \mathbf{L}\underline{\beta} \neq \underline{\mathbf{0}} \end{cases}$$

Con $\mathbf{L}$ está dada por:

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_3 X^*_{3i} + \beta_4 X_{4i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 65$$

Donde $X^*_{3i} = X_{2i} + X_{3i}$ y $X^*_{4i} = X_{4i} + X_{5i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,59} \tag{3}$$

$$F_0 = \frac{(SSE(MR) - 64.178/2}{64.178/59} \overset{H_0}{\sim} f_{2,59} \tag{4}$$

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Para la validación de este supuesto, se planteará la siguiente prueba de hipótesis shapiro wilk, acompañada de un gráfico cuantil-cuantil:

$$\begin{cases} \text{H}_0 : \varepsilon_i \sim \text{Normal} \\ \text{H}_1 : \varepsilon_i \nsim \text{Normal} \end{cases}$$

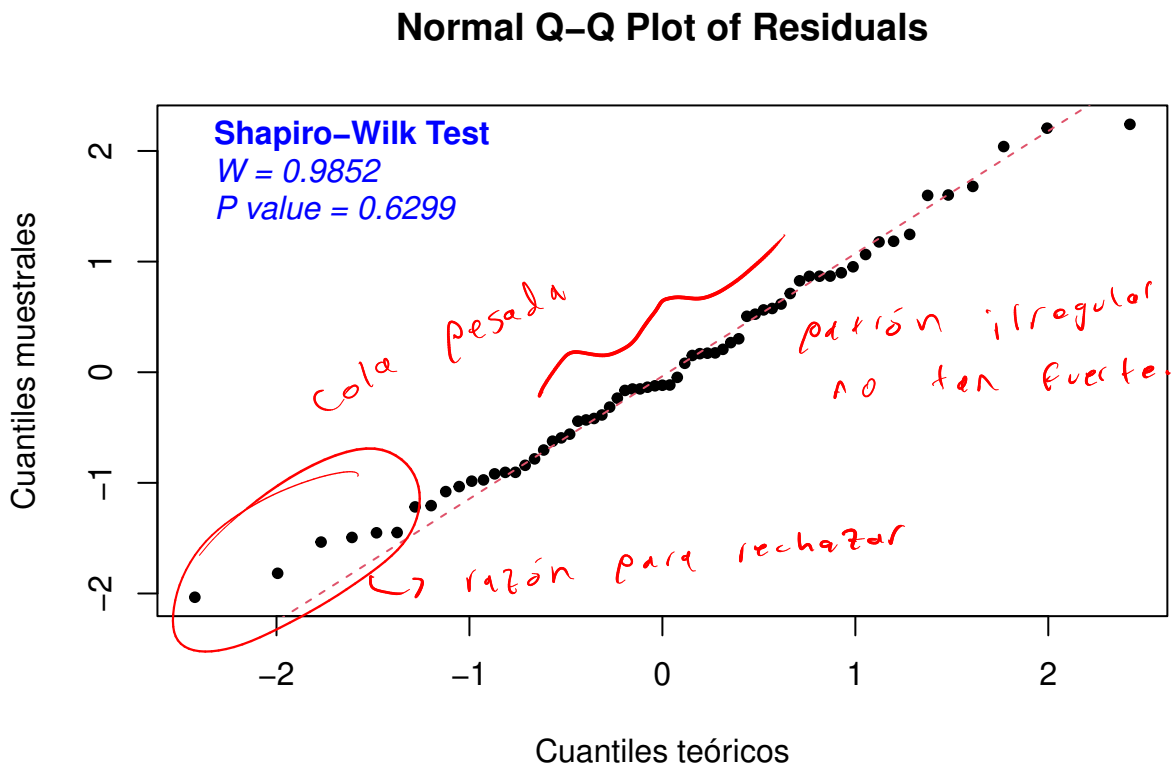

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

De la prueba se tiene un estadistico w=0.9852, este indica que los datos se ajustan bien a una distribucion normal. Al ser el P-valor aproximadamente igual a 0.6299 y teniendo en cuenta que el nivel de significancia $\alpha = 