4,3

Trabajo 1

Estudiantes

Johan Manuel Gutierrez Sabogal
María Paola Andrade Cortes
Juan José Ramírez Monsalve
Rafael Eduardo Rodriguez Muñoz

Equipo # 21

Docente

Julieth Veronica Guarin Escudero

Asignatura

Estadística II

Sede Medellín
30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

Introducción

Se presenta un caso práctico donde se involucra una muestra de 50 hospitales, las variables que se involucran en el modelo son el riesgo de infección(Y) como variable respuesta y las variables duración de la estadía(X_1), rutina de cultivos(X_2), número de camas(X_3), censo promedio diario(X_4) y número de enfermeras(X_5), correspondientes a las variables regresoras. Se busca plantear un modelo de regresión lineal múltiple, que permita identificar el comportamiento de la variable respuesta, respecto a cada una de las variables regresoras; además de realizar la respectiva comprobación de los supuestos del modelo de manera que se compruebe que este funciona correctamente para el caso presentado.

1. Pregunta 1

Teniendo en cuenta la base de datos Equipo21.txt, en la cual hay 5 variables variables regresoras denotadas como:

Y: Riesgo de infección (RI)

X_1: Duración de la estadía (DE)

X_2: Rutina de cultivos (RC)

X_3: Número de camas (NC)

X_4: Censo promedio diario (CPD)

X_5: Número de enfermeras (NE)

Entonces se plantea el módelo de regresión lineal múltiple:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{12i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i \ ; \ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2); \ 1 \leq i \leq 50$$

1.1. Modelo de regresión

Al ajustar el modelo se obtienen los siguientes valores para los parámetros β_0, β_1, β_2,β_3, β_4, β_5:

Cuadro 1: Tabla de valores ajustados

	Valor del parámetro
β_0	-1.4866
β_1	0.1413
β_2	0.0299
β_3	0.0471
β_4	0.0164
β_5	0.0025

Con los datos de los parámetros, el modelo de regresión ajustado es el siguiente:

$$\hat{Y}_i = -1.4866 + 0.1413X_{1i} + 0.0299X_{2i} + 0.0471X_{3i} + 0.0164X_{4i} + 0.0025X_{5i}$$

1.2. Significacia de la regresión

Para analizar la significacia de la regresión se plantea la siguiente prueba de hipótesis:

$$\begin{cases} H_0 : \text{Todos los } \beta_j = 0 \text{ para j=1,2,3,4,5} \\ H_a : \text{Algún } \beta_j \text{ distinto de 0, para j=1, 2,3,4,5} \end{cases}$$

El estadístico de prueba para definir la región de rechazo esta dado por:

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

A continuación se presenta la tabala ANOVA, correspondiente al modelo:

Cuadro 2: Tabla anova para el modelo de regresión

	Sumas de cuadrados	Grados de libertad	Cuadrado medio	F_0	P-valor
Regresión	51.6756	5	10.335117	11.3731	4.44576e-07
Error	39.9844	44	0.908737		

Partiendo de los datos depositados en la tabla ANOVA y usando el criterio de rechazo de valor P, se puede decir que este, es lo suficientemente pequeño para un nivel de significancia α por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con j perteneciente al intervalo discreto $1 \leqslant j \leqslant 5$, aceptando la hipótesis alternativa, en la cual se plantea que algún $\beta_j \neq 0$, por lo tanto se puede afirmar que la regresión es significativa.

Al revisar el segundo criterio de rechazo, se confirma el rechazo de la hipotesis nula, debido a que el estadístico de prueba supera el percentil de la región de rechazo para una significancia de 5 %; es decir que $|F_0| > f_{0.5,5,44}$.

1.3. Significancia de los parámetros

El siguiente cuadro busca denotar cuales de los parámetros β_j son significativos y cuales no; además se plantean los juegos de hipótesis correspondientes para cada parámetro:

$$\begin{cases} H_0 : \beta_j = 0 \text{ para j=1,2,3,4,5} \\ H_a : \beta_j \text{ es distinto de 0, para j=1, 2,3,4,5} \end{cases}$$

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-1.4866	1.6742	-0.8880	0.3794
β_1	0.1413	0.1184	1.1934	0.2391
β_2	0.0299	0.0305	0.9812	0.3319
β_3	0.0471	0.0164	2.8733	0.0062
β_4	0.0164	0.0091	1.8030	0.0782
β_5	0.0025	0.0009	2.8393	0.0068

Por medio del criterio de rechazo de valor P, se concluye que los parámetros β_3 y β_5, son significativos para un valor de α del 5 %, mientras que los demás β_j no superan el criterio de rechazo.

