3,65

Trabajo 1

Estudiantes

Doris Alejandra Ramos Figueroa
Jhon Sebastian Chidiak Olaya
Laura Valentina Rincon Guataquira

Docente

Julieth Veronica Guarin Escudero

Asignatura

Estadística II

Sede Medellín
30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. pregunta 1

Teniendo en cuenta la base de datos equipo 13, nuestro modelo de RLM que explica el porcentaje de la eficacia en el control de infecciones hospitalarias(Y:Riesgo de infeccion) en terminos de las variables:

- Duraccion de la estadia(X_1).
- Rutina de cultivos(X_2).
- Numero de camas(X_3).
- Censo promedio diario(X_4).
- numero de enfermeras(X_5).

El modelo se propone a continuacion:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + + \beta_5 X_{5i} + \varepsilon_i; \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2); \ 1 \leq i \leq 55$$

1.1. Modelo de regresion

Al ajustar el modelo el modelo que explica el porcentaje de la eficacia en el control de infecciones hospitalarias, se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores de los coeficientes estimados

	V.Parámetro
β_0	-0.0058
β_1	0.1675
β_2	-0.0009
β_3	0.0619
β_4	0.0154
β_5	0.0017

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.0058 + 0.1675 X_{1i} + -9 \times 10^{-4} X_{2i} + 0.0619 X_{3i} + 0.0154 X_{4i} + 0.0017 X_{5i}$$

Donde $1 \leq i \leq 55$

1.2. Significancia de la regresión

Se plantea el siguiente Juego de Hipotesis, para analizar la significancia de la regresion:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_a : \text{Algún } \beta_j \neq 0 \; para \; j = 1, 2, 3, 4, 5 \end{cases}$$

Cuyo estadistico de prueba es:

$$F_0 = \frac{MST}{MSE} \overset{H_0}{\sim} f_{5,'rnrow(35)-6'} \tag{1}$$

Para la significancia de la regresión tomaremos los resultados de la tabla Anova presentada a continuacion:

Cuadro 2: Tabla ANOVA significancia de la regresión

	Suma cuadratica	Grados de libertad	Cuadrado medio	F_0	Valor-P
Modelo LRM	69.4276	5	13.885513	19.6988	1.03971e-10
Error	34.5397	49	0.704892		

De acuerdo con la tabla anova, con un nivel de significancia del $\alpha = 0.05$, por el criterio del valor-p encontrado en la anterior tabla anova (tabla 2). Podemos determinar que nuestro modelo rechaza Ho debido a que Vp$< \alpha = 0.05$ yes significativo ya que existe algun parametro distinto de cero. . Al rechazarla probamos que existe una relación de regresión, sin embargo, esto no garantiza que el modelo resulte útil para hacer predicciones.

1.3. Significancia de los parámetros

En el siguiente cuadro se presenta información de los parámetros, la cual permitirá 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	-0.0058	1.6503	-0.0035	0.9972
β_1	0.1675	0.0735	2.2799	0.0270
β_2	-0.0009	0.0309	-0.0304	0.9759
β_3	0.0619	0.0141	4.3929	0.0001
β_4	0.0154	0.0071	2.1749	0.0345
β_5	0.0017	0.0006	2.6766	0.0101

Observando los resultados obtenidos, podemos determinar que con un nivel de significancia $\alpha = 0.05$ los parámetros β_1, β_3,β_4 y β_5 son significativos porque sus valores p son menores a $\alpha = 0.05$. Y los parametros β_0 y β_2 no son significativos debido a que su p es mayor a $\alpha = 0.05$.

1.4. Interpretación de los parámetros

A continuacion se hara la interpretacion de los parametros que son significativos, ya que los otros parametros pueden ser eliminados el no aportar el modelos de la grasa corporal

- $\hat{\beta}_1$= 0.1675: Indica que por cada unidad que aumenta la duracion de la estadia, la probabilidad promedio estimada de adquirir una infeccion en el hospital aumenta en un 0.1675 unidades cuando las demás variables predictoras se mantienen fijas.
- $\hat{\beta}_3$= 0.0619: Indica que por cada unidad que aumenta el numero de camas, la probabilidad promedio estimada de adquirir una infeccion en el hospital aumenta en un 0.0619 unidades cuando las demás variables predictoras se mantienen fijas.
- $\hat{\beta}_4$= 0.0154: Indica que por cada unidad que aumenta el censo promedio diario, la probabilidad promedio estimada de adquirir una infeccion en el hospital aumenta en un 0.0154 unidades cuando las demás variables predictoras se mantienen fijas.
- $\hat{\beta}_5$= 0.0017: Indica que por cada unidad que aumenta el numero de enfermeras, la probabilidad promedio estimada de adquirir una infeccion en el hospital aumenta en un 0.0017 unidades cuando las demás variables predictoras se mantienen fijas.

