Trabajo 1

3,9

Estudiantes

Andres Yair Carvajal Bolivar
Johan Sebastián Robles Rincón
Rusvelt Jose Meza San Martin
Jhoan Sebastian Murillo Villanueva

Equipo 59

Docente

Carlos Mario Lopera

Asignatura

Estadística II

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:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i};\ 1 \leqslant i \leqslant 69$$

De donde las variables del modelo son las siguientes:

- 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

1.1. Modelo de regresión

Al ajustar el modelo, se obtiene la siguiente tabla de coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
$\hat{\beta}_0$	-3.2633
$\hat{\beta}_1$	0.2768
$\hat{\beta}_2$	0.0599
$\hat{\beta}_3$	0.0485
$\hat{\beta}_4$	0.0085
$\hat{\beta}_5$	0.0012

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -3.2633 + 0.2768 X_{1i} + 0.0599 X_{2i} + 0.0485 X_{3i} + 0.0085 X_{4i} + 0.0012 X_{5i} + \varepsilon_i$$

1.2. Significancia de la regresión

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

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

Cuyo estadístico de prueba es:

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

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	58.1698	5	11.633952	12.3894	2.21129e-08
Error	59.1584	63	0.939021		

5 pt

Analizando 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 $0 \leqslant j \leqslant 5$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$ para j = 1, 2, . . . , 5, por lo tanto viendo este valor P, la regresion es significativa.

1.3. Significancia de los parámetros

Primero observemos el juego de hipotesis para la prueba individual de la significancia de los parametros.

$$\begin{cases} H_0 : \beta_j = 0 \\ H_a : \beta_j \neq 0 \text{ con } 0 \leqslant j \leqslant 5 \end{cases}$$

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	-3.2633	1.5600	-2.0918	0.0405
β_1	0.2768	0.0982	2.8195	0.0064
β_2	0.0599	0.0273	2.1960	0.0318
β_3	0.0485	0.0135	3.5883	0.0007
β_4	0.0085	0.0068	1.2524	0.2151
β_5	0.0012	0.0007	1.8486	0.0692

4 pt

También es significativo

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros $\hat{\beta}_1$ y $\hat{\beta}_3$ con 0.0064 y 0.0007 son significativos, pues sus P-valores son menores a el α dado.

1.4. Interpretación de los parámetros

6 pt

El β_0 no es interpretado dado que el 0 no se encuentra dentro de los valores, por lo tanto no tiene una cordenada

$$(X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}) = (0, 0, 0, 0, 0)$$

El β_1 nos indica que por cada dia que aumenta la estadia de los pacientes en el hospital, en promedio la probabilidad de riesgo de infeccion aumenta 0.2768.

El β_3 nos indica que en promedio por cada unidad que aumenta el numero promedio de camas en el hospital, en promedio la probabilidad de riesgo de infeccion aumenta 0.0485.

las demás constantes.

1.5. Coeficiente de determinación múltiple R^2

2pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.4958$, lo que significa que aproximadamente el 49.58 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

como se calcula?

2. Pregunta 2

4pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las variables regresoras con los valores P más alto en el modelo fueron X_2, X_4, X_5, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_2 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 2, 4, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	59.158	X1 X2 X3 X4 X5
Modelo reducido	67.696	X1 X3

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

2.2. Estadístico de prueba y conclusión

2pt

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_1, \beta_3) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5))/3}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)} \overset{H_0}{\sim} F_{3,63} \\ &= \frac{(67.696 - 59.158)/3}{0.939021} \\ &= 3.030816 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,63} = 2.7505$, se puede ver que $F_0 > f_{0.95,3,63}$

Por lo tanto, Se rechaza H_0. Es decir que almenos alguna de la variables (B2, B4 Y B5) tiene un efecto significativo en la variable dependiente en el modelo regresión. Por lo tanto, no descartamos las variables del subconjunto, pues estamos afirmando que almenos una de esas variables es importante y deberia considerarse en el modelo.

