Trabajo 1

3,9

Estudiantes

Brayan Alberto Patiño Alzate
Jenifer Tatiana Atehortua Duque
Samuel David Suarez Tovar
Felipe Andrés Villero Mejia
Equipo # 65

Docente

Javier Armando Lozano Rodriguez

Asignatura

Estadística II

Sede Medellín
5 de Octubre de 2023

Índice

Índice de figuras

Índice de tablas

1. Pregunta 1

20pt

Teniendo en cuenta la base de datos asignada, se realiza el ajuste del modelo de regresión lineal múltiple (RLM), explicando la eficacia sobre el control de infecciones hospitalarias en la cual hay 5 variables regresoras dadas por:

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 54$$

Se asignan las siguientes variables:

$$\begin{cases} Y : \text{Riesgo de Infección} \\ X1 : \text{Duración de estadia} \\ X2 : \text{Rutina de cultivos} \\ X3 : \text{Número de camas} \\ X4 : \text{Censo promedio diario} \\ X5 : \text{Número de enfermeras} \end{cases}$$

1.1. Modelo de regresión

Al ajustar el modelo, se obtienen los siguientes coeficientes:

Tabla 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-0.7722
β_1	0.1952
β_2	0.0255
β_3	0.0552
β_4	0.0076
β_5	0.0016

3pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.7722 + 0.1952 X_{1i} 0.0255 X_{2i} + 0.0552 X_{3i} + 0.0076 X_{4i} + 0.0016 X_{5i}$$

1.2. Significancia de la regresión

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

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

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,48} \tag{1}$$

$$F_0 = \frac{10.82931}{1.01425} = 10.6772 \tag{2}$$

Ahora, se presenta la tabla Anova:

Tabla 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	54.1466	5	10.82931	10.6772	6.18567e-07
Error	48.6838	48	1.01425		

De la tabla Anova, se observa un valor P aproximadamente igual a 0, por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$, en este caso el P-valor nos permite concluir que sí, el modelo de regresión es significativo, rechazando la hipótesis nula planteada anteriormente, esto quiere decir que alguna de las variables predicctorias es significativa en el riesgo de adquirir infección en hospitales.

1.3. Significancia de los parámetros

En el siguiente cuadro respecto a los parametros individuales se presenta información de los los mismos, lo cual permitirá determinar cuáles de ellos son significativos.

Tabla 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-0.7722	1.9008	-0.4063	0.6864
β_1	0.1952	0.0905	2.1575	0.0360
β_2	0.0255	0.0351	0.7247	0.4721
β_3	0.0552	0.0145	3.8224	0.0004
β_4	0.0076	0.0091	0.8416	0.4042
β_5	0.0016	0.0007	2.1533	0.0364

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_5 son significativos, Los P valores mostrados en la tabla me indican cuales de los parámetros está aportando significativamente al modelo de regresión, ya que sus P-valores 0.0360 y 0.0004 y 0.0364 respectivamente son menores al valor alfa.

1.4. Interpretación de los parámetros

Con lo anterior se puede determinar que $\hat{\beta}_1$: 0.0360, indica que por cada unidad que aumente el promedio de duración de la estadia en el hospital X_1, el riesgo de infección aumenta 0.0360 unidades, cuando las demás variables predictoras se mantienen fijas.

Asi mismo $\hat{\beta}_3$: 0.0004 indica que por cada unidad que aumente el promedio de camas en el hospital X_3, el riesgo de infección aumenta 0.0004 unidades, cuando las demás variables predictoras se mantienen fijas.

Y por ultimo $\hat{\beta}_5$: 0.0364 indica que por cada unidad que aumente el promedio de enfermeras en el hospital X_5, el riesgo de infección aumenta 0.0364 unidades, cuando las demás variables predictoras se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2

$$R^2 = \frac{SSR}{SSR + SSE} \tag{3}$$

$$R^2 = \frac{54.1466}{54.1466 + 48.6838} = 0.526562 \tag{4}$$

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.526562$, lo que significa que aproximadamente el 52.66 % de la variabilidad total observada en los resultados de la prueba de riesgo 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, rutina de cultivos X_2 (Valor P= 0.4721) y censo promedio diario X_4 (Valor P= 0.4042), ahora se procede a plantear la prueba de hipótesis para verificar la significancia en simultaneo del subconjunto con valores p más altos del modelo de regresión.