1.3.1. Interpretacion de los parámetros

Se procede a analizar la interpretación de los parámetros que resultaron ser significativos, a partir de la prueba de significancia que se realizó anteriormente.

$\hat{\beta}_3$**:** El significado de la variable es que por cada unidad que aumente el número de camas, el riesgo de contagio se incrementará en 0.0471 unidades; mientras las demás variables permanecen constantes.

$\hat{\beta}_5$**:** El significado de la variable es que por cada unidad que aumente el número de enfermeras, el riesgo de contagio se incrementará en 0.0025 unidades; mientras las demás variables permanecen constantes.

1.3.2. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5638$, lo cual indica que el 56.38 % de la variabilidad total del modelo es explicada por la regresión, mientras que el 43.62 % de la variabilidad total de modelo es explicada por el error.

Siendo más precisos se tiene un R^2 *ajustado* $= 0.5142$, lo que realmente no supone un gran cambio para el número de variables del modelo.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Los parámetros con el valor p más alto en el modelo fueron X_1, X_2, X_4, se desea comprobar la significancia simultanea, a través de la tabla de todas las regresiones posibles; por lo que se plantea la siguiente prueba de hipótesis.

$$\begin{cases} \text{H}_0 : \text{Todos los } \beta_j = 0, \text{ para j=1,2,4} \\ \text{H}_1 : \text{Alguno de los } \beta_j \text{ es distinto de 0, para } j = 1, 2, 4 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	39.984	X1 X2 X3 X4 X5
Modelo reducido	48.106	X1 X3

Entonces, de acuerdo a lo anterior, el modelo reducido es:

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 50$$

2.2. Estadístico de prueba y conclusión

Se establece el estadístico de prueba de la siguiente manera:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,44} \\ &= \frac{48.106 - 39.984}{((48.106)/44)} \\ &= 7.429 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,44} = 2.8165$,con el valor del estadístico de prueba F_0. Dado que el estadístico de prueba supera el percentil $f_{0.95,3,44}$, se puede afirmar que, se rechaza la hipótesis nula, lo que indica que las variables del subconjunto son significativas y las demás variables del modelo se pueden descartar, adoptando el modelo reducido.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se quiere demostrar si β_1 es igual a cuatro veces β_5 y simultaneamente, si β_4 es igual a dos veces β_2, por tal razón se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \text{H}_0 : \beta_1 = 4\beta_5; \beta_4 = 2\beta_2; \text{Todas las igualdades se cumplen} \\ \text{H}_1 : \text{Algunas de las igualdades no se cumplen} \end{cases}$$

reescribiendo la prueba de hipótesis matricialmente:

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

Donde la matriz $\mathbf{L}$ esta dada por:

$$L = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & -4 \\ 0 & 0 & -2 & 0 & 1 & 0 \end{bmatrix}$$

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_2 X^*_{2i} + \beta_3 X_{3i} + \beta_5 X^*_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 50$$

Donde $X^*_{2i} = X_{2i} + 2X_{4i}$ y $X^*_{5i} = 4X_{1i} + X_{5i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 para el modelo reducido esta dado por:

$$F_0 = \frac{((SSE(MR) - SSE(MF)/2)}{MSE(MF)} \overset{H_0}{\sim} f_{2,44}$$
$$F_0 = \frac{((SSE(MR) - 39.984)/2)}{(39.984/44)} \overset{H_0}{\sim} f_{2,44} \quad (3)$$

\section

sección{Pregunta 4}

3.3. Supuestos del modelo

3.3.1. Normalidad de los residuales

Se validará el supuesto por medio de una prueba de hipótesis de ~~shapiro wilk~~, acompañada de un gráfico cuantil - cuantil:

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

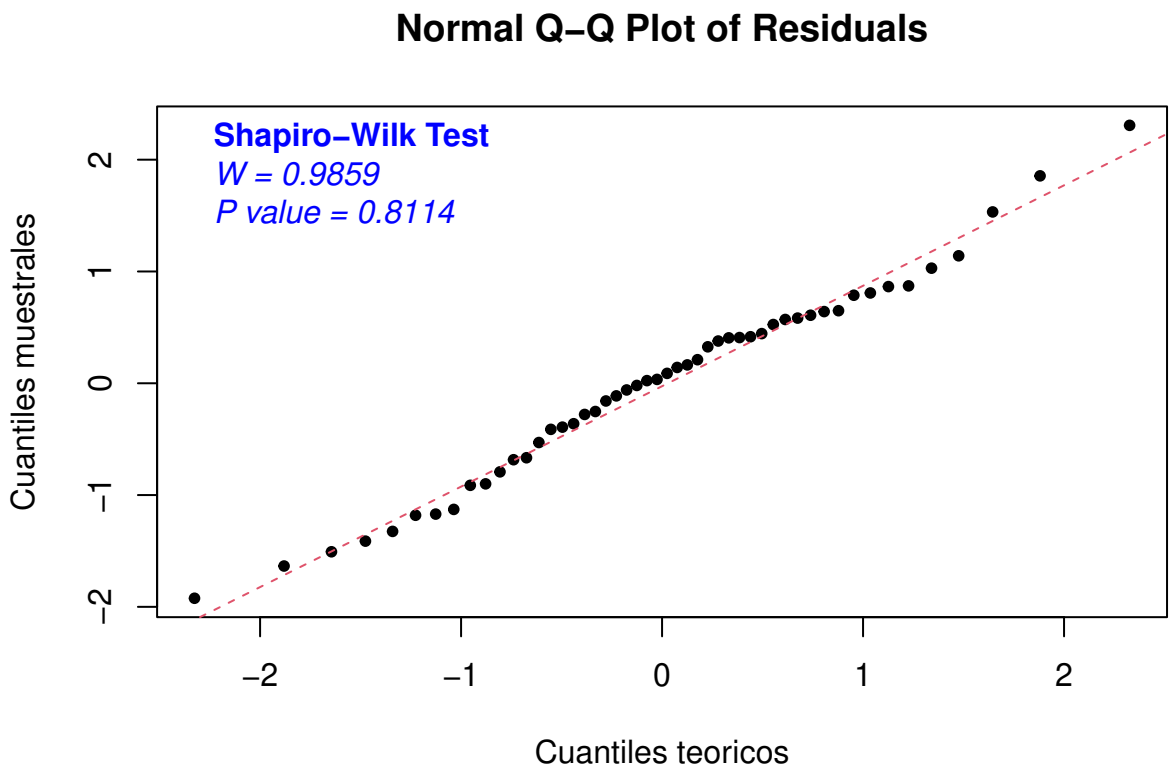


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Por el criterio del valor P que se obtuvo mediante la prueba de shapiro-wilk, en la cual dicho valor fue de aproximadamente 0,8114, para un nivel de significancia α de 5 %, no se rechaza la hipótesis nula con media μ y varianza σ^2, cabe recalcar que mediante la gráfica se logra evidenciar que los valores extremos distan bastant de la línea diagonal, lo que puede significar que realmente no todos los errores se distribuyan de manera normal. ✓

No hablaron del patrón irregular pero concluyen bien, no hay evidencia fuerte contra normalidad