2pt

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si las variables predictoras X_2 y X_4 son colineales y las variables predictoras X_1 y X_3 presentan colinealidad en el modelo. por consiguiente se plantea la siguiente prueba de hipótesis:

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

lo que es equivalente a lo siguiente:

$$\begin{cases} \mathrm{H}_0 : \beta_2 - \beta_4 = 0, \ \beta_1 - \beta_3 = 0 \\ \mathrm{H}_1 : \text{Al menos una de las igualdades no se cumple} \end{cases}$$

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

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1 X_{1i}^* + \beta_2 X_{2,i}^* + \varepsilon_i \ , \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \ \ 1 \leqslant i \leqslant 69$$

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

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

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

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 shapiro-wilk, acompañada de un gráfico cuantil-cuantil:

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

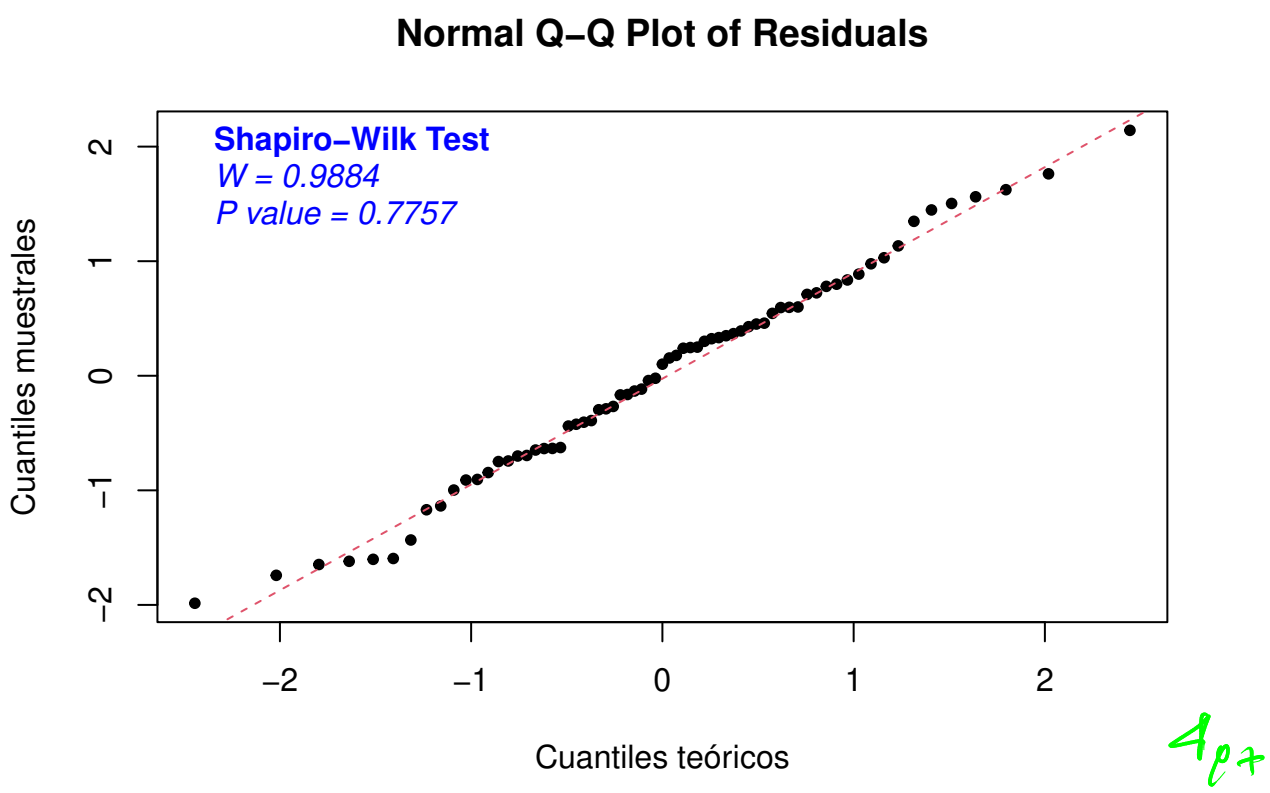


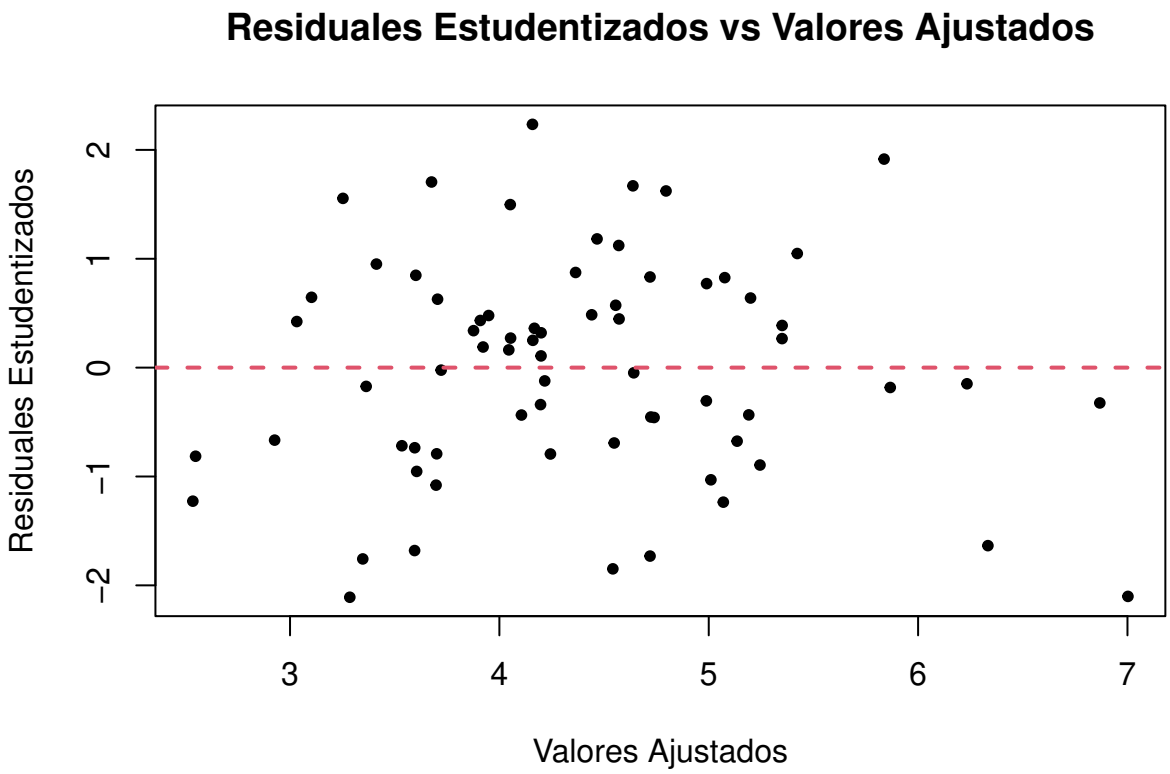
Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

El P-valor, aproximadamente 0.7757, es considerablemente mayor que nuestro nivel de significancia $\alpha = 0.05$. Esto sugiere que no tenemos evidencia suficiente para rechazar la hipótesis nula, indicando que nuestros datos se asemejan a una distribución normal.

Además, al observar el gráfico Q-Q plot, notamos que la mayoría de los puntos se encuentran cerca de la línea roja diagonal, lo que respalda la suposición de que nuestros datos siguen una distribución normal. En resumen, tanto el P-valor como el gráfico Q-Q plot sugieren que nuestros datos son en su mayoría consistentes con una distribución normal.

