4,3

Trabajo 1

Estudiantes

Alejandra Upegui Pajarito
Pedro Aristizabal Alzate
Ramiro Cardenas Mendoza

Equipo # 15

Docente

Julieth Veronica Guarin Escudero

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín

30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

16,5 pt

A partir de la base de datos dada, formulamos el siguiente modelo de regresión lineal multiple, con 5 variables predictoras:

¿Quiénes son $Y, X_1, X_2, X_3, X_4, X_5$?

$$Y_i = \beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i,\ \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ \ 1 \leqslant i \leqslant 50$$

1.1. Modelo de regresión

3 pt

Luego, ajustamos el modelo y obtenemos los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-0.3686
β_1	0.3056
β_2	-0.0007
β_3	0.0486
β_4	0.0104
β_5	0.0013

✓

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.3686 + 0.3056X_{i1} - 7 \times 10^{-4} X_{i2} + 0.0486X_{i3} + 0.0104X_{i4} + 0.0013X_{i5}$$ ✓

$i = 1, 2, \ldots, 50$

1.2. Significancia de la regresión

9 pt

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

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

El cual podemos verificar mediante el estadístico de prueba:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,44} \tag{1}$$ ✓

Ahora, se presenta la tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	39.1580	5	7.83160	6.82004	8.70317e-05
Error	50.5262	44	1.14832		

✓

si lo ponen así es individual

A un nivel de significancia $\alpha = 0.05$, y con el P-valor dado en la tabla, observamos que $vp < \alpha$, por lo que rechazamos la hipotesis nula: $\beta_j = 0$ con $1 \leqslant j \leqslant 5$, aceptando la hipótesis alternativa en la que almenos un $\beta_j \neq 0$, por lo tanto validamos que la regresión es significativa.

1.3. Significancia de los parámetros

6 pt

En la siguiente tabla se darán todos los valores de los parámetros del modelo que nos permitan analizar su significancia individual.

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-0.3686	2.1634	-0.1704	0.8655
β_1	0.3056	0.1212	2.5209	0.0154
β_2	-0.0007	0.0405	-0.0166	0.9868
β_3	0.0486	0.0215	2.2561	0.0291
β_4	0.0104	0.0091	1.1375	0.2615
β_5	0.0013	0.0008	1.7185	0.0927

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

1.4. Interpretación de los parámetros

2 pt

la probabilidad promedio

?

$\hat{\beta}_1$: Por cada día que aumente la duración de la estadía en el hospital, en promedio el riesgo de infección aumenta en 0.3056 unidades, cuando los valores en las demas predictoras se mantienen fijos.

$\hat{\beta}_3$: Por cada unidad que aumente el número de camas en el hospital, en promedio el riesgo de infección aumenta en 0.0486 unidades, cuando los valores en las demas predictoras se mantienen fijos.

1.5. Coeficiente de determinación múltiple R^2

1,5 pt

Hacemos uso de las sumas cuadráticas de la regresión y el error para calcular el coeficiente de determinación $R^2 = SSR/SST = \cancel{0.3737}$, lo que significa que aproximadamente el ~~37.37 %~~ de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto.

Su $R^2 = 0,9366$

2. Pregunta 2

9pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con el P-valor más alto en el modelo fueron X_2, X_4, X_5, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0: \beta_2 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_1: \text{Algún } \beta_j \text{ distinto de 0 para } j = 2, 4, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	50.526	X1 X2 X3 X4 X5
Modelo reducido	55.133	X1 X3

Donde asumiendo verdadera la hipotésis nula planteada anteriormente, el modelo reducido esta dado por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \varepsilon_i;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 50$$

2.2. Estadístico de prueba y conclusión

Entonces es posible determinar el estadístico de prueba como:

3pt

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_3) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,44} \\ &= \frac{55.133 - 50.526}{1.14832} \\ &= 4.011 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,44} = 2.8165$, se puede ver que $F_0 > f_{0.95,3,44}$ 1pt

Por lo que en este caso no es posible descartar las variables del subconjunto.

