Trabajo 1

4,7

Estudiantes

Felipe Cabeza Pareja
Samuel Lopera Jaramillo
Jose David Gallego Zapata
Santiago Sosa Garcia

Equipo 41

Docente

Francisco Javier Rodriguez Cortes

Asignatura

Estadística II

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medellín
30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

Teniendo en cuenta la base de datos Equipo41, en la cual hay 5 variables regresoras, denominadas por:

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

Por lo tanto, se plantea el modelo de regresión lineal múltiple:

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

1.1. Modelo de regresión

Al ajustar el modelo, se estiman los siguientes coeficientes:

Cuadro 1: Valor de los parámetros estimados

	Valor de parámetro
β_0	1.9206
β_1	0.2494
β_2	-0.0276
β_3	0.0767
β_4	-0.0035
β_5	0.0024

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 1.9206 + 0.2494 X_{1i} - 0.0276 X_{2i} + 0.0767 X_{3i} + -0.0035 X_{4i} + 0.0024 X_{5i}$$

1.2. Significancia de la regresión

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

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

Donde el estadístico de prueba corresponde a:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,44} \quad (1)$$

Teniendo la información anterior, se presenta la tabla Anova correspondiente a este modelo:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	Grados de libertad	Cuadrado medio	F_0	Valor-p
Regresión	67.4392	5	13.487831	13.8587	3.86349e-08
Error	42.8226	44	0.973242		

Como Valor-p $< 0.05 = \alpha$, se rechaza H_0 concluyendo que el modelo de RLM propuesto es significativo. Esto quiere decir que, el riesgo de infección es afectado significativamente por al menos una de las predictorias consideradas.

1.3. Significancia de los parámetros

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

$$\begin{cases} H_0 : \beta_j = 0 \text{ para j = 1,2,...,5} \\ H_a : \beta_j \neq 0 \end{cases}$$

Donde el estadístico de prueba corresponde a:

$$T_{j,0} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \overset{H_0}{\sim} t_{44} \quad (2)$$

Cuadro 3: Tabla de parámetros estimados

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	valor-P
β_0	1.9206	1.9555	0.9822	0.3314
β_1	0.2494	0.0945	2.6395	0.0114
β_2	-0.0276	0.0357	-0.7739	0.4431
β_3	0.0767	0.0187	4.0907	0.0002
β_4	-0.0035	0.0093	-0.3735	0.7106
β_5	0.0024	0.0009	2.7874	0.0078

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 cada uno en presencia de los demás paramentros.

Por otro lado, se encuentra que β_0, β_2, y β_4 son individualmente no significativos en presencia de los demás paramentros.

1.4. Interpretación de los parámetros

$\hat{\beta}_1 = 0.2494$ **indica que por cada unidad de aumento en la duración de la estadia, el promedio del riesgo de infección aumenta en** 0.2494 **unidades, cuando las demás variables se mantienen fijas**

$\hat{\beta}_3 = 0.0767$ **indica que por cada unidad de aumento en el número de camas, el promedio del riesgo de infección aumenta en** 0.0767 **unidades, cuando las demás variables se mantienen fijas**

$\hat{\beta}_5 = 0.0024$ **indica que por cada unidad de aumento en el número de enfermeras, el promedio del riesgo de infección aumenta en** 0.0024 **unidades, cuando las demás variables se mantienen fijas**

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.6116$, lo que significa que aproximadamente el $61.16\,\%$ de la variabilidad total en el riesgo de infección es explicada por el modelo de regresión propuesto en el presente informe.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más alto en el modelo fueron X_1, X_2, X_4, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_4 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 2, 4 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	42.823	X1 X2 X3 X4 X5
Modelo reducido	50.473	X3 X5

