Trabajo 1

Estudiantes

John Jairo Alpala Cumbalaza

3,4

Equipo 55

Docente

Carlos Mario Lopera

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín

5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

Teniendo en cuenta la base de datos brindada, en la cual hay 5 variables regresoras dadas por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 113$$

Donde:

- Y (Riesgo de infección): probabilidad promedio estimada de adquirir una infección en el hospital, expresada en porcentaje.
- X1 (Duración de la estadía): duración promedio de la estadía de todos los pacientes en el hospital, medida en días.
- X2 (Rutina de cultivos): Mide la razón del número de cultivos realizados en pacientes sin síntomas de infección hospitalaria por cada 100 pacientes.
- X3 (Número de camas): Número promedio de camas en el hospital durante el periodo del estudio.
- X4 (Censo promedio diario): Indica el número promedio de pacientes en el hospital por día durante el periodo del estudio.
- X5 (Número de enfermeras): Promedio de enfermeras equivalentes a tiempo completo durante el periodo del estudio.

1.1. Modelo de regresión

Tras estimar el modelo de regresión lineal múltiple utilizando los datos de nuestra base de datos, los resultados de los coeficientes estimados son los siguientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	0.1644
β_1	0.2292
β_2	-0.0017
β_3	0.0339
β_4	0.0128
β_5	0.0015

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 0.1644 + 0.2292 X_{1i} + -0.0017 X_{2i} + 0.0339 X_{3i} + 0.0128 X_{4i} + 0.0015 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 113$$

1.2. Significancia de la regresión

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

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2,..., 5} \end{cases}$$

Cuyo estadístico de prueba es:

$$F_0 = \frac{MST}{MSE} \overset{H_0}{\sim} f_{5,63} \tag{1}$$

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	48.1300	5	9.62600	10.685	1.83601e-07
Error	56.7561	63	0.90089		

De la tabla Anova, se observa un valor P extremadamente pequeño (1.83601e-07), lo que significa que es improbable obtener un estadístico F tan grande como 10.685 bajo la hipótesis nula. por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $0 \leqslant j \leqslant 5$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$, por lo tanto la regresión es significativa.

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	0.1644	1.8626	0.0883	0.9300
β_1	0.2292	0.1004	2.2833	0.0258
β_2	-0.0017	0.0344	-0.0486	0.9614
β_3	0.0339	0.0133	2.5372	0.0137
β_4	0.0128	0.0065	1.9657	0.0537
β_5	0.0015	0.0007	2.2603	0.0273

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, pues sus P-valores son menores a α, lo que indica que estas variables tienen un impacto estadísticamente significativo en el Riesgo de infección.

1.4. Interpretación de los parámetros

$\hat{\beta}_1$: coeficiente para X_1 (Duración de la estadía) es 0.2292. Esto indica que, manteniendo todas las demás variables constantes, un aumento de una unidad en la Duración de la estadía está asociado con un aumento de 0.2292 en el Riesgo de infección.

$\hat{\beta}_3$: coeficiente para X_3 (Número de camas) es 0.0339. Esto sugiere que, manteniendo todas las demás variables constantes, un aumento de una unidad en el Número de camas está asociado con un aumento de 0.0339 en el Riesgo de infección.

$\hat{\beta}_5$: coeficiente para X_5 (Número de enfermeras) es 0.0015. Esto indica que, manteniendo todas las demás variables constantes, un aumento de una unidad en el Número de enfermeras está asociado con un aumento de 0.0015 en el Riesgo de infección.

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.290$, lo que significa que aproximadamente el 29 % de la variabilidad en el Riesgo de infección puede ser 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_3, X_5, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_3 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 3, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	56.756	X1 X2 X3 X4 X5
Modelo reducido	60.237	X1 X3 X5

Luego un modelo reducido para la prueba de significancia del subconjunto es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 113$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_3, \beta_5) - SSE(\beta_0, \cdots, \beta_5))/4}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{4,63} \\ &= \frac{60.237 - (56.756)/3}{0.90089} \\ &= \frac{41.318333333333}{0.90089} \\ &= 45.863807406196 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,63} = 2.7505$, se puede ver que $F_0 ? f_{0.95,1,45}$ y por tanto qué se rechaza ...