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

Tabla 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	48.684	X1 X2 X3 X4 X5
Modelo reducido	83.101	X2 X4

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \cdots, \beta_5))/2}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{2,48} \\ &= \frac{(83.101 - 48.684/2)}{1.01425} \\ &= 17.4537 \end{aligned} \tag{5}$$

Ahora, comparando el F_0 con $F_{0.95,2,48} = 17.4537$, se puede ver que $F_0 < f_{0.95,2,48}$

Por consiguiente, con un nivel de significancia de 0.05 comparando F_0 con $F_{0.95,2,48} = 17.4537$, se puede ver que $F_0 < F_{0.95,2,48}$, como F_0 no esta dentro de la region de rechazo, no se rechaza H_0 y se concluye que las variables X_2 y X_4 no afectan de forma conjunta a la probabilidad de riesgo de infección.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se plantea la hipotesis de que la variable (X1) Duración de la estadia es igual de significativa a (X3) número de camas; ademas (X2) rutina de cultivos y (X4) censo promedio diario son igual de significativas Por consiguiente se plantea la siguiente prueba de hipótesis:

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

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 $\mathbf{L}$ dada por

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_2 X^*_{2i} + \beta_3 X^*_{3i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 54$$

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

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,48} \tag{6}$$

$$F_0 = \frac{(SSE(MR) - 48.684/2}{1.01425} \tag{7}$$

4. Pregunta 4

4.1. Supuestos del modelo

Normalidad Tomando en cuenta el valor p no se rechaza H_0 siendo se podría decir que hay normalidad, sin embargo la gráfica nos muestra patrones irregulares, además las colas están alejadas de la línea de tendencia marcada en la gráfica lo que supone que el supuesto de normalidad no se este cumpliendo como lo indica el valor p, por consiguiente al tener mas peso el método grafico frente al valor se rechaza el supuesto de normalidad por medio del grafico de cuanti – cuantil.

Varianza En la gráfica observamos que la varianza no tiene una tendencia marcada hacia la linealidad, o hacia valores constantes, ya que los datos se muestran dispersos a lo largo de la gráfica, se puede decir que no se cumple el supuesto de varianza constante.