Luego un modelo reducido para la prueba de significancia del subconjunto 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 construye el estadístico de prueba como:

$$F_0 = \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \ \cdots, \ \beta_5))/3}{MSE(\beta_0, \ \cdots, \ \beta_5)} \overset{H_0}{\sim} f_{3,44}$$

$$F_0 = \frac{(50.473 - 42.823)/3}{0.97325} \overset{H_0}{\sim} f_{3,44} \tag{3}$$

$$F_0 = 2.620087$$

comparando el F_0 obtenido con el cuantil$f_{0.95,3,44} = 2.8165$, se puede ver que $F_0 < f_{0.95,3,44}$ esta esta en la región de aceptación, por ende, no se tiene suficiente evidencia estadística para rechazar la hipótesis nula, por lo que se concluye que el subconjunto no es significativo, siendo posible descartar del modelo las variables del subconjunto.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

$$\begin{cases} \mathrm{H}_0 : \beta_1 = 4\beta_2; \ \beta_3 = 2\beta_4 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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:

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_2 X^*_{2i} + \beta_4 X^*_{4i} + \beta_5 X_{5i} + \varepsilon_i, \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); \ 1 \leqslant i \leqslant 50$$

Donde $X^*_{2i} = 4X_{1i} + X_{2i}$ y $X^*_{4i} = 2X_{3i} + X_{4i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - 42.823)/2}{0.97325} \overset{H_0}{\sim} f_{2,44} \tag{4}$$

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

En esta parte parte se revisaran dos criterios, uno es la siguiente prueba de hipotesis de ~~shapiro wilk~~ y la otra es la prueba grafica cuantil-cuantil:

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

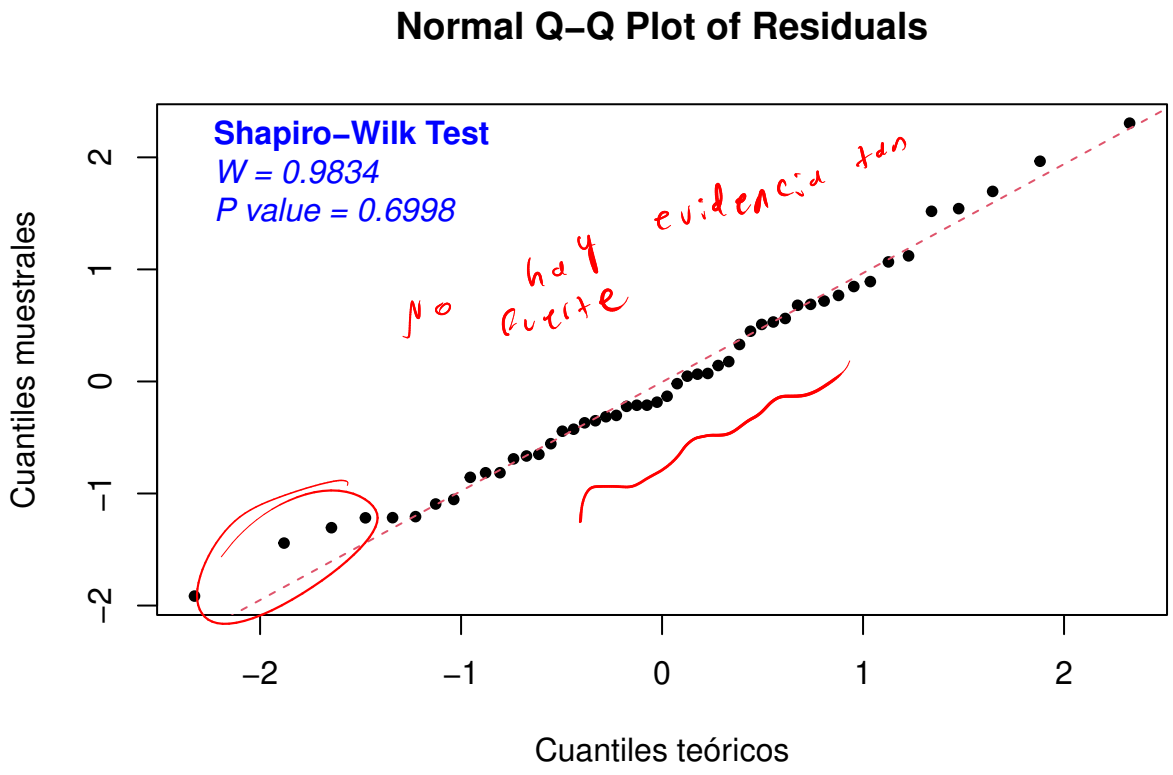

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

lo primero será analizar la prueba de hipótesis de shapiro-wilk, esta nos arroja un valor-p de 0.6998, esta cifra es bastante más grande que el nivel de significancia de $\alpha = 0.05$, por lo tanto la prueba nos indicaría que no se rechaza la hipótesis nula, es decir que los datos se