0.05$, el P-valor es mucho mayor y por lo tanto, no se rechazaría la hipótesis nula, es decir que los datos distribuyen normal con ~~media~~ μ y varianza σ^2, sin embargo, al observar la gráfica de comparación de cuantiles, se pueden observar colas más pesadas y patrones irregulares, lo que sugiere que el análisis gráfico tiene más poder que el valor P. Por lo tanto, se rechaza la hipótesis de que los datos se distribuyen normalmente. Ahora se procederá a evaluar si la varianza cumple con el supuesto de ser constante.

- No están probando media constante μ y Var cte σ^2
- Se los vetgo, sin embargo la evidencia no era tan fuerte

4.1.2. Varianza constante

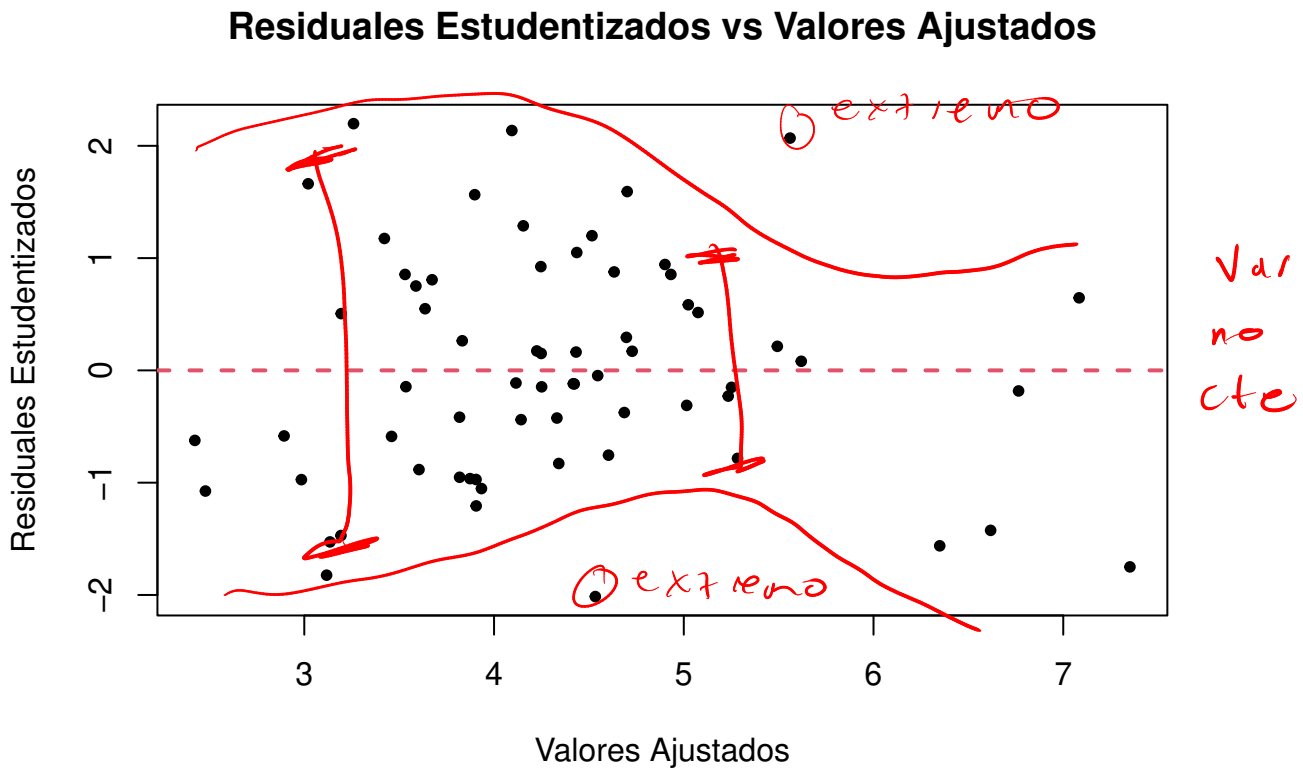

Figura 2: Gráfico residuales estudentizados vs valores ajustados

No se detectan patrones que sugieran una varianza no constante en el gráfico de residuales estudentizados contra los valores ajustados. Por lo tanto, se acepta este supuesto al no contar con evidencia suficiente para rechazarlo. La media es cero, como se puede ver en el gráfico.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