Es posible o no descartar las variables del subconjunto?

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

¿La duration de la estadía de todos los pacientes en el hospital (en días) depende del número de camas en el hospital? y ¿Número de pacientes en el hospital es equivalente al número de enfermeras? por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_3;\ \beta_4 = \beta_5 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1 X_{1i}^* + \beta_2 X_{2i} + \beta_4 X_{4i}^* + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 113$$

Donde $X_{1i}^* = X_{1i} + X_{3i}$ y $X_{4i}^* = X_{4i} + X_{5i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{100,63} \tag{3}$$

Aquí deben igualar después esa ecuación a sí misma con los valores conocidos reemplazados, es decir, el SSE(MF) y el MSE(MF).

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

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

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

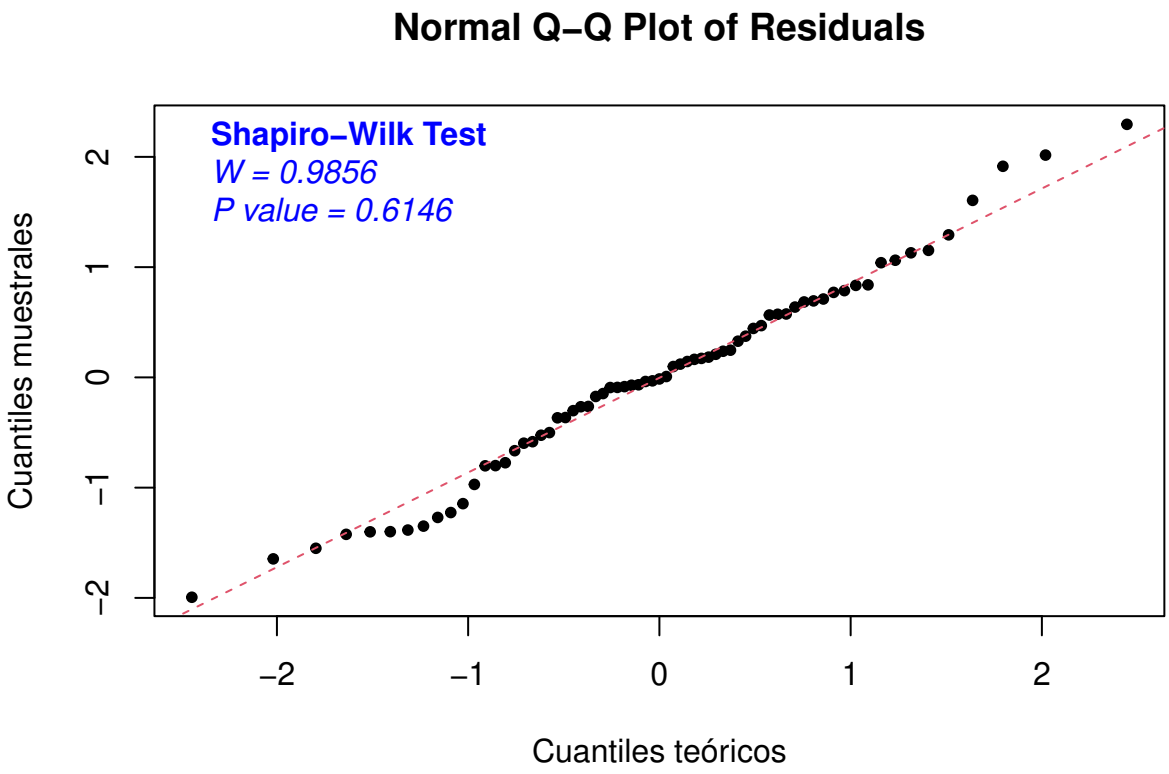

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.6146 y teniendo en cuenta que el nivel de significancia $\alpha = 0.05$, el P-valor es mucho mayor y por lo tanto, no se rechazaría la hipótesis nula, es decir que los datos distribuyen normal con media μ y varianza σ^2, sin embargo la gráfica de comparación de cuantiles permite ver colas más pesadas y patrones irregulares, al tener más poder el análisis gráfico, se termina por rechazar el cumplimiento de este supuesto. Ahora se validará si la varianza cumple con el supuesto de ser constante.