3.3.2. Varianza constante

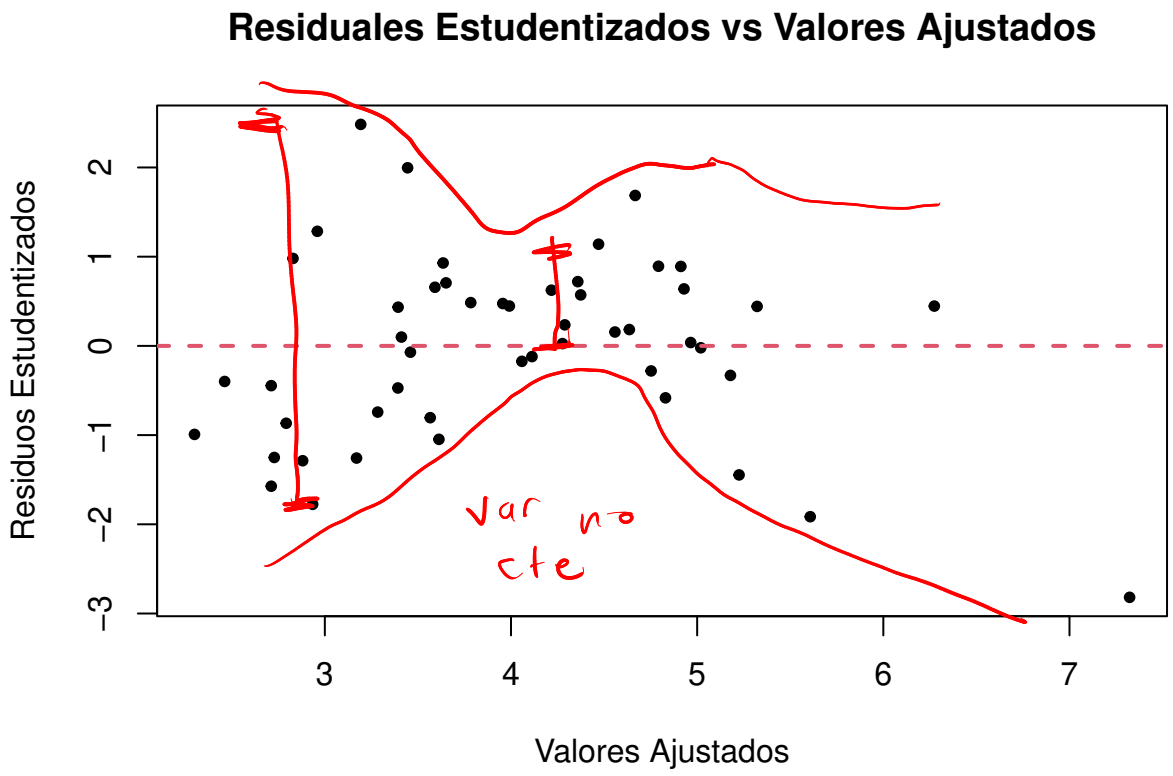


Figura 2: Gráficos residuales estudentizados vs valores ajustados

Se puede denotar una especie de cuello de botella, en el cual primero la disperción de los residuales es mayor, luego se hace pequeña y finalmente vuelve a expandirse, por tal razón, se puede concluir que la varianza de los errores no es constante y con ello el modelo de regresión lineal multiple no cumple dicho supuesto.