distribuyen normal con media μ y varianza σ^2, pero al revisar la gráfica podemos notar que hay varios datos sobretodo en las colas que se alejan del comportamiento esperado de una distribución normal y al ser la prueba grafica más determínante que la analítica se termina por rechazar el supuesto de normalidad, a continuación pasamos a verificar el supuesto de varianza constante.

4.1.2. Varianza constante

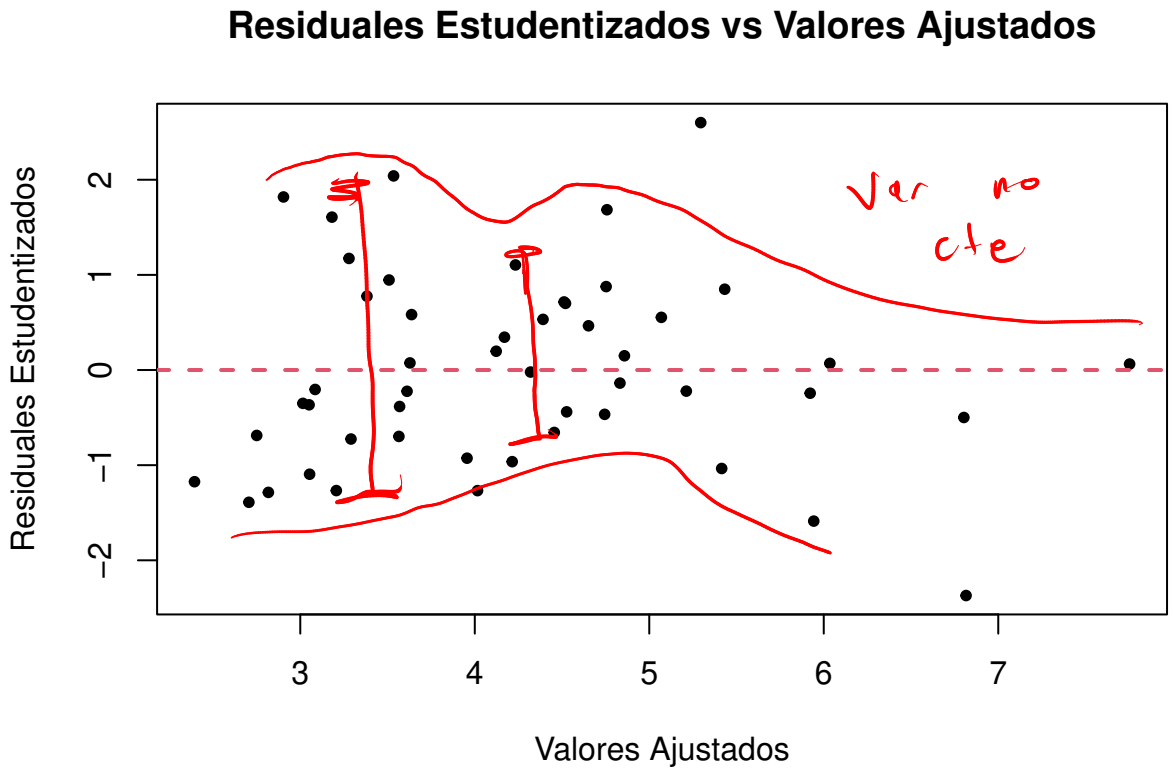

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Al observar la gráfica de residuales estudentizados vs valores ajustados se puede ver que los puntos parecen distribuirse de forma homogénea en el espacio de la gráfica, no se observan patrones tan marcados de heterocedasticidad, es decir, cambios abruptos en la distribución de la varianza en el plano de la gráfica lo que nos sugiere que los residuales cumplen con el supuesto de varianza constante.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