4.1.2. Varianza constante

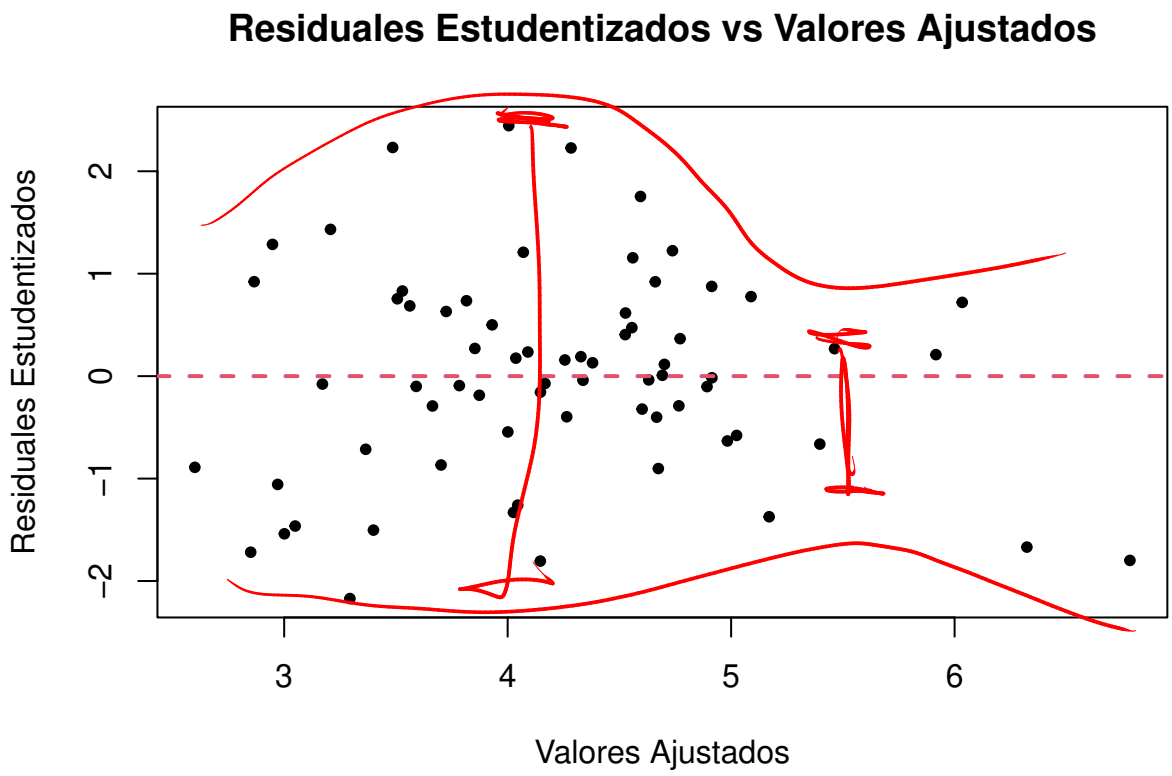

Figura 2: Gráfico residuales estudentizados vs valores ajustados

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.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