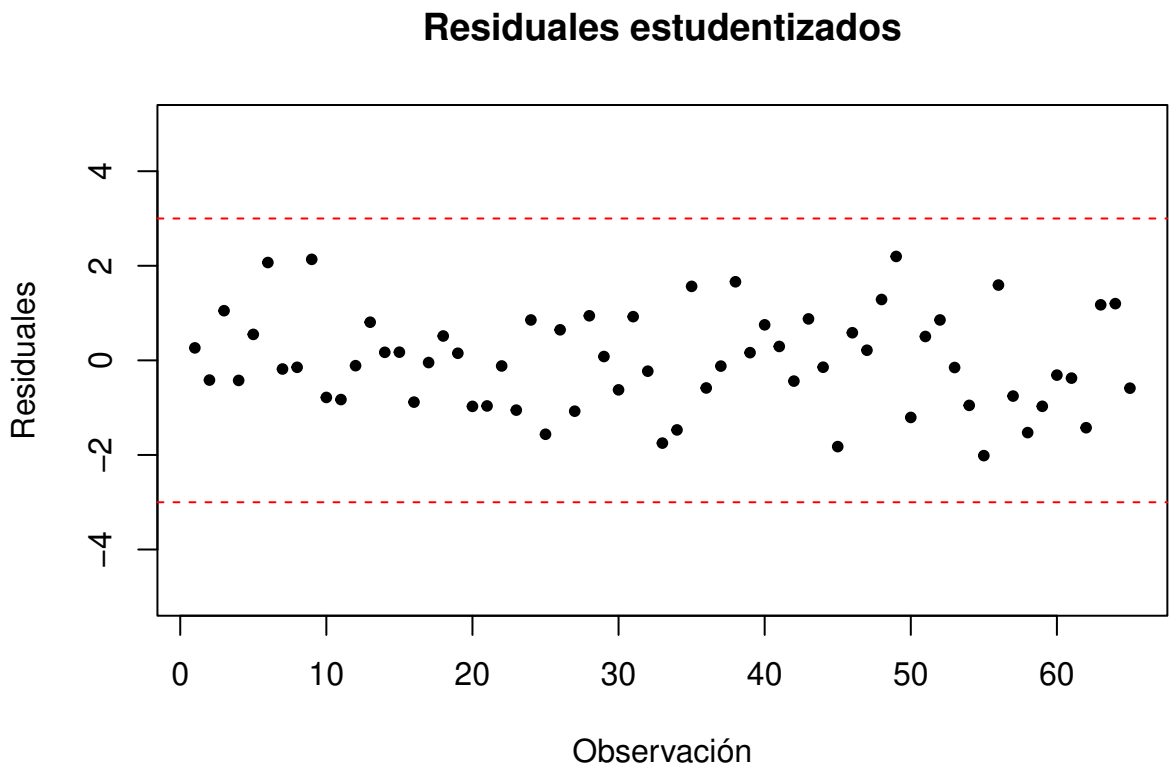

Figura 3: Identificación de datos atípicos

Como se puede observar en la gráfica anterior, no hay datos atípicos en el conjunto de datos pues ningún residual estudentizado sobrepasa el criterio de $|r_{estud}| > 3$.