4.1.2. Varianza constante

$$H_0 : \mathbf{V}[\varepsilon_i] = \sigma^2 \qquad \textbf{vs} \qquad H_a : \mathbf{V}[\varepsilon_i] \neq \sigma^2$$



3pt

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se puede observarla mayoría de los datos entre los dos primeros tercios, especialmente el segundo tercio del gráfico, que el supuesto de varianza constante se cumple y que además se organizan alrededor del 0. En general, el supuesto se cumple pero vale mencionar de que en último tercio hay presencia de algunos datos que puedan afectar este supuesto dado que se encuentran alejados de los demás, y como mencionamos en la prueba de normalidad anterior, son posibles valores extremos.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

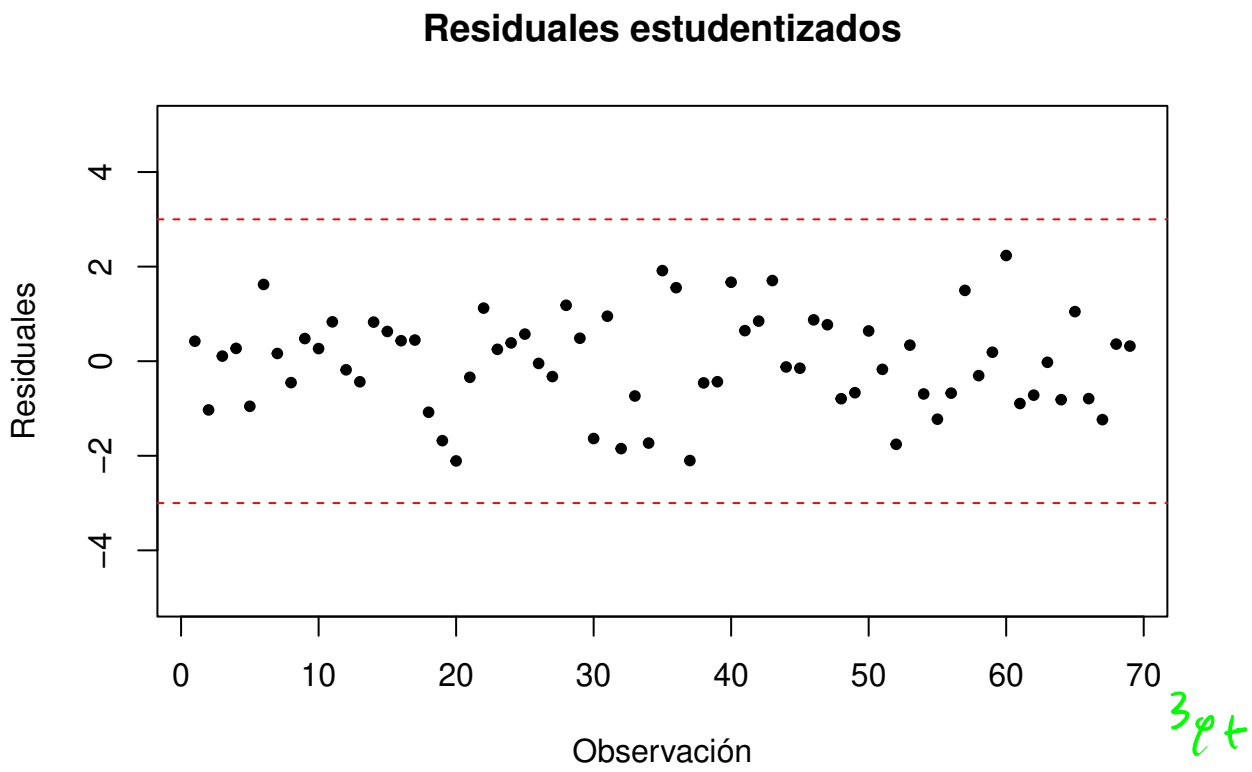


Figura 3: Identificación de datos atípicos

Como se puede observar en la gráfica de dispersión anterior, no hay datos atípicos en el modelo, ya que ninguno de los datos se encuentra por fuera del rango (-3,3), es decir que no cumplen que -3 $< r_i >$ 3.