3.4. Verificación de las observaciones

3.4.1. Datos atípicos

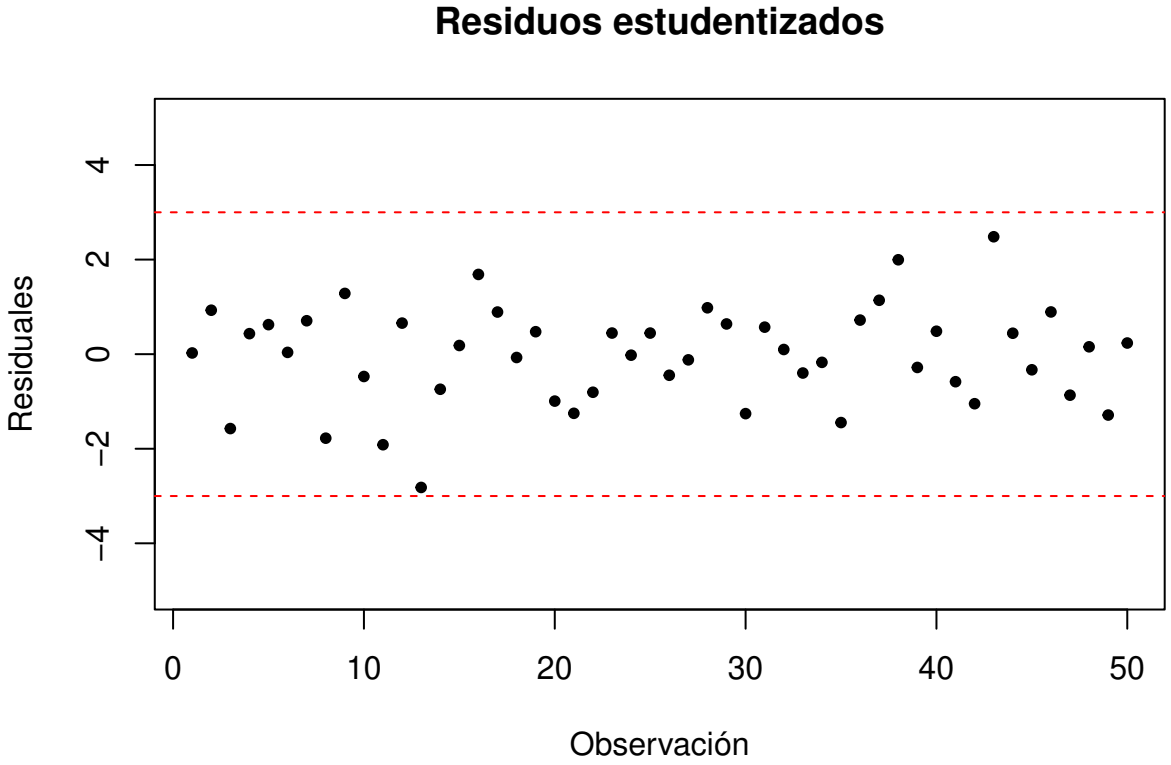
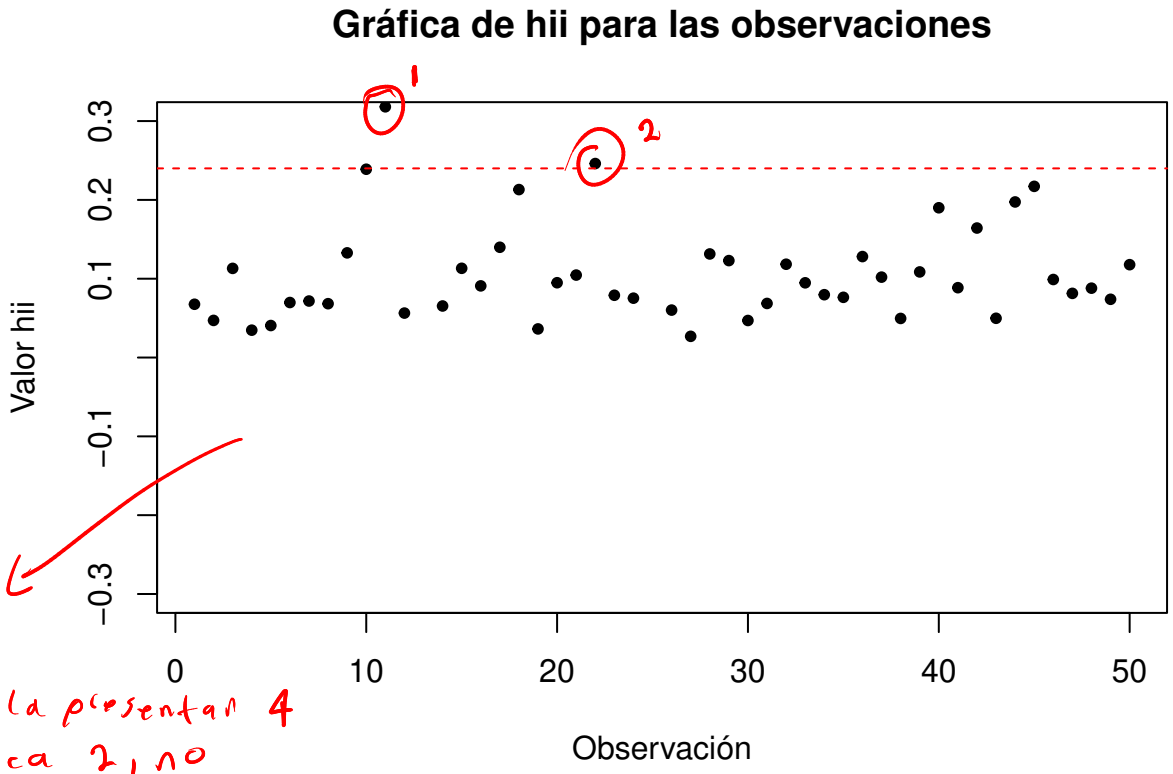


Figura 3: Identificación de datos atípicos

Para el caso estudiado, no se presentan datos atípicos que superen los límites estándar; es decir, que se cumple que $|r_{estud}| < 3$.

3.4.2. Puntos de balanceo

2,5 pt



En tabla presentan 4 y gráfica 2, no son congruentes.