La conclusión directa es que el subconjunto es significativo y por tanto ahí sí se pueden descartar

3. Pregunta 3

5pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Deseamos probar si $\beta_1 = \beta_2, \beta_3 = \beta_4$, mediante prueba de hipótesis lineal general, por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2;\ \beta_3 = \beta_4 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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 **L** dada por

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

2 pt

Asumiendo verdadera la hipótesis nula, el modelo reducido esta dado por:

$$Y_i = \beta_o + \beta_1 X_{i1}^* + \beta_3 X_{i3}^* + \beta_5 X_{i5} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 50$$

1 pt

Donde $X_{i1}^* = X_{i1} + X_{i2}$ y $X_{i3}^* = X_{i3} + X_{i4}$

3.2. Estadístico de prueba

17,5

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)}$$

$$F_0 = \frac{(SSE(MR) - 50.5262)/2}{1.14832} \overset{H_0}{\sim} f_{2,44} \tag{3}$$

2 pt

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

4 pt

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

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

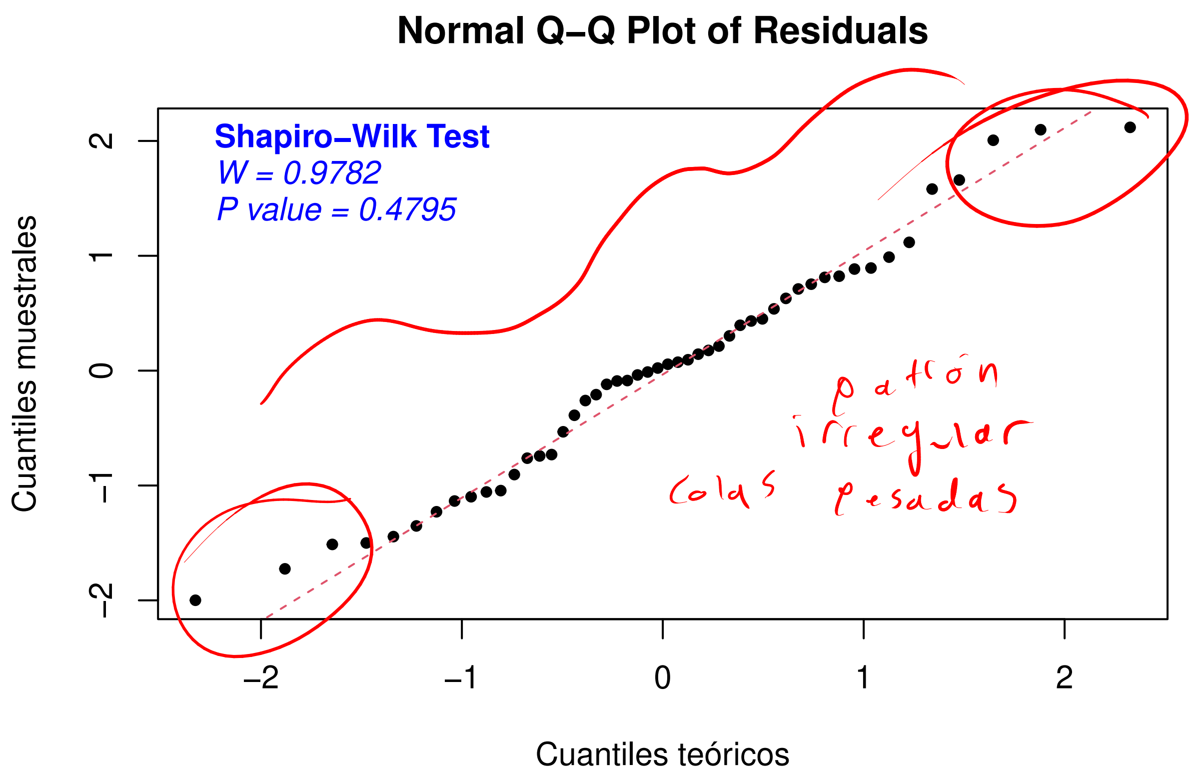

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Dado que el valor de P es aproximadamente 0.4795 y el nivel de significancia α es 0.05, el valor de P es mucho mayor que α, lo que significa que no se puede rechazar la hipótesis nula de que los datos se distribuyen normalmente con una media μ. Sin embargo, al observar la gráfica de comparación de cuantiles, se pueden notar colas más pesadas y patrones irregulares. Como el análisis gráfico tiene más poder, se decide rechazar la hipótesis de que los datos se distribuyen normalmente. El siguiente paso será validar si la varianza cumple con el supuesto de ser constante.

Excelente!