4.2.2. Puntos de balanceo

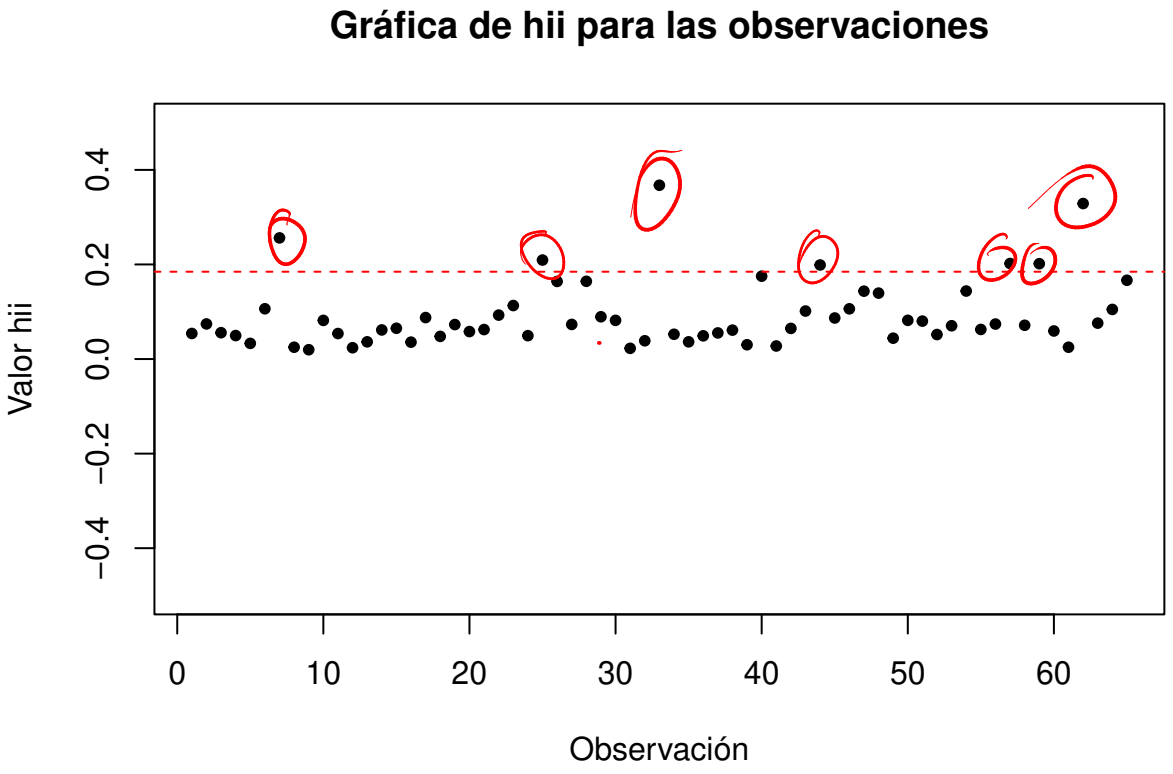

Figura 4: Identificación de puntos de balanceo

Cuadro 5: Tabla de puntos de balanceo

	Residuales estudentizados	Cooks.D	hii.value	Dffits
7	-0.1829	0.0019	0.2562	-0.1064
25	-1.5624	0.1077	0.2093	-0.8139
33	-1.7500	0.2966	0.3675	-1.3583
44	-0.1452	0.0009	0.1989	-0.0718
57	-0.7548	0.0240	0.2020	-0.3784
59	-0.9714	0.0397	0.2016	-0.4879
62	-1.4246	0.1657	0.3288	-1.0061

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 0.184615$, se puede apreciar que existen 7 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{p}{n}$, los cuales son los presentados en la tabla.