4.1.1. Normalidad de los residuales

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

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

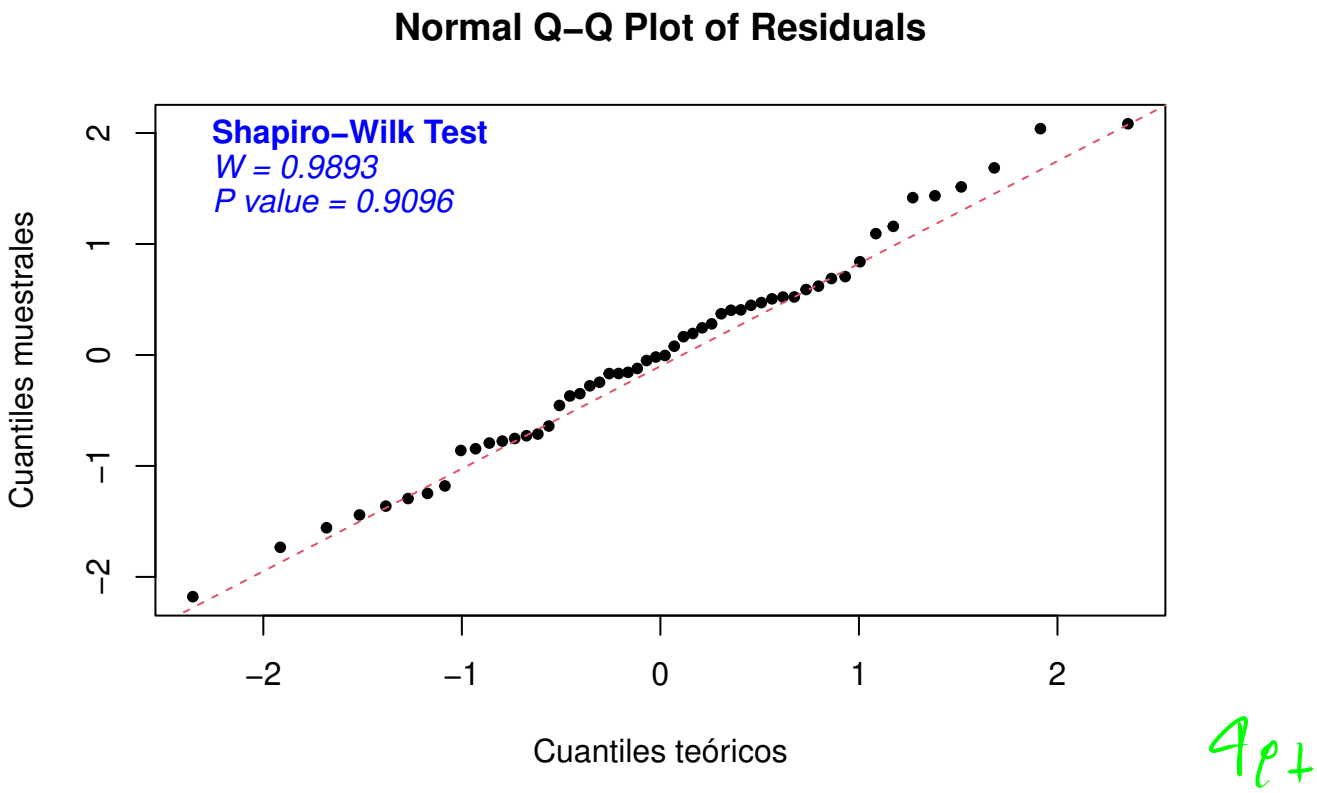

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Tomando en cuenta el valor p no se rechaza H0, se podre decir que hay normalidad, sin embargo la gráfica nos muestra patrones irregulares en la parte superior de recta, se evidencia que la cola esta un poco alejada de la línea de tendencia marcada en la gráfica lo que supone que el supuesto de normalidad no se este cumpliendo como lo indica el valor p. Sin embargo la gráfica de comparación de cuantiles permite ver patrones irregulares que posiblemente se deben a datos atípicos, pero son leves, así que, a partir del gráfico tampoco se rechaza la hipótesis nula.