Coeficiente de determinación múltiple R^2

El coeficiente de determinacion del modelo es $R^2 = 0.6678$ lo que indica que el modelo explica 66.78 % de la variabilidad total en el modelo riesgo de infeccion en el hospital y es explicada por el modelo propuesto de regresion lineal multiple.

2. Pregunta 2

2.1. Planteamiento prueba de hipotesis y modelo reducido

Los tres parametros cuyos valores P fueron los más altos corresponden a β_1, β_2,β_4. Debido a esto, se propone la siguiente prueba de hipótesis:

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

Se presenta la siguiente tabla con el resumen de todas las regresiones para plantear el estadístico de prueba:

Cuadro 4: Resumen de todas las regresiones

	Suma cuadratica del error	Predictores del modelo
M.Completo	34.540	X1 X2 X3 X4 X5
M.Reducido	45.963	X3 X5

El modelo completo es el definido en la sección 1.1, y el modelo reducido para la prueba de significancia del subconjunto es:

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 55$$

2.2. Estadístico de prueba y conclusiones

Se construye el estadístico de prueba como:

$$F_0 = \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,49}$$

$$F_0 = \frac{(45.963 - 34.540)/3}{0.9380204081632} \overset{H_0}{\sim} f_{3,49}$$

$$F_0 = 4.06 \overset{H_0}{\sim} f_{3,49}$$

Comparamos F_0 con $f_{0.95,3,49} = 2.7939$, se evidencia que $F_0 = 4.09 > f_{0.95,3,49}$, por lo cual a partir de la prueba de F realizada se concluye que es mayor, entonces se rechaza la hipotesis nula planteada y tambien nos indica que el subconjunto es significativo. Por lo tanto, es posible descartar las variables del M.Reducido.

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:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_5;\ \beta_2 = \beta_3 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

reescribiendo matricialmente:

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

2 pt

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

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1 X^*_{1i} + \beta_2 X^*_{2i} + \beta_4 X_{4i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 55$$