4.2.3. Puntos influenciales

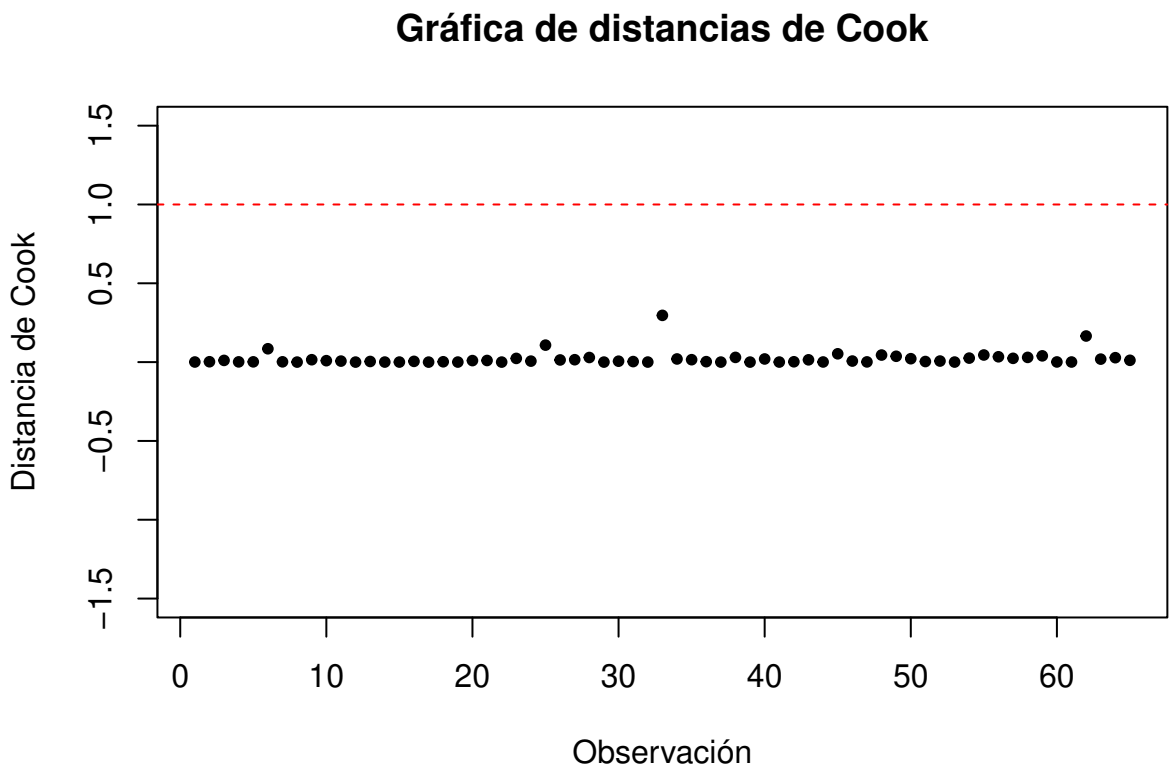

Figura 5: Criterio distancias de Cook para puntos influenciales

De la anterior gráfica se puede concluir bajo el criterio de Cook que no hay puntos influenciales en el conjunto de datos, es decir, no hay puntos que tengan un efecto desproporcionado en la regresión y que puedan afectar la presición del modelo.