4.1.2. Varianza constante

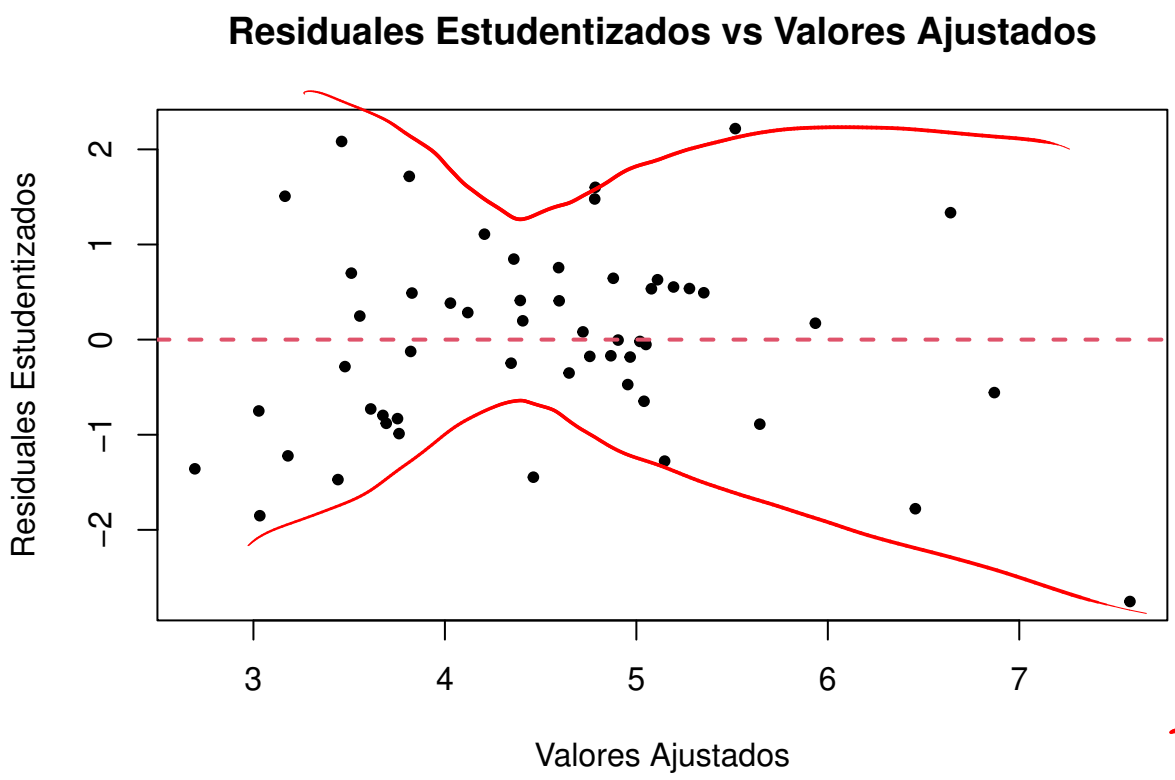

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que no hay patrones en los que la varianza aumente, decrezca ni un comportamiento que permita descartar una varianza constante, al no haber evidencia suficiente en contra de este supuesto se acepta como cierto. Además es posible observar media 0. Aunque La Varianza efectivamente no se cumple pero el análisis hecho, no explica el hecho expuesto lo suficientemente claro.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