4.2.2. Puntos de balanceo

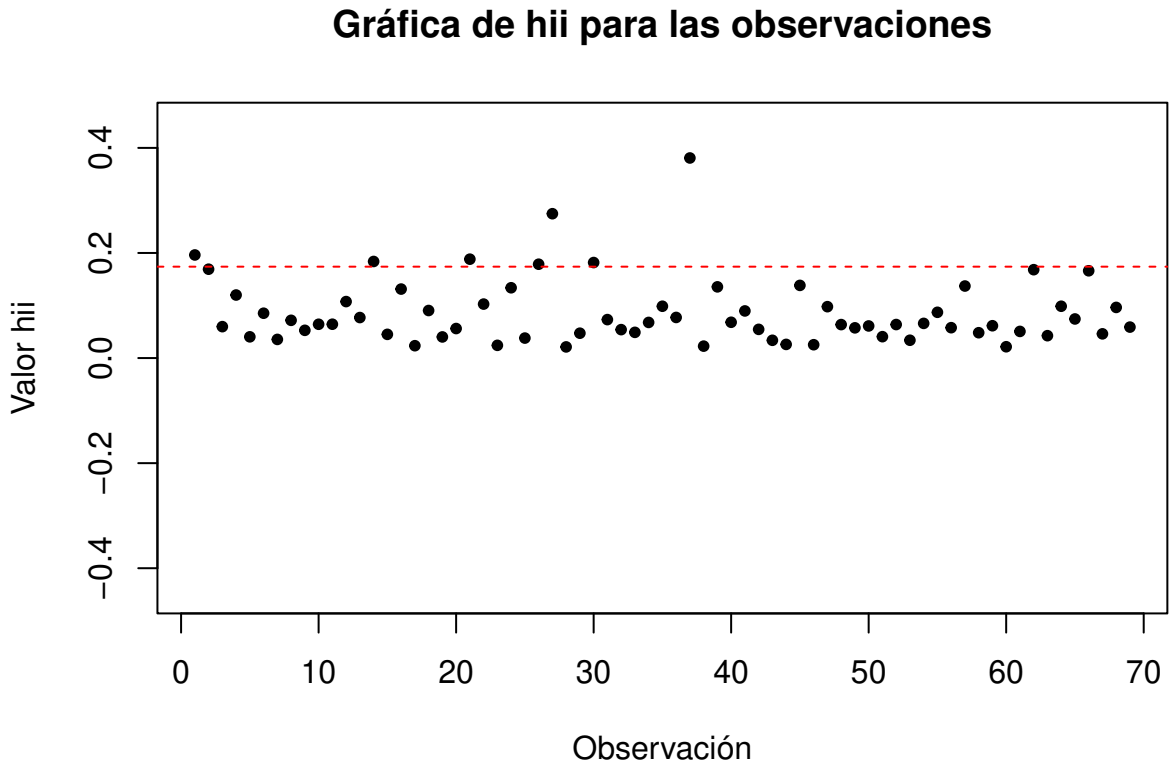


Figura 4: Identificación de puntos de balanceo

Cuadro 5: Puntos de balanceo

Observación	Valor hii
1	0.1962
14	0.1838
21	0.1882
26	0.1786
27	0.2747
30	0.1818
37	0.3808

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$, es decir $h_{ii} = 0.1846$, se reconocen 7 puntos de balanceo bajo el criterio que su respectivo $h_{ii} > 0.1846$, los cuales están presentados en la tabla