Figura 4: Identificación de puntos de equilibrio

Cuadro 5: Resumen de los coeficientes de balanceo

	Res.studen	Cooks.d	Hii.value	Dffits
Dato 11	-1.9150	0.2851	0.3181	-1.3505
Dato 13	-2.8185	1.2604	0.4877	-3.0032
Dato 22	-0.8050	0.0353	0.2461	-0.4581
Dato 25	0.4463	0.0233	0.4125	0.3705

✓ Bien por hacer la tabla

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 0.24$, se puede apreciar que existen 4 datos del conjunto que son puntos de balance según el criterio bajo el cual $h_{ii} > 2\frac{p}{n}$,dichos valores se presentan en la tabla anterior y se presentan en las observaciones 11, 13, 22 y 25; estos valores podrían llegar a afectar estadísticas de resumen como los errores estándar de los coeficientes estimados y el R^2.

✓

¿No, sólo se 2

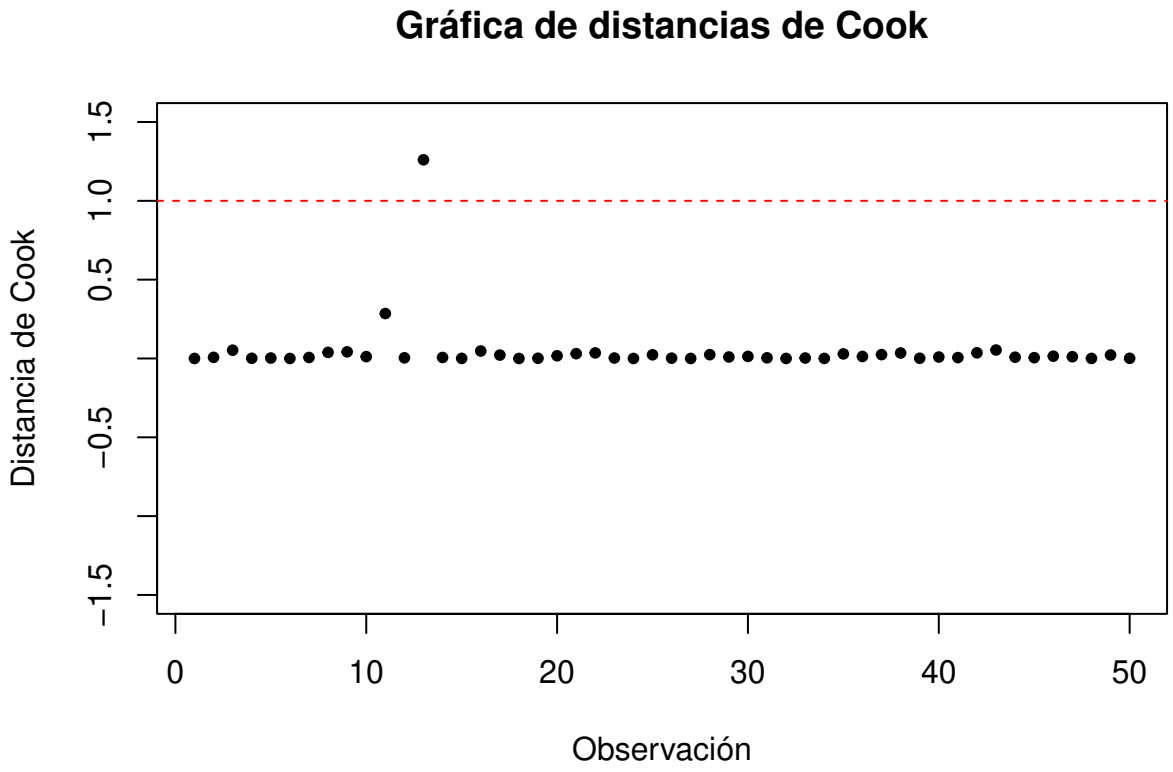


Figura 5: Criterio distancias de Cook para puntos influenciales

Cuadro 6: Resumen de los coeficientes de distancias de cook

	Res.studen	Cooks.d	Hii.value	Dffits
Dato 13	-2.8185	1.2604	0.4877	-3.0032

Como se denota en la anterior tabla, en el dato obtenido de la observación 13, resulta ser un valor influencial i, dado $D_i > 1$,dicho valor interfiere con las estimaciones que se puedan realizar a $\hat{Y}$, sus respectivos análisis y conslusiones.