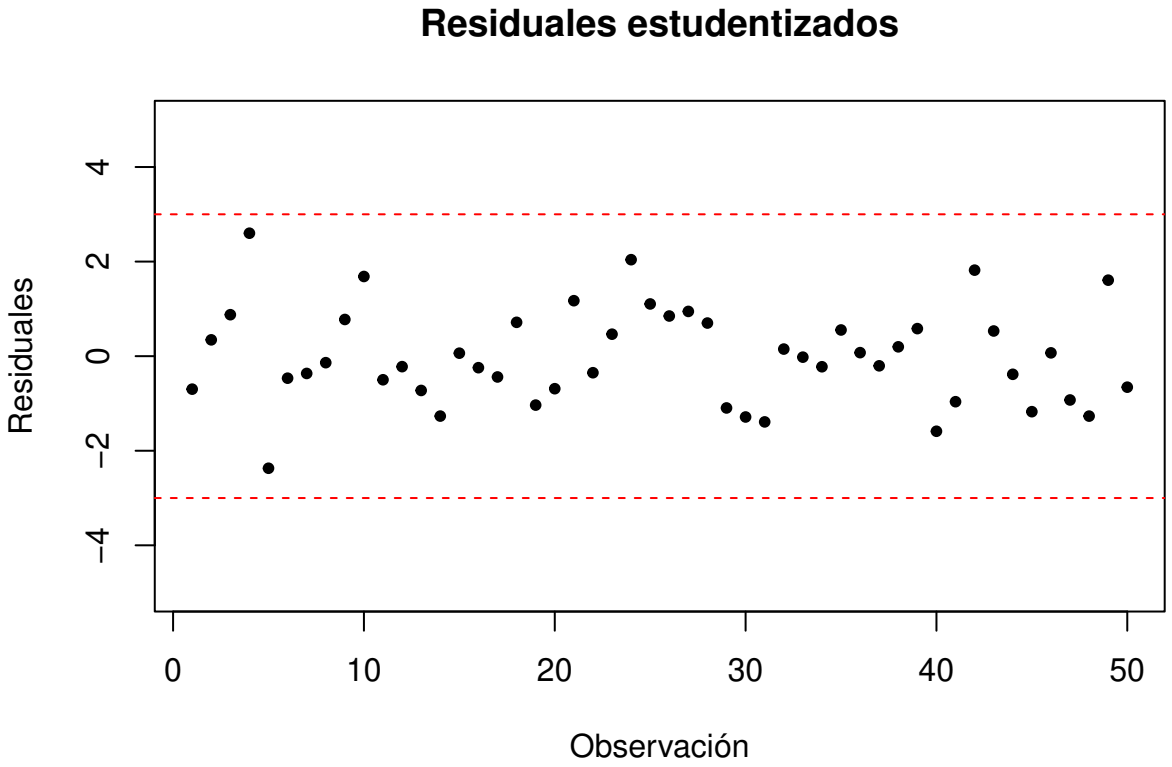

Figura 3: Identificación de datos atípicos

Al analizar la gráfica anterior podemos ver que ninguno de los residuales estudentizados cumple con el criterio de $|r_{estud}| > 3$, por lo tanto podemos decir que no se registran datos atípicos en la muestra estudiada.