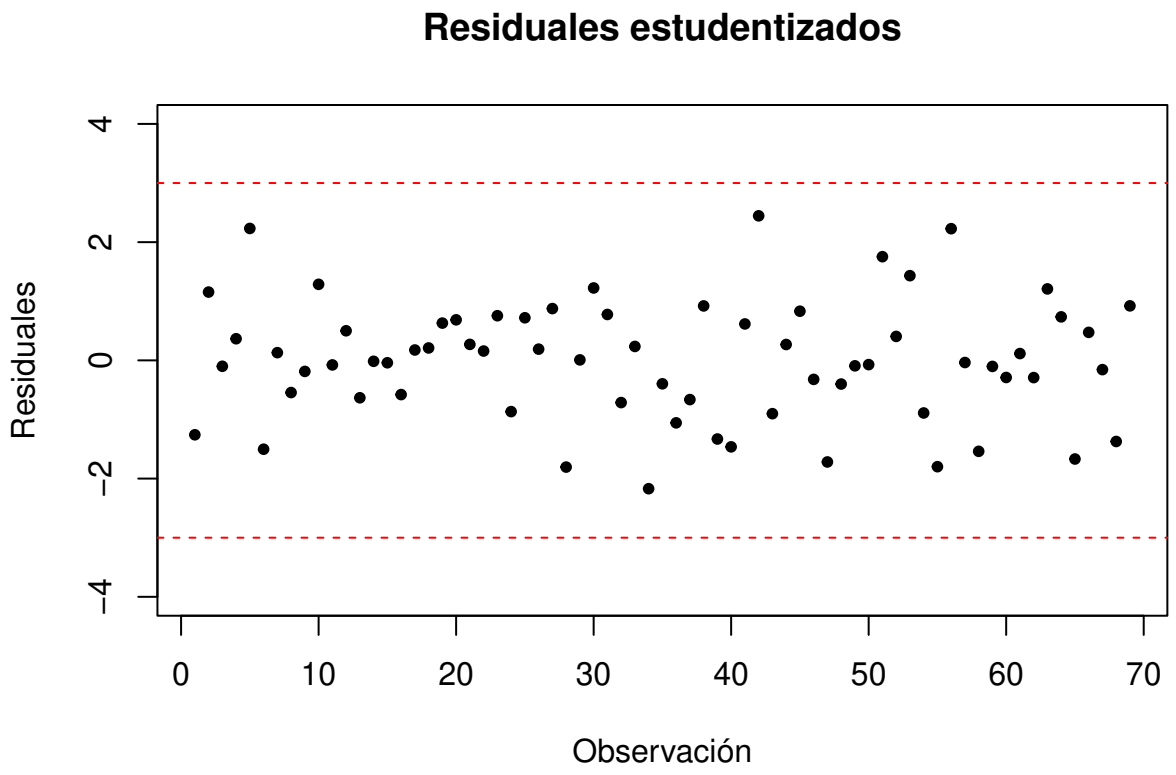

Figura 3: Identificación de datos atípicos

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

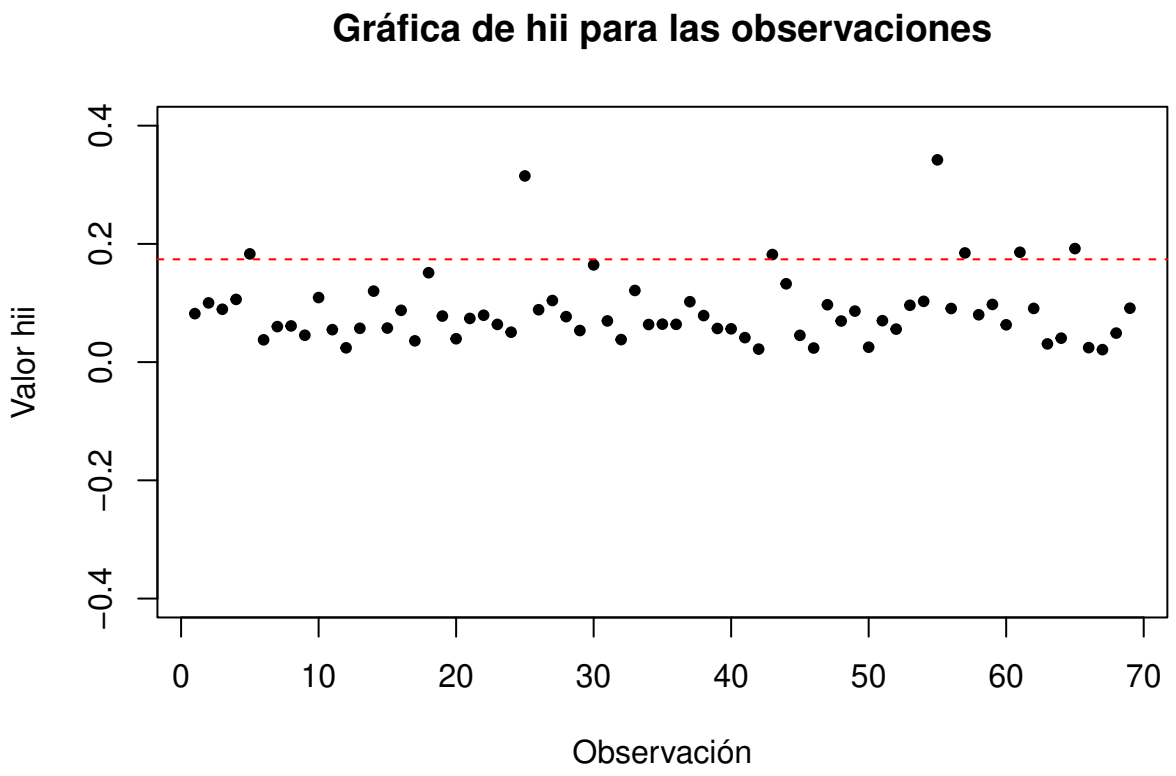

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 5    2.2318  0.1860    0.1831  1.0921
## 25   0.7203  0.0398    0.3150  0.4866
## 43  -0.9020  0.0302    0.1820 -0.4249
## 55  -1.7988  0.2806    0.3422 -1.3216
## 57  -0.0366  0.0001    0.1847 -0.0173
## 61   0.1151  0.0005    0.1858  0.0545
## 65  -1.6687  0.1103    0.1921 -0.8257
```

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$, se puede apreciar que existen 7 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.

4.2.3. Puntos influenciales

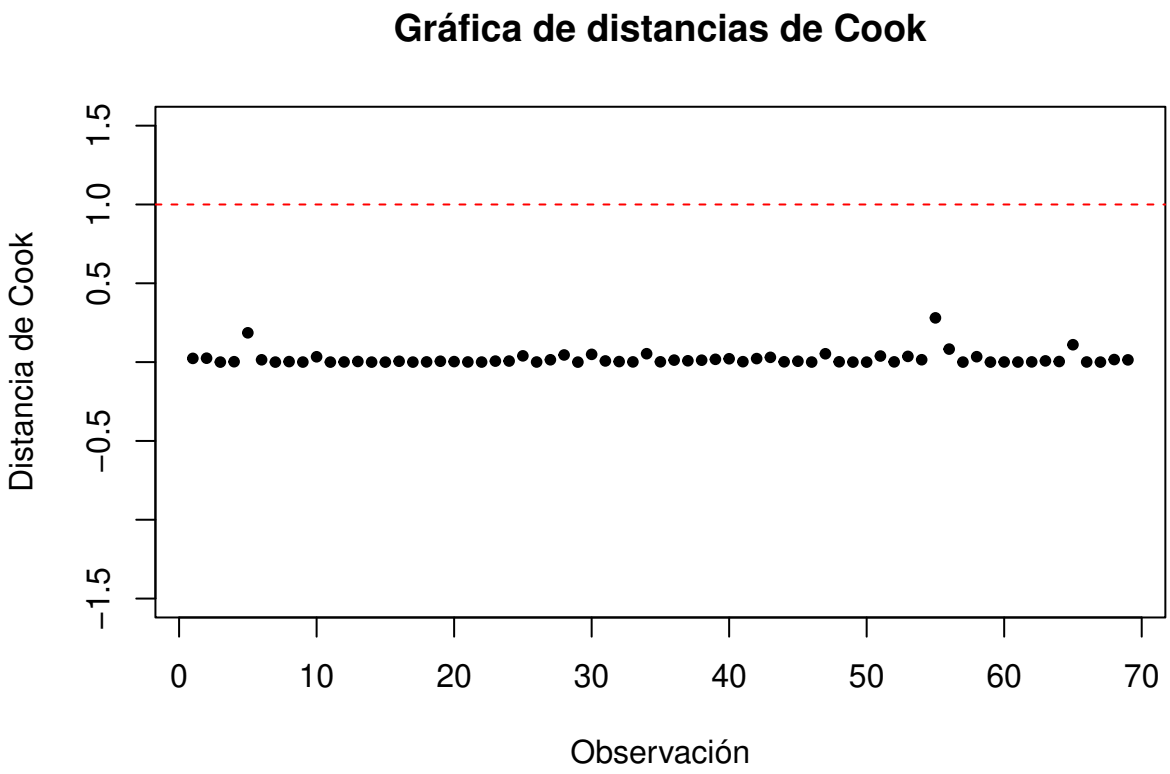