4.1.2. Varianza constante

3pt

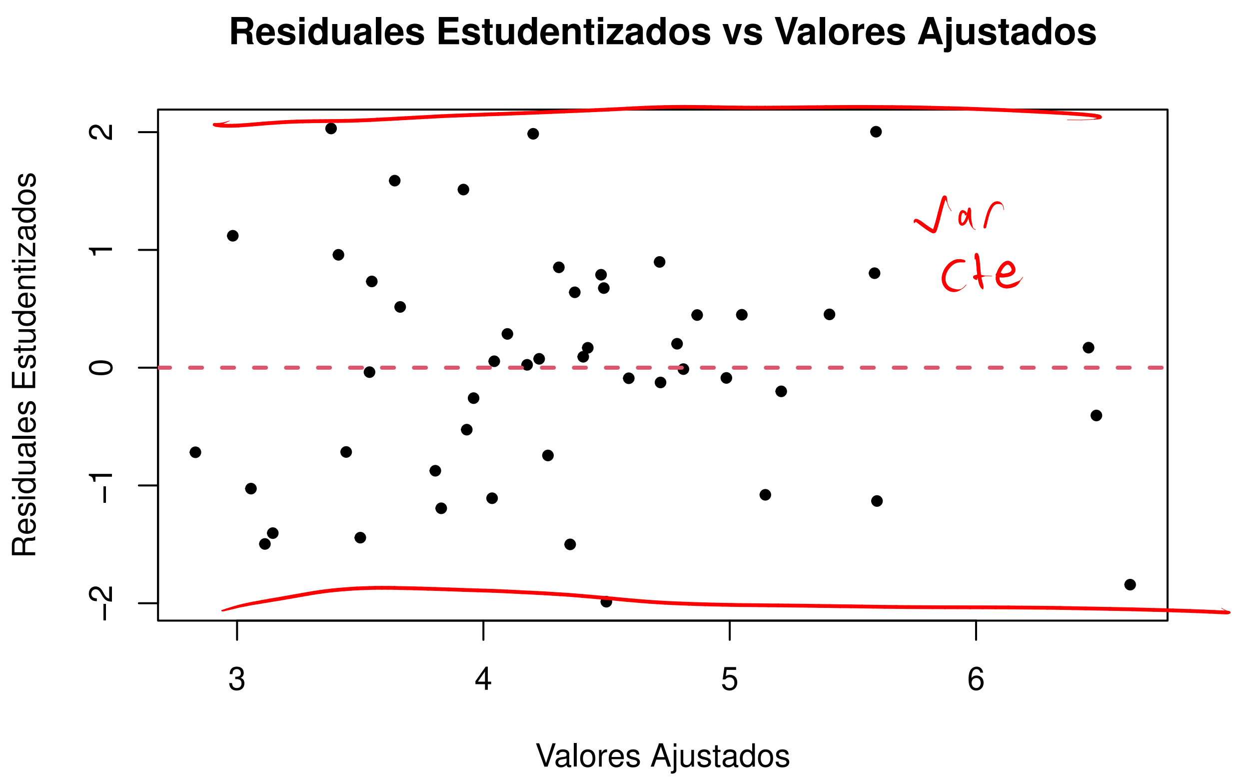

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Al observar el gráfico de residuos estudentizado en función de los valores ajustados, se puede notar que no hay patrones que sugieran un aumento o disminución en la varianza. Tampoco hay patrones que indiquen que la varianza no sea constante. Debido a la falta de evidencia en contra de la suposición de una varianza constante, se acepta como verdadera. Además, se observa que la media es igual a cero.

Es de hecho siempre sucede con los residuales estandarizados y estudentizados, para ver media 0 se hace sobre los residuales crudos