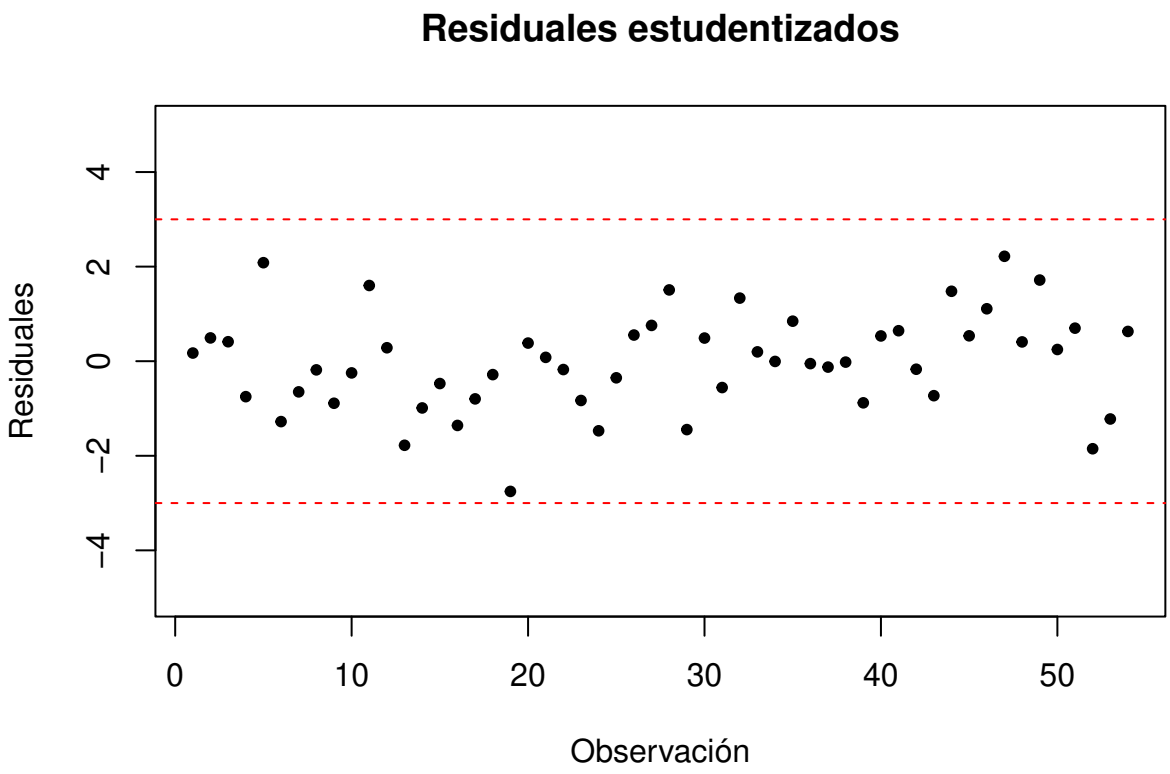

Figura 3: Identificación de datos atípicos

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

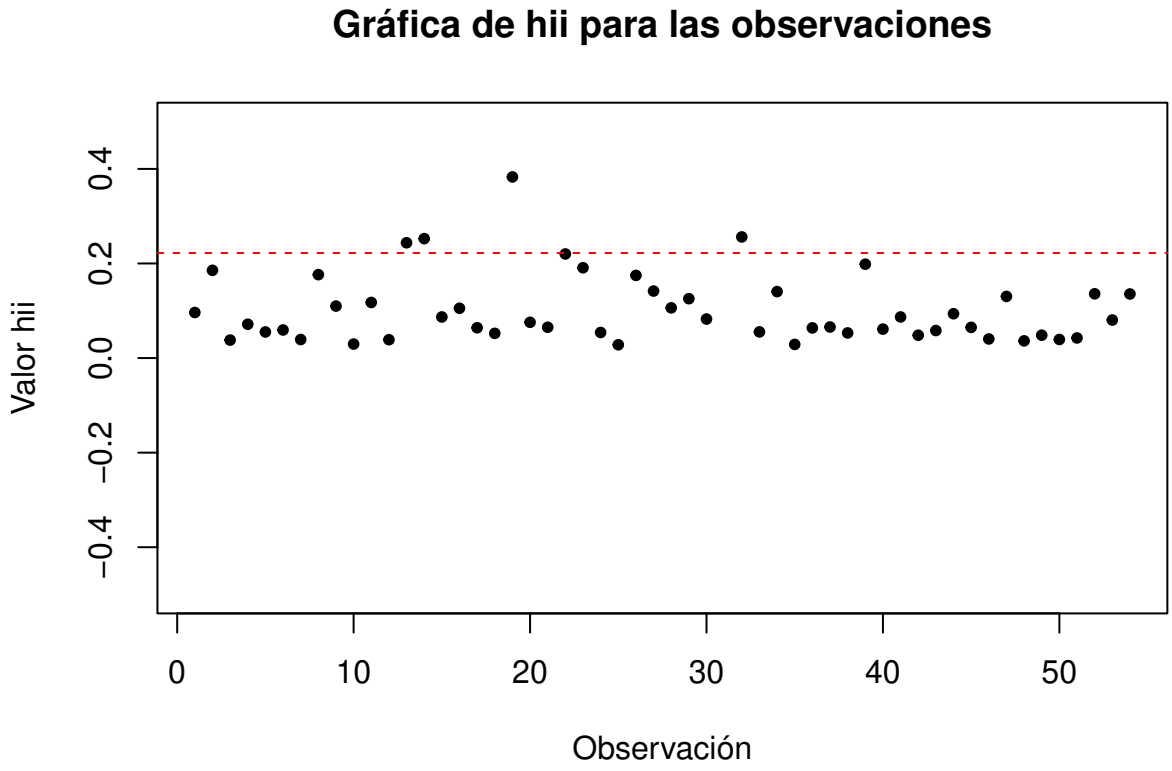

Figura 4: Identificación de puntos de balanceo

Tabla 5: Resumen de diagnostico

	$Res.stud$	$Cook.D_i$	$h_{ii}value$	$Dffits$
13	-1.7782	0.1699	0.2438	-1.0338
14	-0.9884	0.0550	0.2525	-0.5744
19	-2.7533	0.7844	0.3830	-2.3394
31	-0.5571	0.0674	0.5658	-0.6313
32	1.3343	0.1022	0.2561	0.7895