4.2.2. Puntos de balanceo

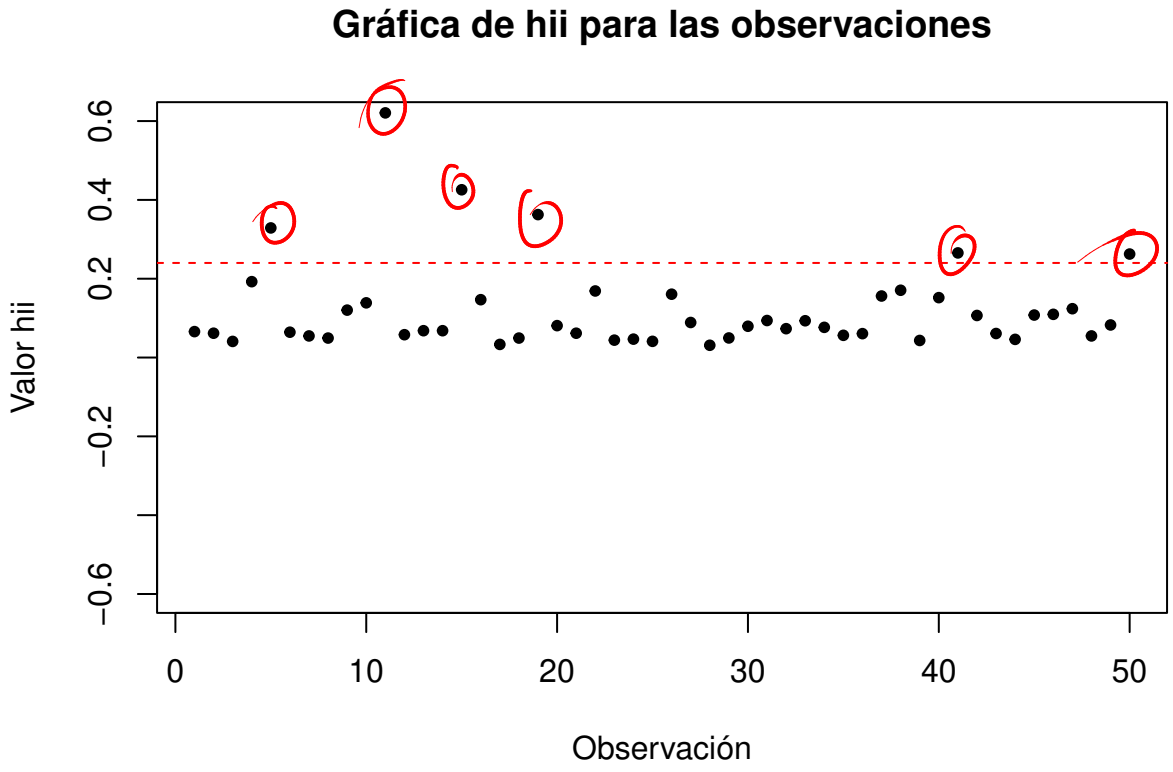

Figura 4: Identificación de puntos de balanceo

Cuadro 5: Tabla de puntos de balanceo

	res.stud	Cooks.D	hii.value	Dffits
5	-2.3701	0.4588	0.3289	-1.7562
11	-0.4988	0.0679	0.6208	-0.6327
15	0.0630	0.0005	0.4257	0.0536
19	-1.0348	0.1016	0.3627	-0.7813
41	-0.9633	0.0559	0.2655	-0.5787
50	-0.6559	0.0255	0.2626	-0.3888

Al analizar la gráfica de observaciones vs valores h_{ii} se puede apreciar que hay seis observaciones que están cumpliendo con el criterio definido en los puntos de balanceo, el cual es $h_{ii} > 2\frac{p}{n}$, donde $h_{ii} = 2\frac{p}{n} = 0.24$, y al revisar la tabla vemos que dichas observaciones son: la quinta observación con un valor $h_{ii} = 0.3289$, la onceava observación con un valor de $h_{ii} = 0.6208$, la quinceava observación con un valor de $h_{ii} = 0.4257$, la decimonovena observación con un valor de $h_{ii} = 0.3627$, la observación número 41 con un valor de $h_{ii} = 0.2655$ y por último la observación número 50 con un valor de $h_{ii} = 0.2626$, es muy importante identificar estos

puntos porque pueden afectar las estadísticas resumen del modelo como el R^2 y los errores estándar de los coeficientes estimados, lo cual se puede traducir en una mala interpretación del modelo y de la significancia de los parámetros estimados.