a los parámetros, no a $\hat{Y}$

1,5 pt

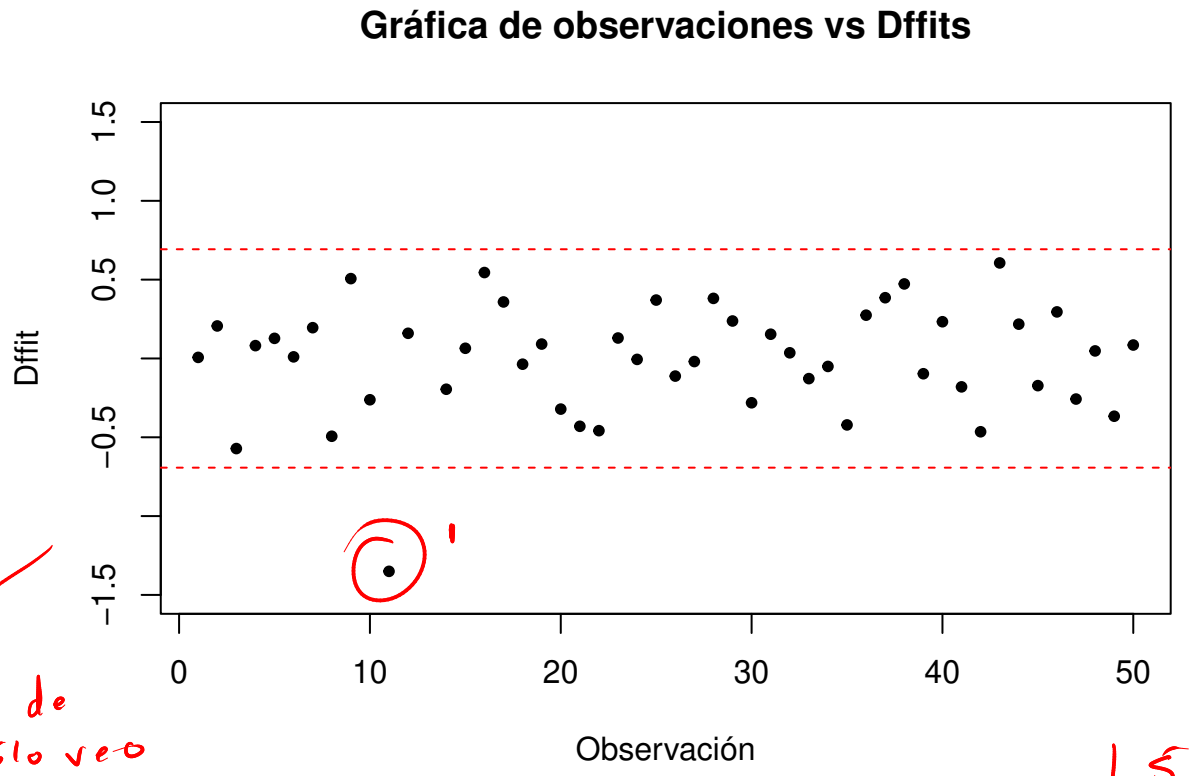


Figura 6: Criterio Dffits para puntos influenciales

Cuadro 7: Resumen de los coeficientes Observaciones vs Dffit

	Res.studen	Cooks.d	Hii.value	Dffits
Dato 11	-1.9150	0.2851	0.3181	-1.3505
Dato 13	-2.8185	1.2604	0.4877	-3.0032

Como se puede ver en la tabla anterior,los datos recopilados en las observaciones 11 y 13 son 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 influyente. Cabe destacar también que con el criterio de distancias de Cook, en el cual para cualquier punto cuyo $D_i > 1$ se considera un punto influencial, dado esto, la observación número 13 también se confirma como un punto influyente; dichos puntos pueden afectar las posibles estimaciones que se realicen a $\hat{Y}$.

3.5. Conclusión

Según la comprobación de supuestos, se puede decir que el modelo de regresión lineal planteado para este caso, no sirve para realizar estimaciones o predicciones sobre $\hat{Y}$, ya que no se tiene una certeza completa sobre la distribución normal de los errores, la variaza de los mismos, demuestra tener un comportamiento irregular en lugar de constante y hay presencia

de algunos datos atípicos, así como también datos de equilibrio y de influencia que interfieren con la precision de las diferentes inferencias que se desean realizar con este tipo de modelos.