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

3 pt

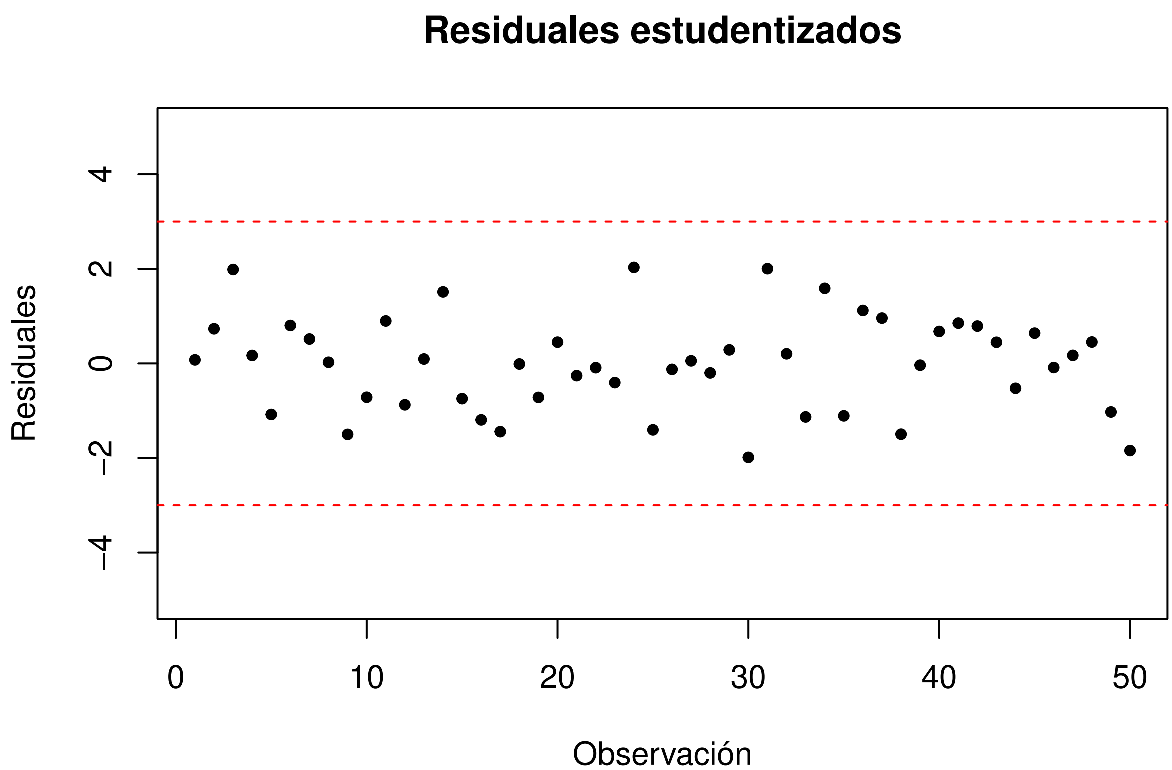

Figura 3: Identificación de datos atípicos

La gráfica previa muestra que no existen datos atípicos en el conjunto de datos, ya que ningún residual estudentizado supera el umbral de $|r_{estud}| > 3$.