1 pt

Donde $X^*_{1i} = X_{1i} + X_{5i}$ y $X^*_{2i} = X_{2i} + X_{3i}$

3.2. Estadístico de prueba

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,49}$$

2 pt

$$F_0 = \frac{(SSE(MR) - 34.540)/2}{0.705} \overset{H_0}{\sim} f_{2,49}$$

4. Pregunta 4

11 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

3 pt

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

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

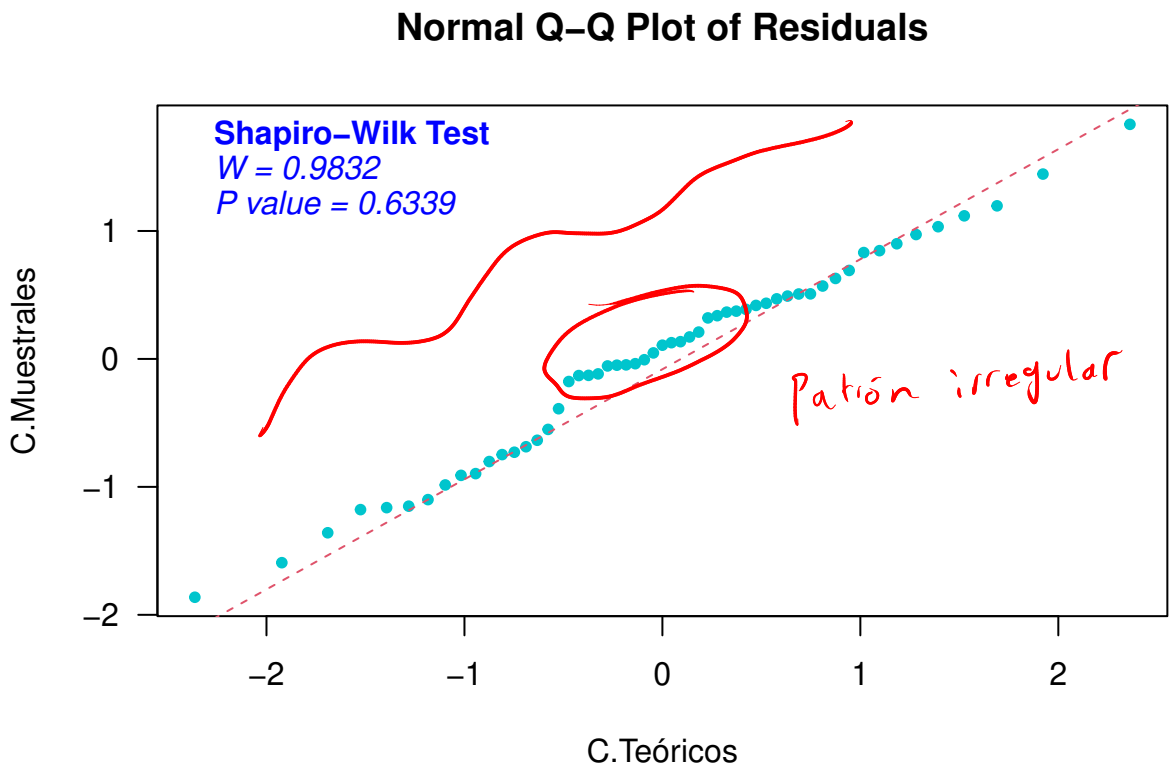

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Todo modelo de regresión debe cumplir con sus supuestos, en este caso analizaremos el supuesto de normalidad con la prueba de shapiro wilk y para asegurar; con una gráfica.

En la gráfica de normalidad Q-Q plot of residuals Miramos en la parte superior el test y la gráfica. Según el test de shapiro wilk el valor p es aproximadamente 0.6339, se trabaja con un $\alpha = 0.05$ entonces 0.6339>0.05, se acepta H_0 cero y se supone que debe distribuir normal, sin embargo, cuándo miramos esa gráfica se observa horas más pesadas y patrones anormales lo que nos haría suponer que no se cumple el supuesto de normalidad. Teniendo dos análisis acerca de la normalidad optamos por elegir la gráfica ya que tiene más qué divinidad un análisis gráfico qué un test generalizado.