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 0.22$, se puede apreciar que existen 6 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 Bajo el criterio de puntos de balanceo y generan un aumento en la variabilidad en el modelo RLM.

4.2.3. Puntos influenciales

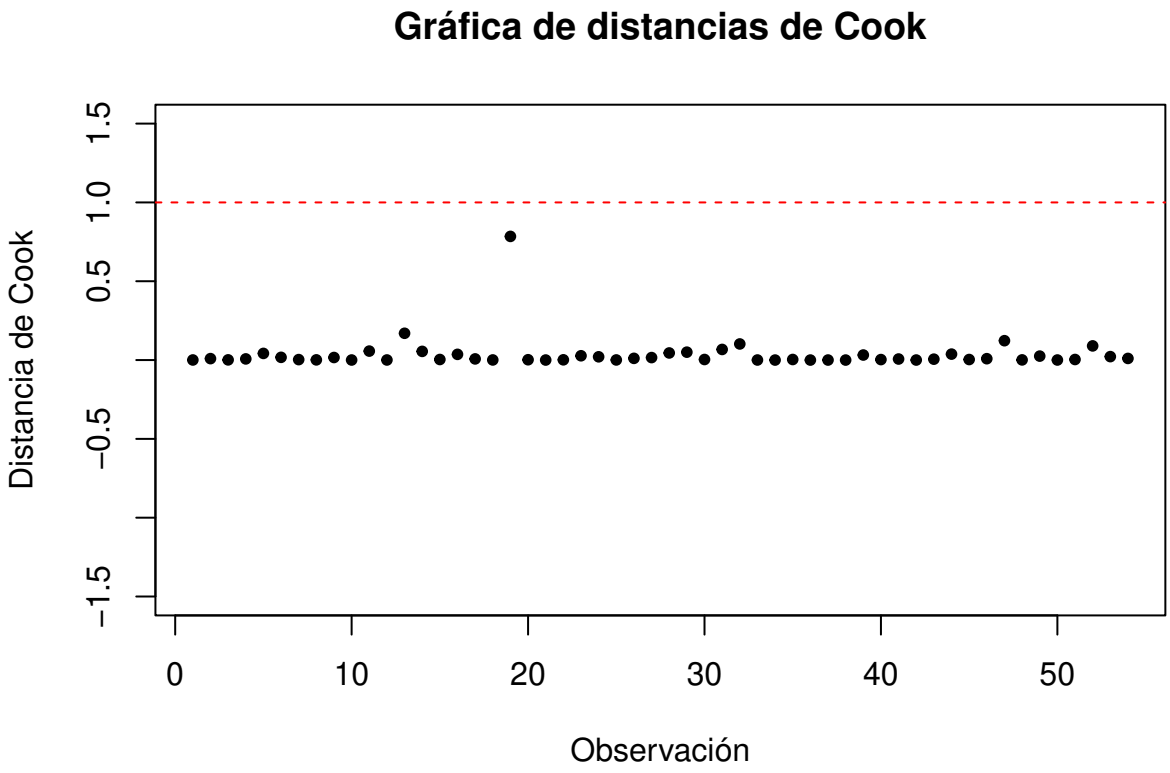

Figura 5: Criterio distancias de Cook para puntos influenciales

Como se puede visualizar en la gráfica de distancias de Cook, donde la línea punteada roja representa el valor 1, se puede inferir que no hay medidas de influencia a partir del criterio $Di > 1$