4.2.3. Puntos influenciales

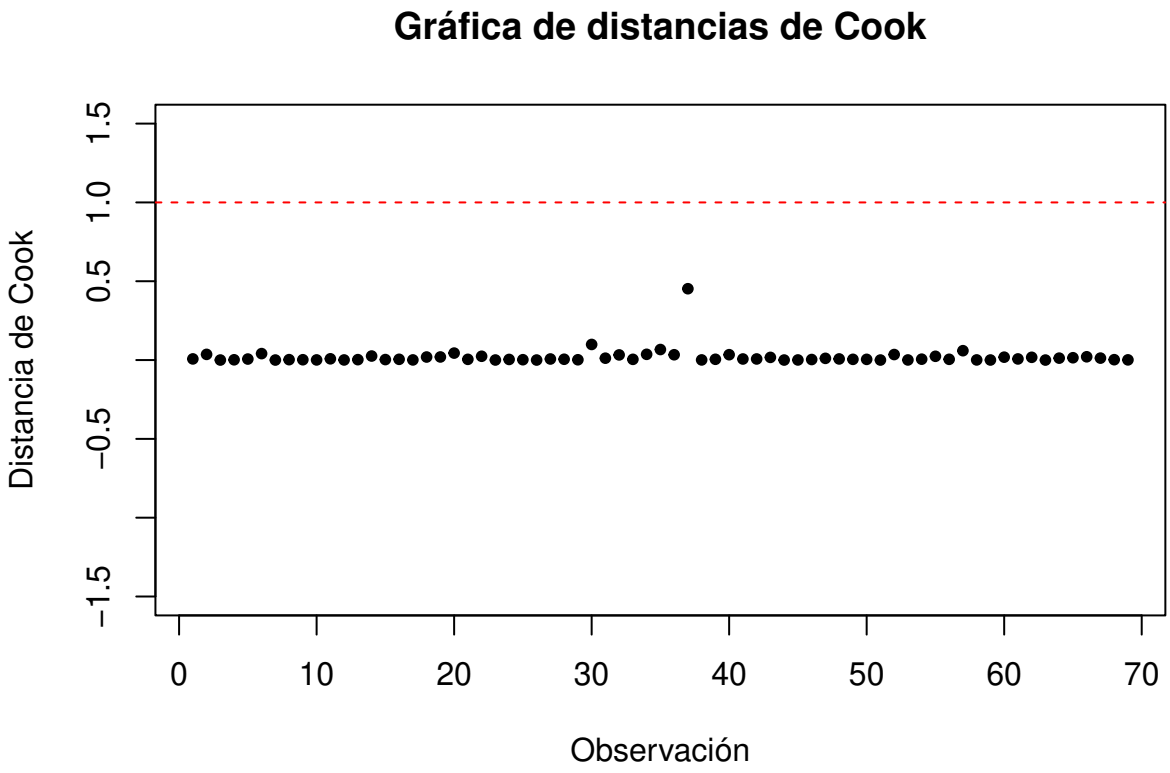


Figura 5: Criterio distancias de Cook para puntos influenciales

Como no hay puntos que superen el rango en el criterio de distancias de Cook, se concluye que no hay observaciones extremadamente influyentes en el modelo de regresión. Esto sugiere que el modelo es robusto y que las observaciones individuales no afectan significativamente los resultados. Sin embargo, sigue siendo importante considerar otros posibles problemas en el análisis de regresión para poder concluir de manera mas segura.