4.1.2. Varianza constante

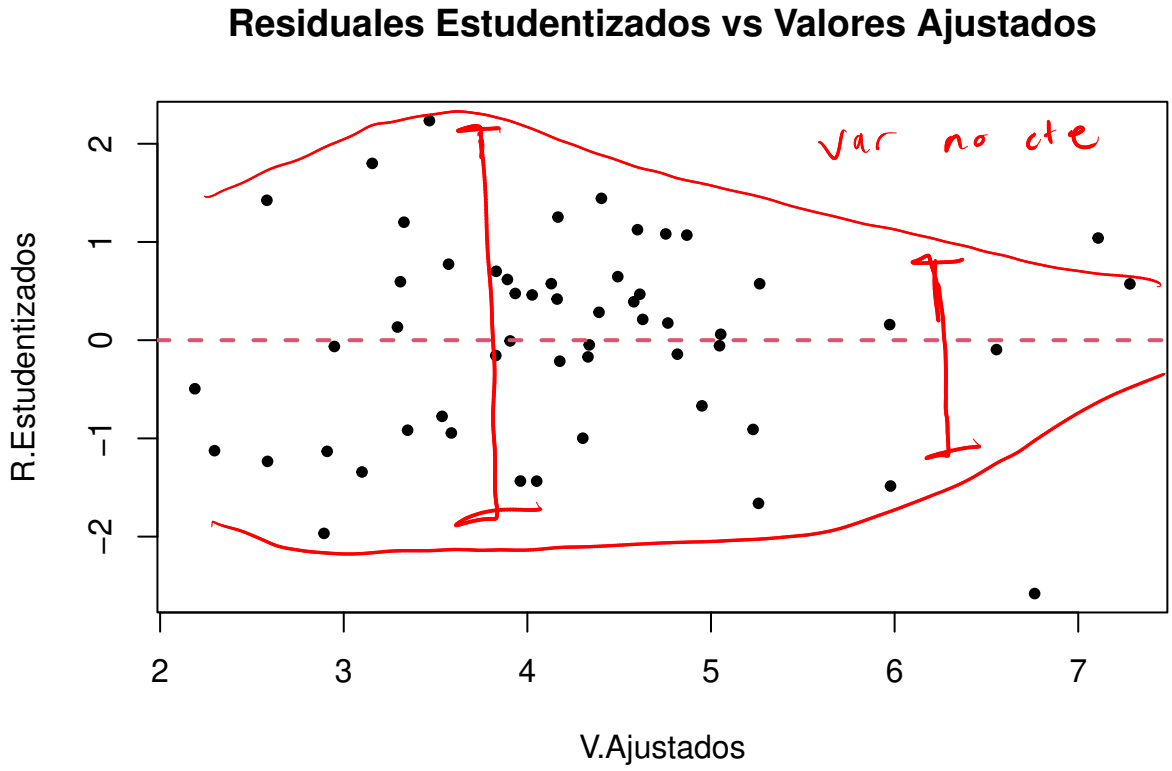

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Después de analizar el supuesto de normalidad, miraremos el otro supuesto acerca de si el modelo de regresión múltiple tiene varianza constante. Analizamos la gráfica, observamos la forma de nube de puntos y no hay un patrón claro, sospecha de una posible heterocedasticidad. Identificando el patrón en la forma de la nube de puntos hay una línea que cruza los residuos; esto definitivamente indica la presencia de heterocedasticidad en el modelo y la ausencia del supuesto de varianza constante supone que la variabilidad de los errores no es la misma en todas las combinaciones de los predictores. Debido a todo lo expuesto, llegamos a la conclusión que no tiene varianza constante.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

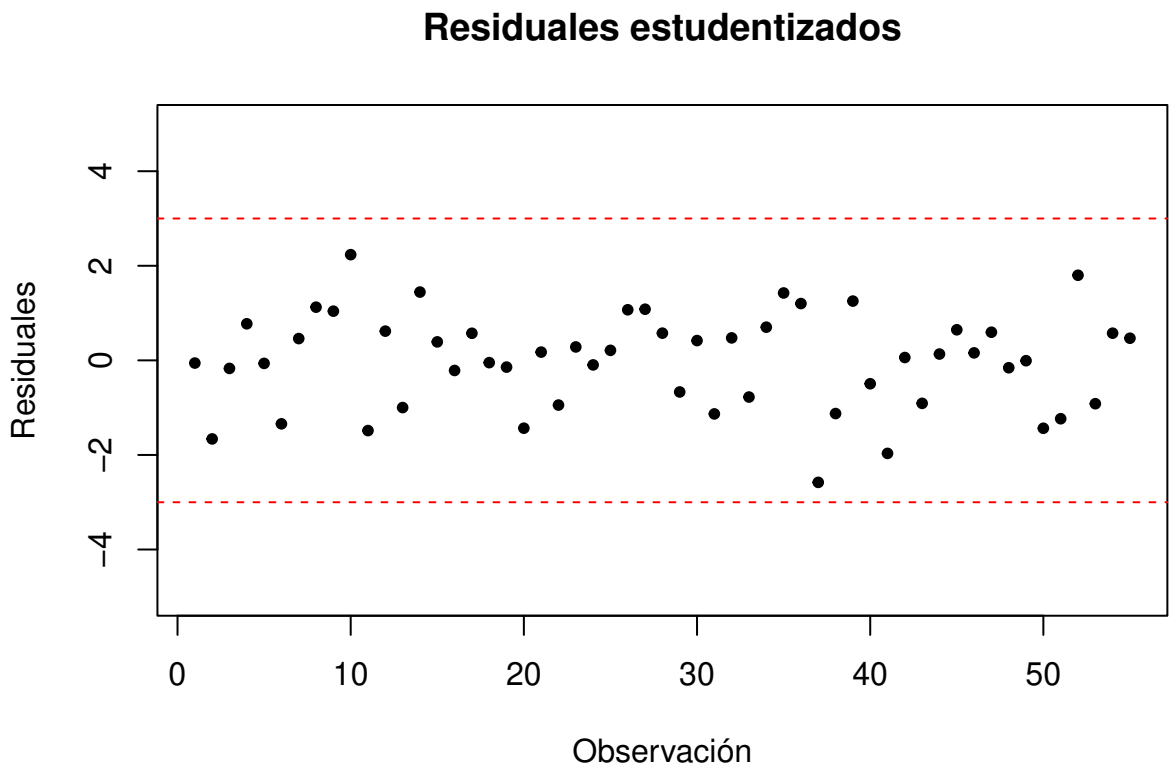

Figura 3: Identificación de datos atípicos

Verificaremos las observaciones con una gráfica que presenta datos atípicos de los residuales estudentizados por las observaciones en el eje x y los residuales en el eje y. El criterio nos dice que |restud|>3, en la gráfica los puntos de nubes se mantienen entre (-3,3), a causa de eso no hay datos atípicos en el conjunto de datos.