Perfecto :3

4.2.3. Puntos influenciales

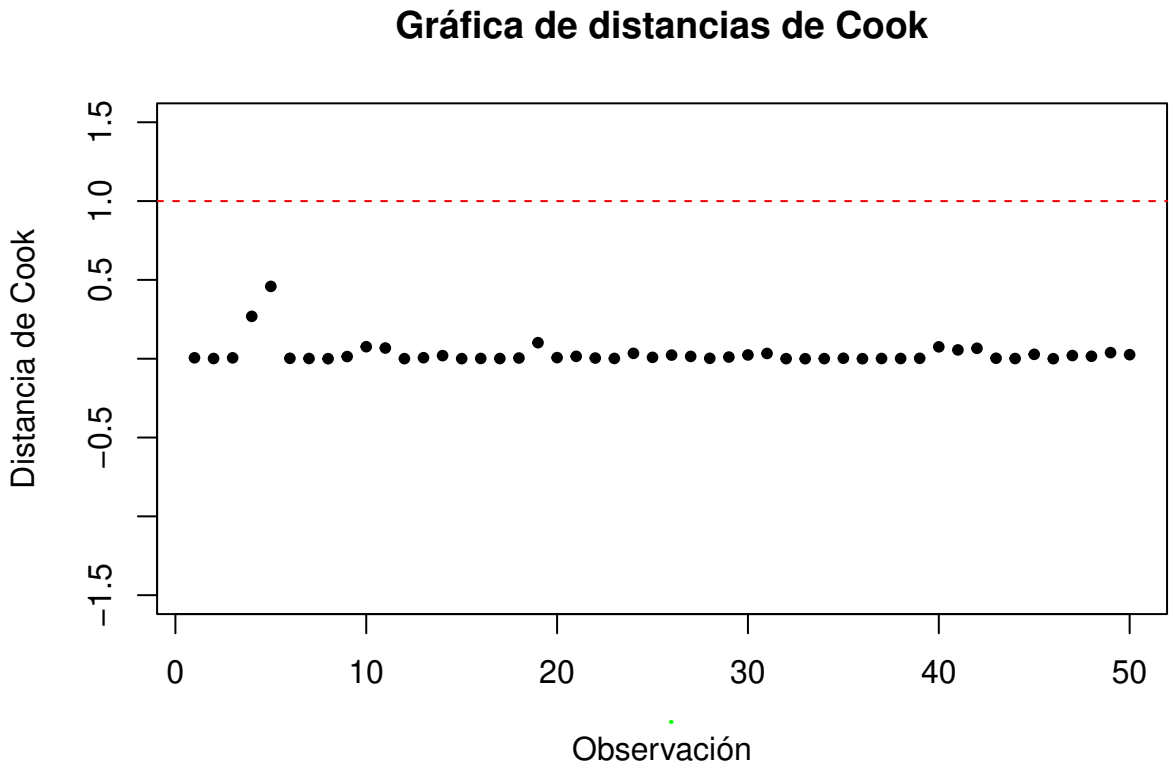

Figura 5: Criterio distancias de Cook para puntos influenciales

Al hacer la evaluación de puntos influénciales con la distancia de cook se obtiene que ninguna de las observaciones cumple con el criterio de $D_i > 1$, es decir que en cada una de las observaciones no se obtuvo una gran diferencia de los estimadores por mínimos cuadrados sin incluir esa i-ésima observación, o visto de otra forma la influencia del punto sobre el vector de parámetros no es suficiente para considerarlo un punto influencial.

Excelente!