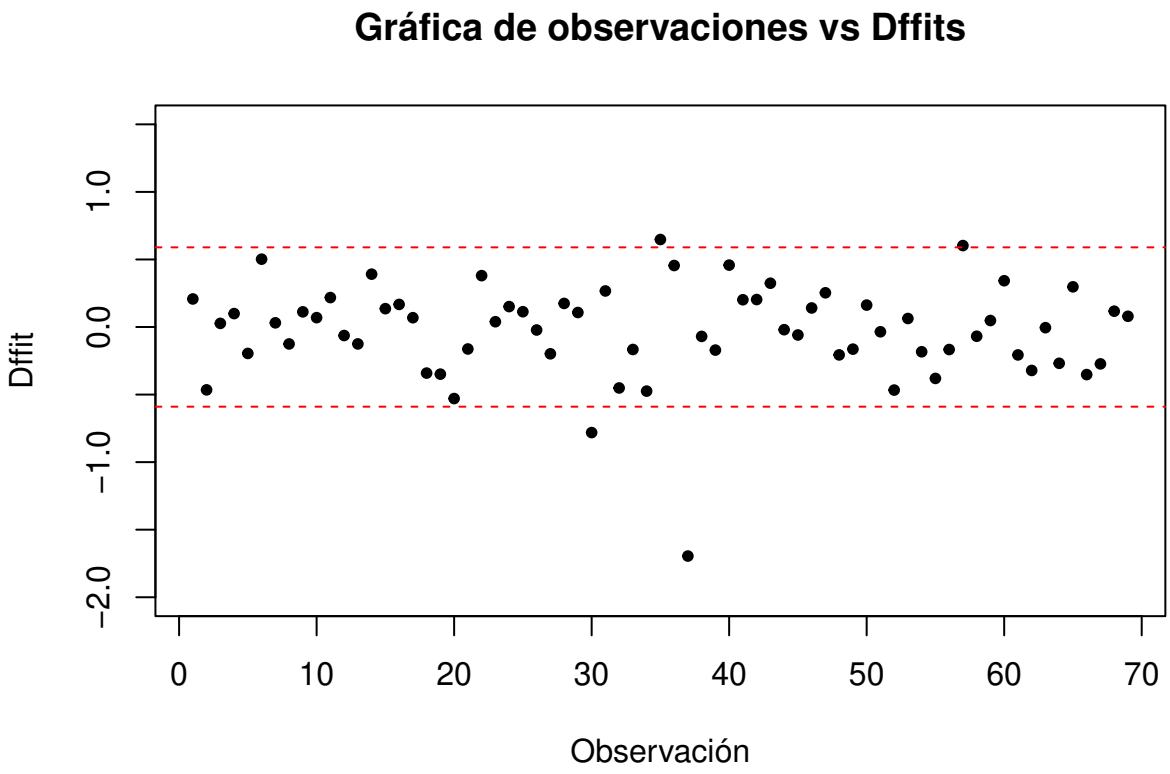


Figura 6: Criterio Dffits para puntos influenciales

Cuadro 6: Puntos influenciales

Observación	Cook (D_i)	Dffits
30	0.0990	-0.7814
35	0.0669	0.6476
37	0.4525	-1.6951
57	0.0594	0.6028

- De acuerdo al criterio de Dffits las observaciones {30, 35, 37, 57} son puntos influenciales, ya que estas observaciones cumplen con el siguiente criterio: $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$.
- Respecto a el criterio de las distancias de Cook en el cual toda observacion cuya $D_i > 1$, se considera un punto influenciable, podemos decir que en este casi ninguno de las observaciones cumple con este criterio

4.3. Conclusión

Aunque, el R^2 que obtuvimos en el modelo no es ni siquiera 0.5, siendo exactos $R^2 = 0.495787$ por lo que, de entrada sabemos que nuestras variables regresoras solo estan explicando alrededor del 49 % de la variabibilidad total del riesgo de infeccion en el problema estudiado. Ademas, aunque concluimos que el modelo tiene varianza constante y media cero, y aceptamos el supuesto de normalidad de los errores y por ende el modelo. Esta decisión se tomo teniendo en cuenta que no obtuvimos ninguna observacion atipica en nuestro conjunto

de datos, obtuvimos 7 puntos de balanceo(Estos puntos pueden cambiar un poco la pendiente “gradiente” de nuestro modelo) y 4 puntos influenciales, estos pueden afectar significativamente la estimacion de los coeficientes y resumen del modelo por consiguiente la estimacion en nuestra variable dependiente, en este caso el rieso de infeccion, Ádemas pueden alterar las pruebas de los supuestos.

Es recomendable estar en compañia de un experto en el area para decidir si descartar o no las observaciones que nos estan generando problemas al modelo y asi poder tratar de obtener el mejor modelo posible.

Válido o no?