4.2.2. Puntos de balanceo

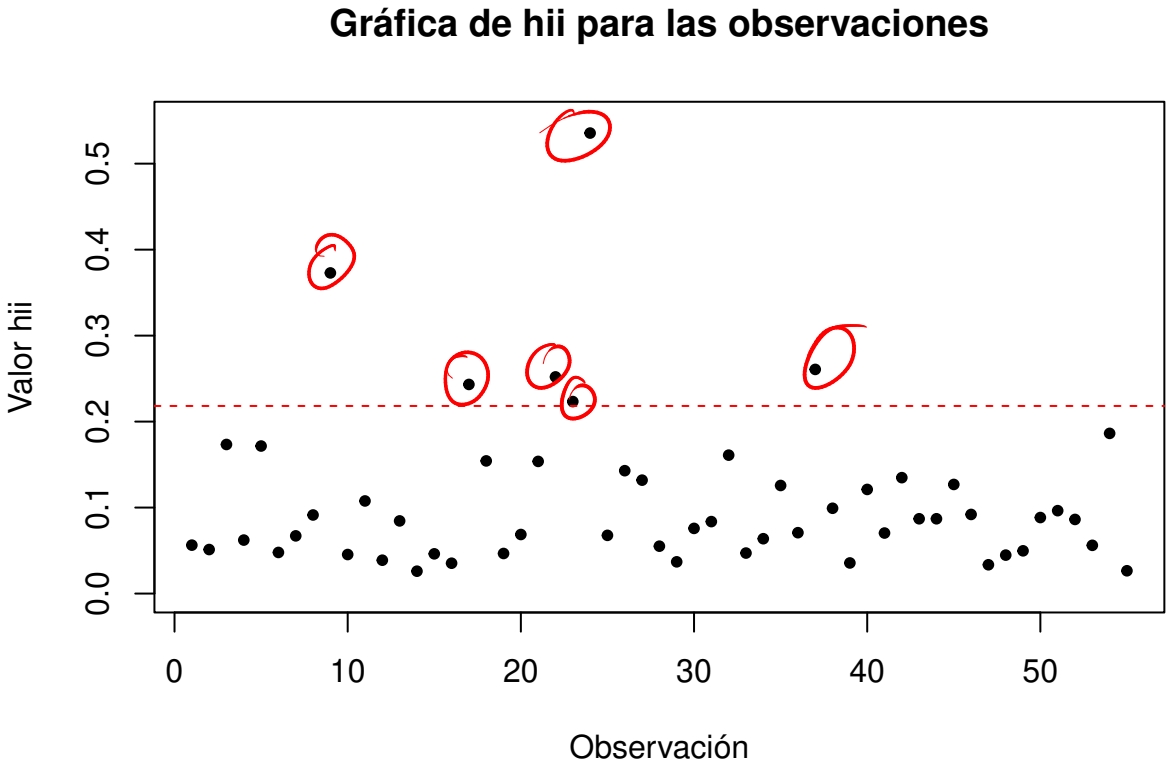

Figura 4: Identificación de puntos de balanceo

Cuadro 5: Tabla de observaciones de Balanceo

	Errores Estudentizados	D.Cook	Valor hii	DFFITS
9	1.0402	0.1072	0.3729	0.8027
17	0.5728	0.0176	0.2432	0.3225
22	-0.9450	0.0502	0.2520	-0.5480
23	0.2837	0.0039	0.2233	0.1507
24	-0.0958	0.0018	0.5356	-0.1018
37	-2.5800	0.3911	0.2607	-1.6311

Para la figura 4 examinamos que sí hay datos atípicos porque hay puntos de balanceo fuera de las líneas y de (-3,3), estos datos según la tabla de puntos de balanceo son 9,17,22,23,24 y 37. Y es que estos valores extremos desvían significativamente el valor de la muestra, por lo tanto, su inclusión puede sesgar la interpretación de los resultados y en lo posible, sería

mejor eliminarlos aunque es importante tenerlos; siempre y cuando no sean muchos ya que eso representa un error en la medición o recolección de datos, y es necesario que hallan para que proporcionen información valiosa sobre la variabilidad de los datos.