Figura 5: Criterio distancias de Cook para puntos influenciales

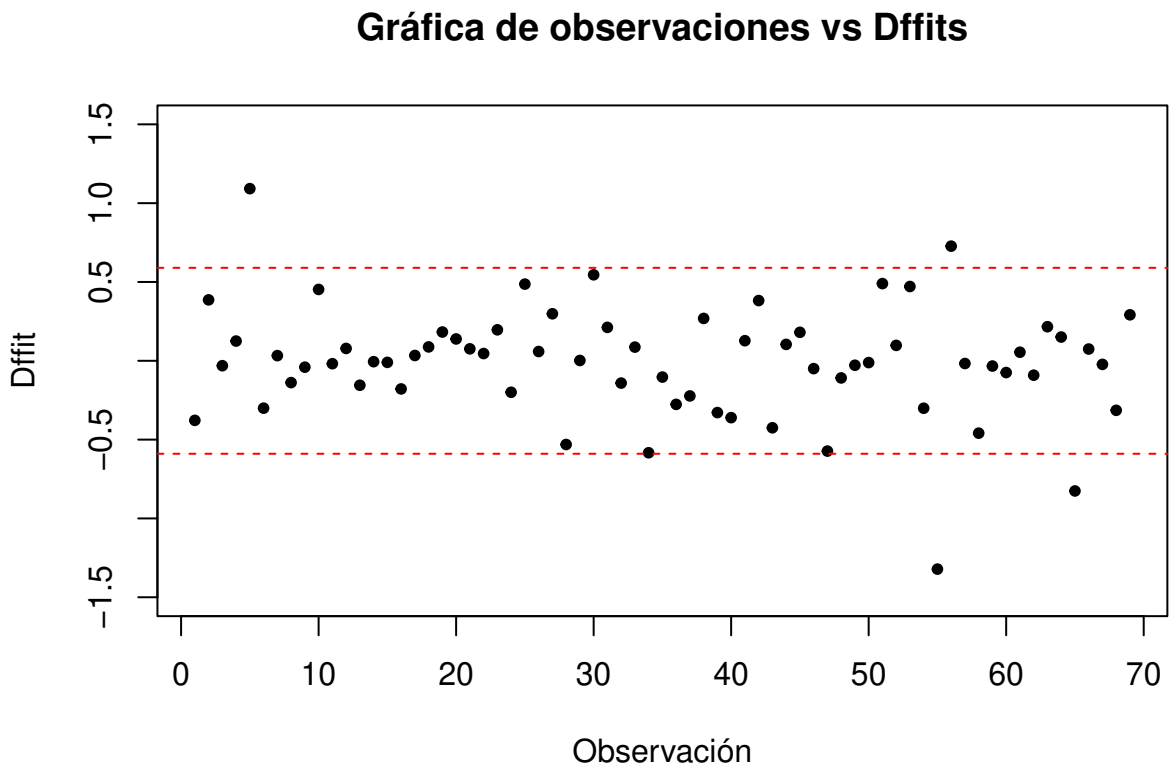

Figura 6: Criterio Dffits para puntos influenciales

Como se puede ver,las observaciones 5, 55, 56 y 65 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 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

1. El modelo de regresión lineal múltiple es estadísticamente significativo en general, y las variables predictoras X_1, X_3, X_5 son significativas para predecir el Riesgo de infección. El coeficiente de determinación múltiple R^2 de 0.290 indica que aproximadamente el 29 % de la variabilidad en el Riesgo de infección se explica mediante el modelo. Las interpretaciones de los coeficientes proporcionan información sobre cómo cada variable predictora contribuye al Riesgo de infección.

2. Se realizó un análisis de todas las regresiones posibles para evaluar la combinación de variables predictoras. Las combinaciones que incluían X1, X3 y X5 resultaron en los valores más altos de R^2 ajustado, lo que sugiere que estas tres variables son las más relevantes para explicar el riesgo de infección.

3. el modelo de regresión lineal múltiple desarrollado proporciona una herramienta valiosa para la predicción del riesgo de infección. Aunque el modelo es significativo y explica una parte importante de la variabilidad, se requieren análisis adicionales y una validación exhaustiva de supuestos para garantizar su validez y fiabilidad en aplicaciones clínicas o de investigación..

Válido o no?