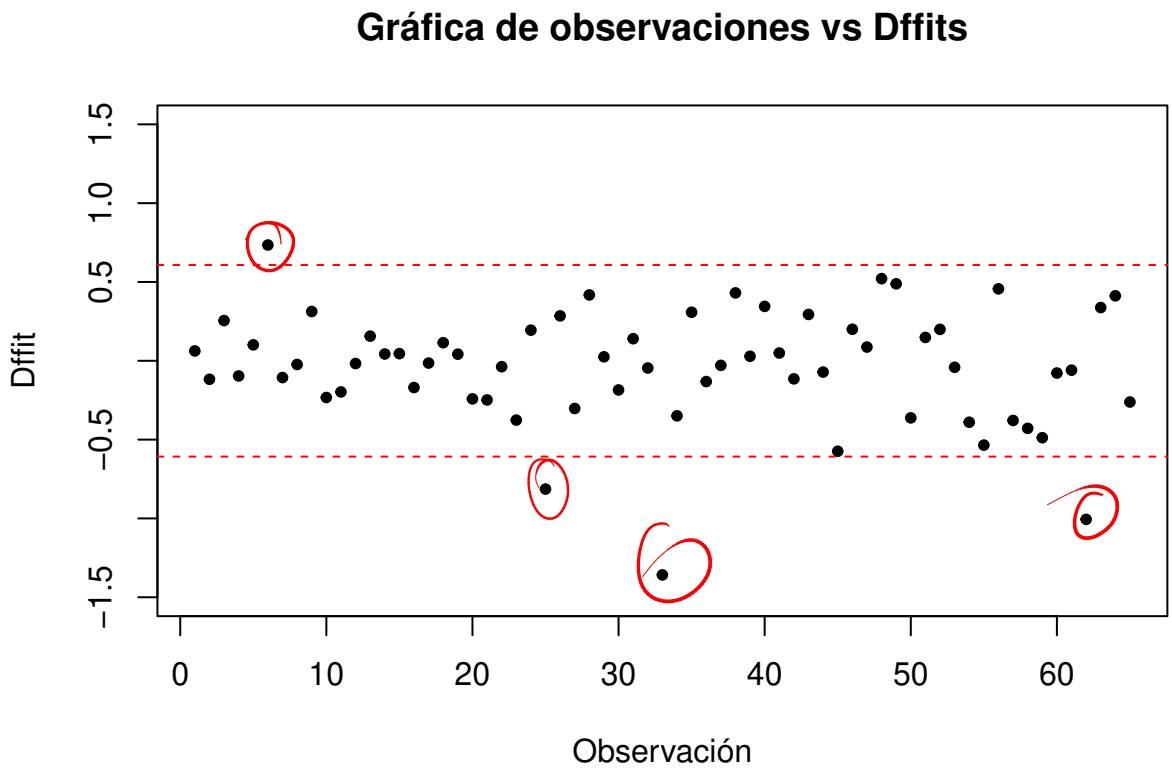

Figura 6: Criterio Dffits para puntos influenciales

Cuadro 6: Tabla de

	Residuales estudentizados	Cooks.D	hii.value	Dffits
6	2.0688	0.0848	0.1063	0.7345
25	-1.5624	0.1077	0.2093	-0.8139
33	-1.7500	0.2966	0.3675	-1.3583
62	-1.4246	0.1657	0.3288	-1.0061

Como se puede ver en la tabla anterior, son 4 puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, es un punto influencial. Cabe destacar también que con el criterio de distancias de Cook, en el cual para cualquier punto cuya $D_i > 1$, es un punto influencial, ninguno de los datos cumple con serlo.

4.3. Conclusión

Los supuestos del modelo son importantes para validar la regresión lineal múltiple. En este caso, se verificó que los residuales siguen una distribución normal y que la varianza es

constante. Además, se identificaron los puntos de balanceo y los puntos influenciales para verificar la validez de las observaciones.

La normalidad de los residuales resulta muy importante porque garantiza que el modelo es adecuado para los datos. En este caso, se verificó que los residuales siguen una distribución normal mediante el gráfico cuantil-cuantil y la prueba de normalidad de Shapiro-Wilk.

La varianza constante garantiza que el modelo es adecuado para los datos. En este caso, se verificó que la varianza es constante mediante el gráfico residuales estudentizados vs valores ajustados. → No lo es, pero son congruentes al menos :D

Se indica con los puntos de balanceo si una observación tiene un efecto significativo en el modelo. Se identificaron siete puntos de balanceo mediante el gráfico de hii

Los puntos influénciales indican que una observación tiene un efecto significativo en el modelo. No se identificaron puntos influenciales mediante el criterio distancias de Cook y se encontraron 4 puntos influénciales con el criterio Dffits.

Se puede afirmar que el modelo de regresión lineal múltiple es adecuado para los datos y que las observaciones son válidas. Sin embargo, se deben tener en cuenta los puntos de balanceo y los puntos influenciales al interpretar los resultados del modelo.

11
✓

Nooo, adecuado en qué sentido, cómo así que observaciones válidas?

Modelo no es válido, aunque según sus conclusiones debían decir que sí