4.2.2. Puntos de balanceo

2pt

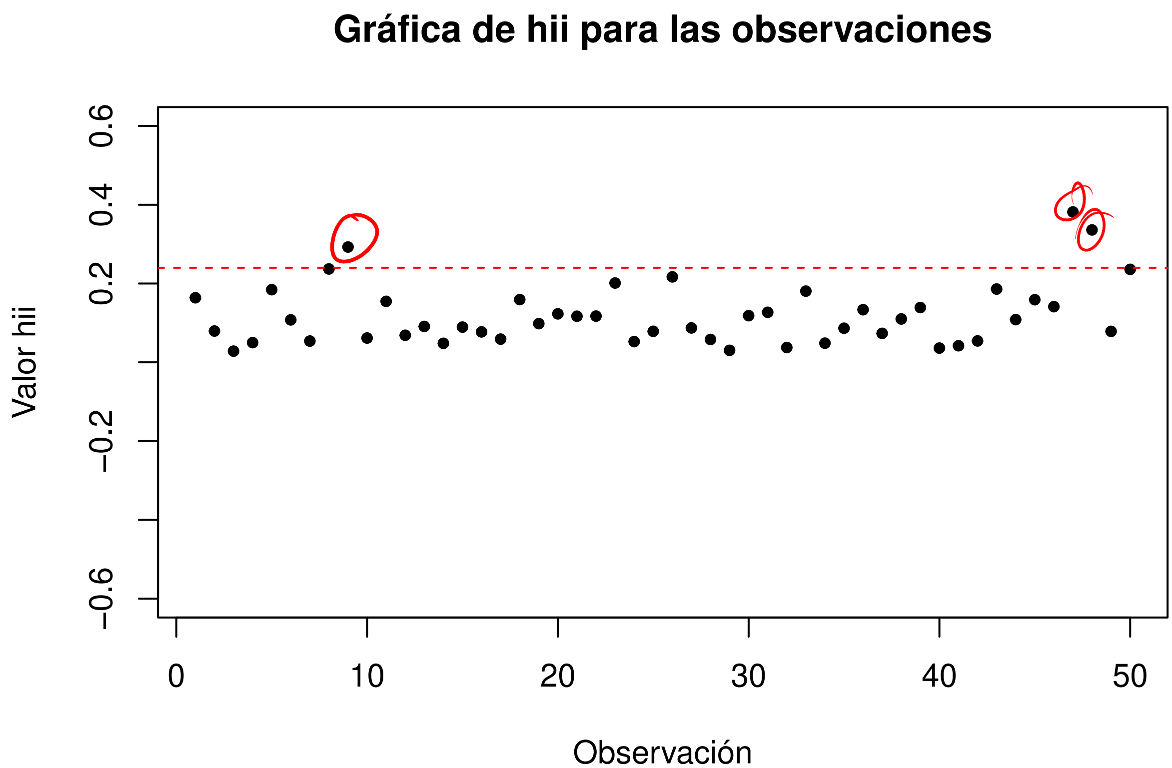

Figura 4: Identificación de puntos de balanceo

Cuadro 5: Resumen tabla de residuales

	ri	Di	hii values	DFFITS
9	-1.5004	0.1553	0.2927	-0.9795
47	0.1702	0.0030	0.3817	0.1323
48	0.4525	0.0173	0.3361	0.3190

Muy bien por hacer la tabla

¿cuánto da?

Al analizar la gráfica de observaciones en función de los valores h_{ii}, donde la línea punteada roja indica el valor $h_{ii} = 2\frac{p}{n}$, se puede notar que hay 3 datos en el conjunto que son considerados puntos de balanceo según el criterio de $h_{ii} > 2\frac{p}{n}$. Estos datos se detallan en la tabla.

¿Qué causan estos puntos?

4.2.3. Puntos influenciales

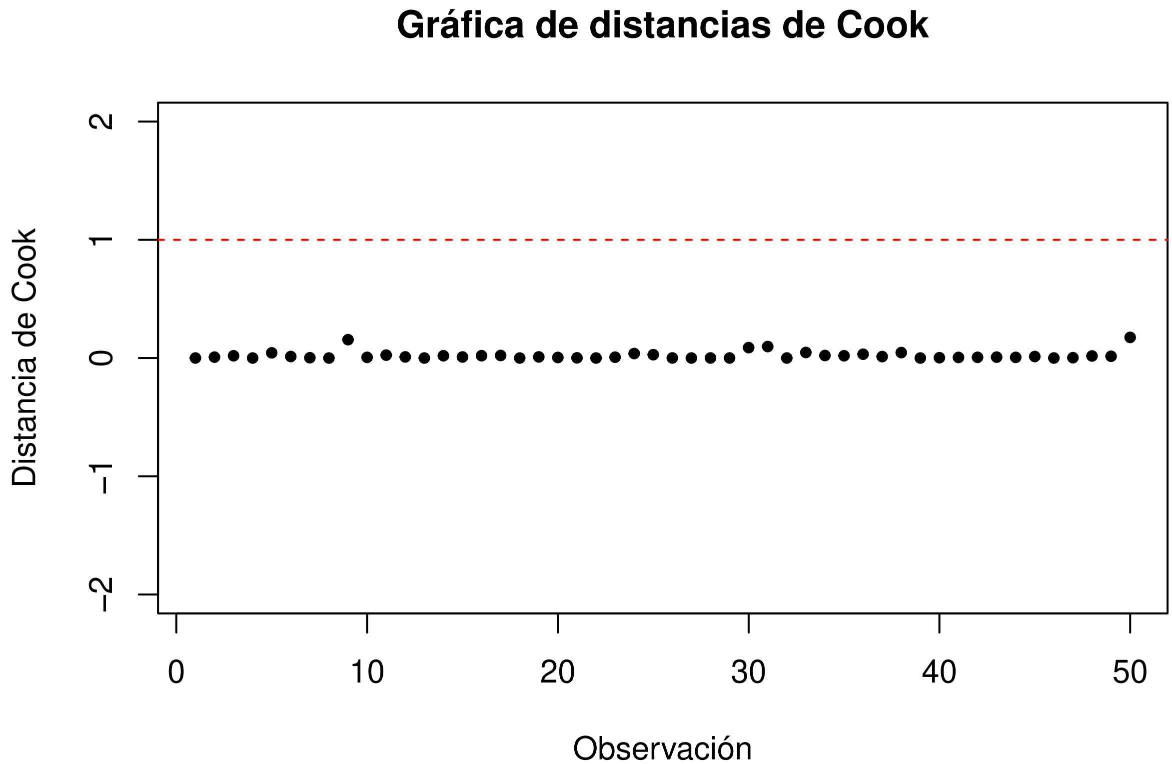

Figura 5: Criterio distancias de Cook para puntos influenciales

Mediante la gráfica de este criterio es posible ver que no existe ningún punto influencial que esté dado por Di > 1.