Lean la teoría de qué es un punto de balanceo, lo que causa y cómo identificarlos.

4.2.3. Puntos influenciales

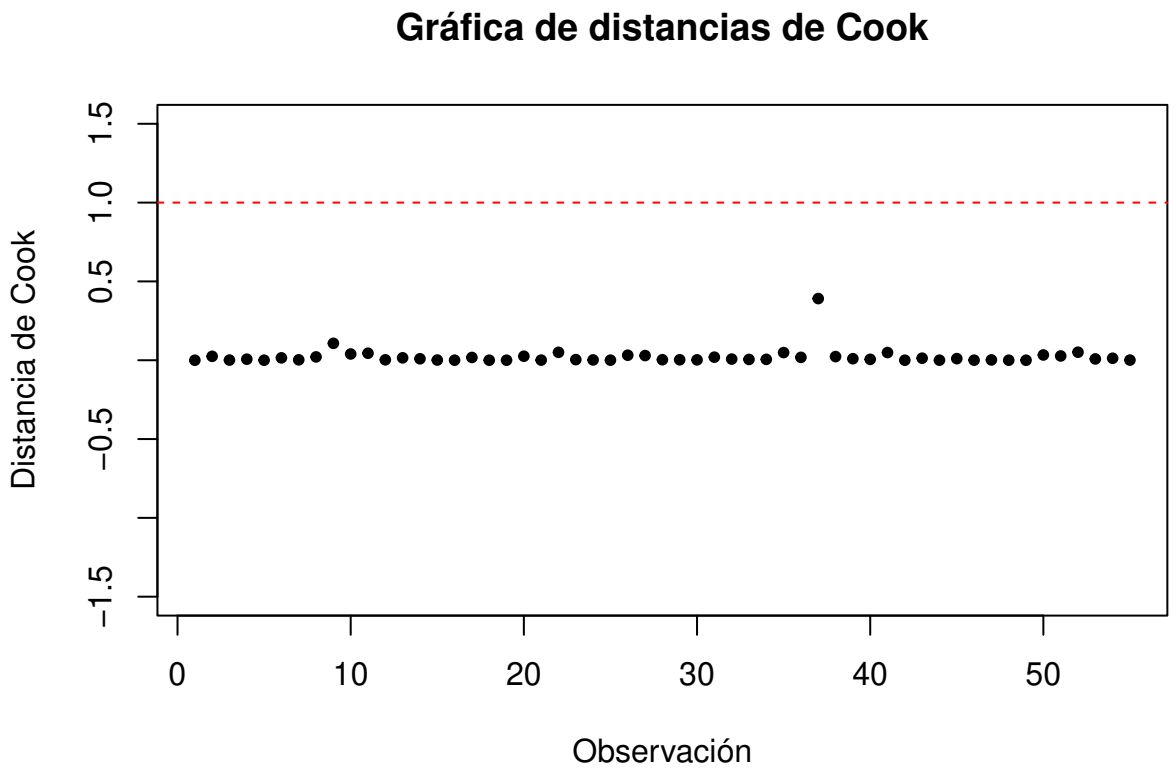

Figura 5: Criterio distancias de Cook para puntos influenciales

Con la figura 5 de distancia de Cook, con observaciones en el eje x y distancia Cook(y). El criterio de Cook dice que observaciones i será influencia le si Di>1, en la gráfica ningún punto es mayor a 1, por lo tanto no hay puntos influenciados.