2pt

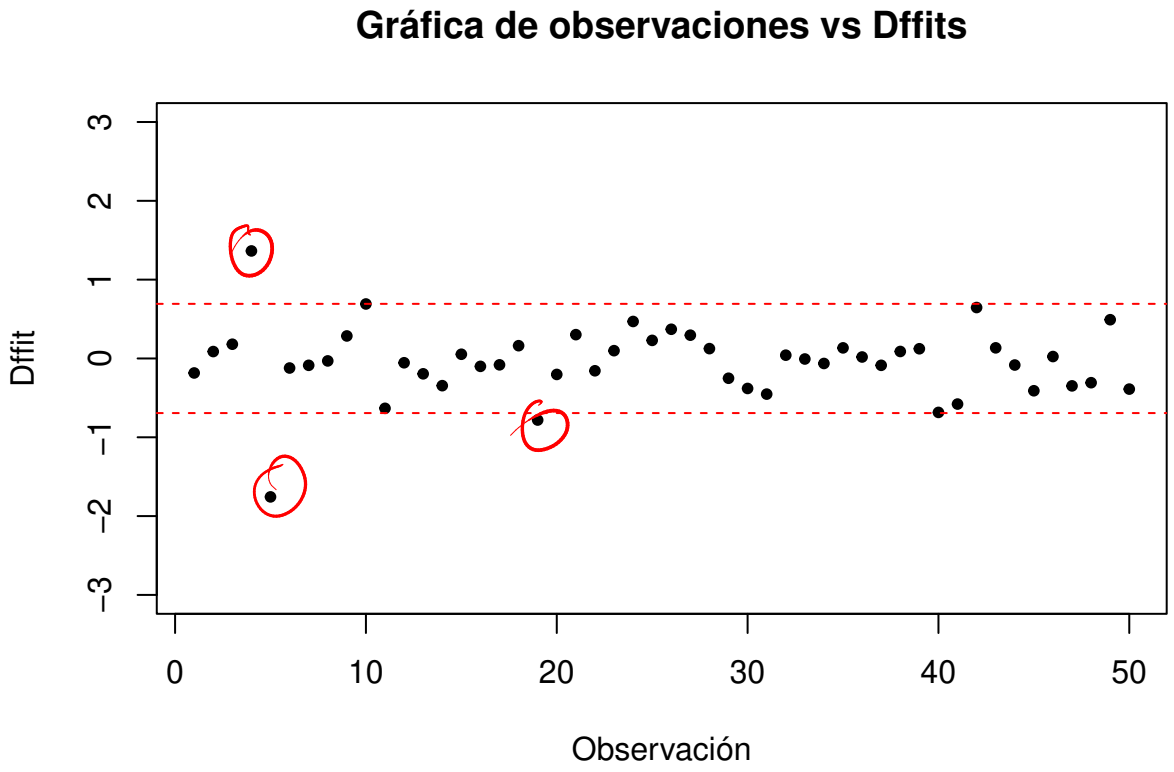

Figura 6: Criterio Dffits para puntos influenciales

Cuadro 6: Tabla de puntos influenciales

	res.stud	Cooks.D	hii.value	Dffits
4	2.6002	0.2686	0.1925	1.3641
5	-2.3701	0.4588	0.3289	-1.7562
19	-1.0348	0.1016	0.3627	-0.7813

Vemos que las observaciones 4, 5 y 19 están cumpliendo con el criterio de la prueba Dffits $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, donde $2\sqrt{\frac{p}{n}} = 0.6928$, lo que significa que dichas observaciones son influénciales, esta vez se usa el vector con los valores ajustados para ver si hay una diferencia significativa al no incluir la observación, para estos caso se obtiene gracias a la evaluación Dffits que si hay diferencias importantes en el modelo al no incluir los puntos enunciados por la prueba.

Estos puntos influénciales tienen un impacto muy importante en el modelo de regresión ya que lo halan en su dirección, haciendo que no sea el que mejor se ajuste a la mayoría de datos proporcionados.

4.3. Conclusión

Acerca de la validez de la regresión se podría concluir que el modelo no está ajustando los datos de la manera más óptima, por lo que se ve en la prueba grafica no se está cumpliendo

con el supuesto de normalidad y esto de entrada hace que el ajuste pierda un importante fundamento teórico para su verificación, esto probablemente se deba a los puntos influénciales que observamos anteriormente, estas observaciones desajustan el modelo ya que lo halan en direcciones en las cuales no se logra acomodar a la mayoría de los datos y por lo tanto da estimaciones de la variable respuesta alejadas de los valores reales o esperados.