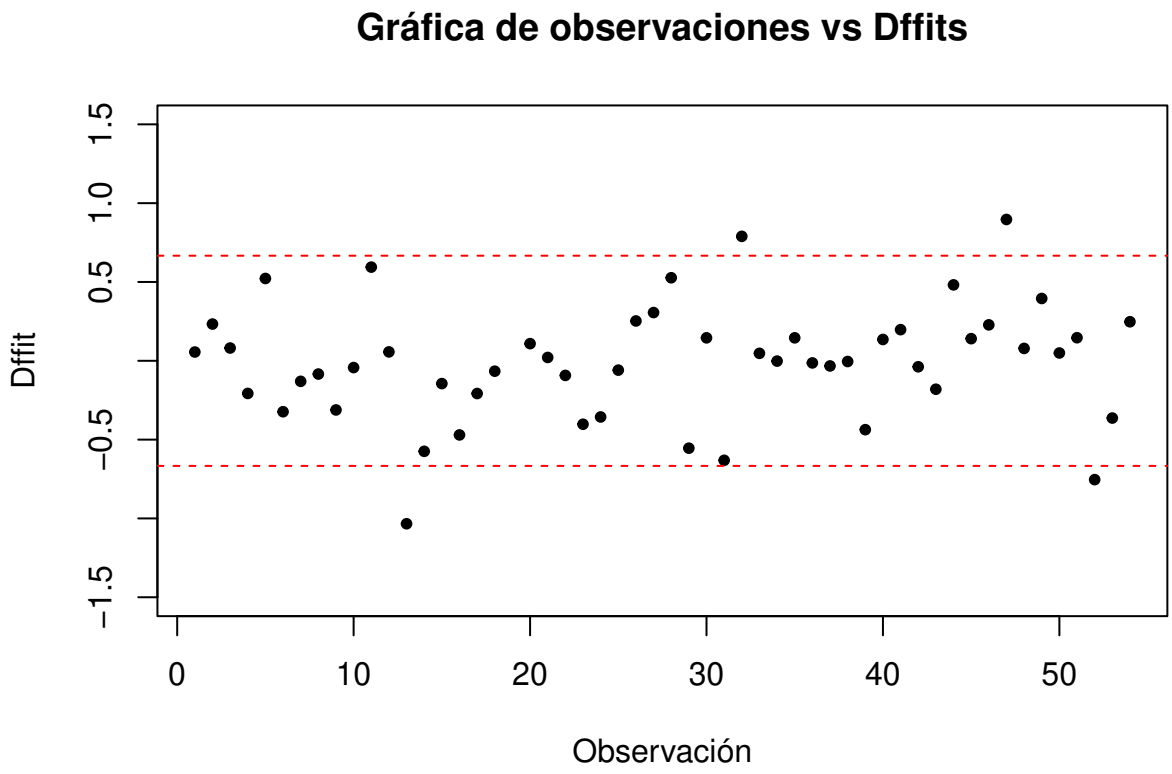

Figura 6: Criterio Dffits para puntos influenciales

2 pt

Tabla 6: Resumen de diagnostico

	$Res.stud$	$Cook.D_i$	$h_{ii}value$	$Dffits$
13	-1.7782	0.1699	0.2438	-1.0338
19	-2.7533	0.7844	0.3830	-2.3394
32	1.3343	0.1022	0.2561	0.7895
47	2.2182	0.1228	0.1303	0.8967
52	-1.8514	0.0898	0.1359	-0.7539

→ No se ve

Cook son...?

son 4

Como se puede ver,las observaciones 11, 14, 20 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 0.42070311619$, 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

3 pt

El modelo es válido, ya que se cumplen los supuestos del error, también cumple con los supuestos extremos donde el análisis de este puede decir si pueden afectar en cierta medida a los mismo el modelo y por lo que se puedo evidenciar no afecta significativamente en estos.