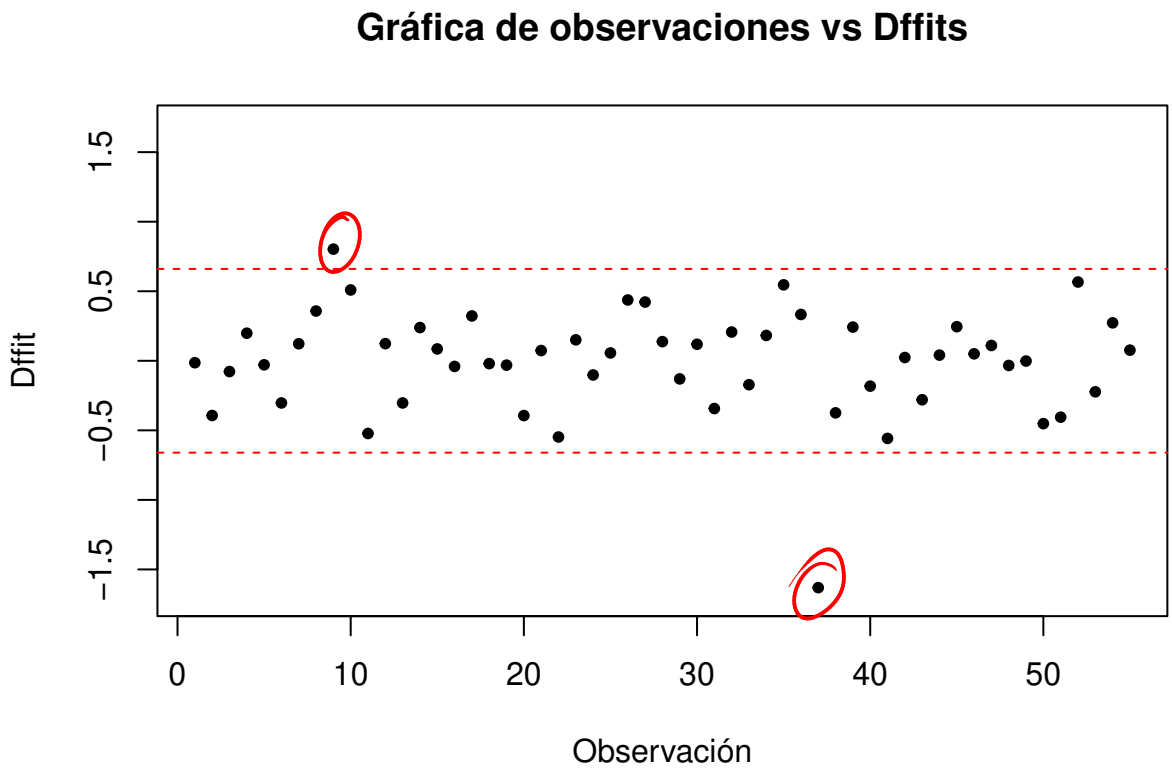

Figura 6: Criterio Dffits para puntos influenciales

0,5 pt

Cuadro 6: Punto influencial

	Errores Estudentizados	D.Cook	Valor hii	DFFITS
9	1.0402	0.1072	0.3729	0.8027
37	-2.5800	0.3911	0.2607	-1.6311

¿influenciales

En la figura 6 por medio deo criterio Dffits para puntos influenciables, se presenta una gráfica de observaciones vs Dffits, con observación en eje x y Diffits en el eje y. El criterio de Dffits nos expresa que para cualquier punto () por lo que las observaciones 9 y 37 son puntos influenciados.

¿Qué es un punto influencial ¿copy + paste! ¿plagio de nuevo? según este criterio y qué causa!

4.3. Conclusión

1 pt

- De este modelo

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + +\beta_5 X_{5i} + \varepsilon_i; \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2); \ 1 \leq i \leq 55$$

se puede deducir que $beta_1$ $beta_4$ $beta_4$ $beta_5$ que respetivamente son ((Duraccion de la estadia(X_1).Numero de camas(X_3),Censo promedio diario(X_4),numero de enfermeras(X_5)) Nos explican el riesgo de y son indispensables para explicar el riesgo de infectarce en el hospital ya que son significativas.

- Si la causa de los valores atípicos se puede identificar y corregir, se puede proceder a realizar un análisis normal de los datos restantes. Sin embargo, si la causa de los valores atípicos no se puede identificar, puede ser necesario considerar el uso de técnicas estadísticas más avanzadas, como la regresión Robusta o el análisis de resúmenes no paramétricos.

  En cualquier caso, es importante tener en cuenta que los valores atípicos no siempre deben eliminarse de la muestra, ya que pueden proporcionar información valiosa sobre la variabilidad de los datos. En lugar de eliminarlos, es recomendable tratarlos en el análisis para entender su impacto en los resultados y determinar si son valores extremos legítimos o errores de medición.

No responden nada sobre validez del modelo