2 p+

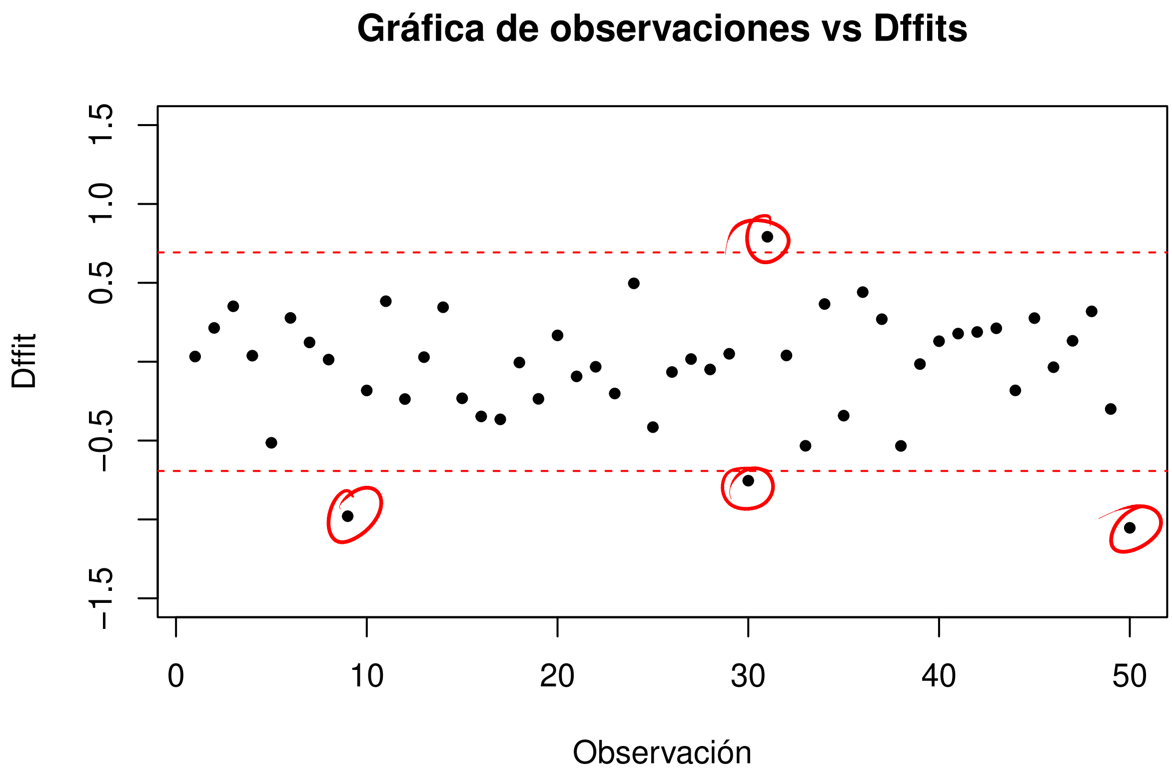

Figura 6: Criterio Dffits para puntos influenciales

10+

Cuadro 6: Resumen tabla de residuales

	ri	Di	hii values	DFFITS
9	-1.5004	0.1553	0.2927	-0.9795
30	-1.9869	0.0883	0.1184	-0.7544
31	2.0036	0.0972	0.1269	0.7921
50	-1.8422	0.1746	0.2359	-1.0532

Se puede observar que las observaciones 9, 30, 31 y 50 son consideradas puntos influenciales según el criterio de Dffits. Este criterio establece que cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, en este caso $2\sqrt{\frac{p}{n}} = 0.6928$ es considerado un punto influencial.

¿Qué causan estos puntos?

4.3. Conclusión

2,5 pt

En conclusión, el modelo de regresión lineal múltiple demostró ser significativo, y cumple con todos los supuestos necesarios, excepto el de normalidad de los errores. Esta falta de normalidad podría deberse a la presencia de los datos de balanceo e influenciales que logramos detectar. Sería recomendable identificar y analizar estos datos para determinar si pueden ser removidos o transformados para mejorar la normalidad de los errores y, por lo tanto, la validez del modelo.

No dicen claramente si el modelo es válido o no