3,75

Trabajo 1

Estudiantes

María Camila López de mesa Acevedo
Leydi Yolanda Torres Chamorro
Julián Orrego Martínez
Isabel Carmona Alzate

Equipo 27

Docente

Francisco Javier Rodriguez

Asignatura

Estadística II

Sede Medellín
30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

16,5 pt

Considerando la base de datos número 27, la cual contiene 5 variables predictoras identificadas como:

X: Riesgo de infección

X_1: Duración de estadía

X_2: Rutina de cultivos

X_3: Número de camas

X_4: Censo promedio diario

X_5: Número de enfermeras

Por lo tanto es pobile plantear el modelo de regresión lineal múltiple

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

1.1. Modelo de regresión

Al realizar ajustes al modelo, obtenemos los siguientes coeficientes

Cuadro 1: Tabla de valores de coeficientes

	Valor del parámetro o dato
β_0	-1.8972
β_1	0.1007
β_2	0.0629
β_3	0.0462
β_4	0.0076
β_5	0.0024

3 pt

Como resultado tenemos el modelo de regresión ajustado:

$$\hat{Y}_i = -1.8972 + 0.1007 X_{1i} + 0.0629 X_{2i} + 0.0462 X_{3i} + 0.0076 X_{4i} + 0.0024 X_{5i}$$

1.2. Significancia de la regresión

9 pt

Con el fin de evaluar la significancia estadística de la regresión, se formula el siguiente conjunto de hipótesis:

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

Por lo tanto, su estadístico de prueba es:

$$F_0 = \frac{MST}{MSE} \overset{H_0}{\sim} f_{5,44} \tag{1}$$

Continuamos con la respectiva tabla anova:

Cuadro 2: Tabla ANOVA

	SC	GL	CM	F_0	Valor-p
Regresión	44.4748	5	8.894962	9.4524	3.57209e-06
Error	41.4052	44	0.941027		

Con los datos suministrados en la tabla ANOVA, se observa un valor P aproximadamente igual a 0. Debido a que, el valor de P se aproxima a un número cercano a 0 se concluye que la hipótesis nula se rechaza en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$, de esta forma aceptando la hipótesis alternativa que cuando $\beta_j \neq 0$ indica que hay una relación lineal significativa entre la variable predictora y la variable de respuesta. Por lo tanto, si $\beta_j \neq 0$ la regresión también es significativa (Sin embargo, es importante tener en cuenta que otros factores también pueden influir en la significación de una regresión, como el tamaño de la muestra).

1.3. Significancia de los parámetros

El cuadro a continuación contiene información sobre los parámetros, lo que facilitará la identificación de los valores significativos.

Cuadro 3: Resumen coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	Valor-p
β_0	-1.8972	2.0356	-0.9320	0.3564
β_1	0.1007	0.0785	1.2828	0.2063
β_2	0.0629	0.0388	1.6202	0.1123
β_3	0.0462	0.0156	2.9678	0.0048
β_4	0.0076	0.0083	0.9157	0.3648
β_5	0.0024	0.0010	2.4840	0.0169

Los Valores-p que se ven en la tabla dan como conclusión, que con un nivel de significancia $\alpha = 0.05$, los parámetros β_3 y β_5 son significativos, pues su valor-p es igual a 0.0048 y 0.0169 los cuales son menores a α.

1.4. Interpretación de los parámetros

$\hat{\beta}_3$: El valor de 0.0628 indica que por cada unidad de cambio en la variable explicativa asociada a $\hat{\beta}_3$, se espera un aumento de 0.0628 unidades en la variable de respuesta, manteniendo todas las otras variables constantes. El valor del estadístico T_{0j} de 3.7108 sugiere que este coeficiente es significativamente diferente de cero, lo que sugiere que la variable explicativa asociada con $\hat{\beta}_3$ es un predictor importante de la variable de respuesta.

$\hat{\beta}_5$: El valor de 0.0022 indica que por cada unidad de cambio en la variable explicativa asociada a $\hat{\beta}_5$, se espera un aumento de 0.0022 unidades en la variable de respuesta, manteniendo todas las otras variables constantes. El valor del estadístico T_{0j} de 2.8909 sugiere que este coeficiente también es significativamente diferente de cero, lo que sugiere que la variable explicativa asociada con $\hat{\beta}_5$ es un predictor importante de la variable de respuesta. Sin embargo, el valor de p es un poco más alto que el de $\hat{\beta}_3$, lo que sugiere que hay una probabilidad ligeramente mayor de que este resultado sea debido al azar.

1.5. Coeficiente de determinación múltiple R^2

El modelo de regresión lineal utilizado tiene un valor de coeficiente de determinación múltiple $R^2 = 0.5179$. Lo que indica que alrededor del 51.79 % de la variabilidad total en la variable de respuesta se puede explicar mediante las variables predictoras incluidas en el modelo propuesto. Esto sugiere que el modelo es útil para para explicar la relación entre las variables y que es capaz de capturar una cantidad significativa de la variabilidad observada de los datos. Sin embargo, todavía queda un 48.21 % de la variabilidad que no se puede explicar con las variables predictoras incluidas en el modelo y es posible que se deba a otros factores no considerados en el análisis.

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 regresiones posibles

	SSE	Covariables en el modelo
Modelo completo	41.405	X1 X2 X3 X4 X5
Modelo reducido	50.201	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

El estadístico de prueba es construido de la siguiente manera:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,44} \\ &= \frac{50.201 - 41.405}{0.9410227} \\ &= 9.347277 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,44} = 2.8165$, se puede ver que $F_0 > f_{0.95,1,45}$

Con base en esto se rechaza la hipótesis nula lo que indica que las variables son significativas en el modelo y por ende no es posible descartar las variables del subconjunto.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

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

reescribiendo matricialmente:

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

Con $\mathbf{L}$ dada por

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

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

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - 41.405)/2}{0.9410227} \overset{H_0}{\sim} f_{2,44} \tag{3}$$

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

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

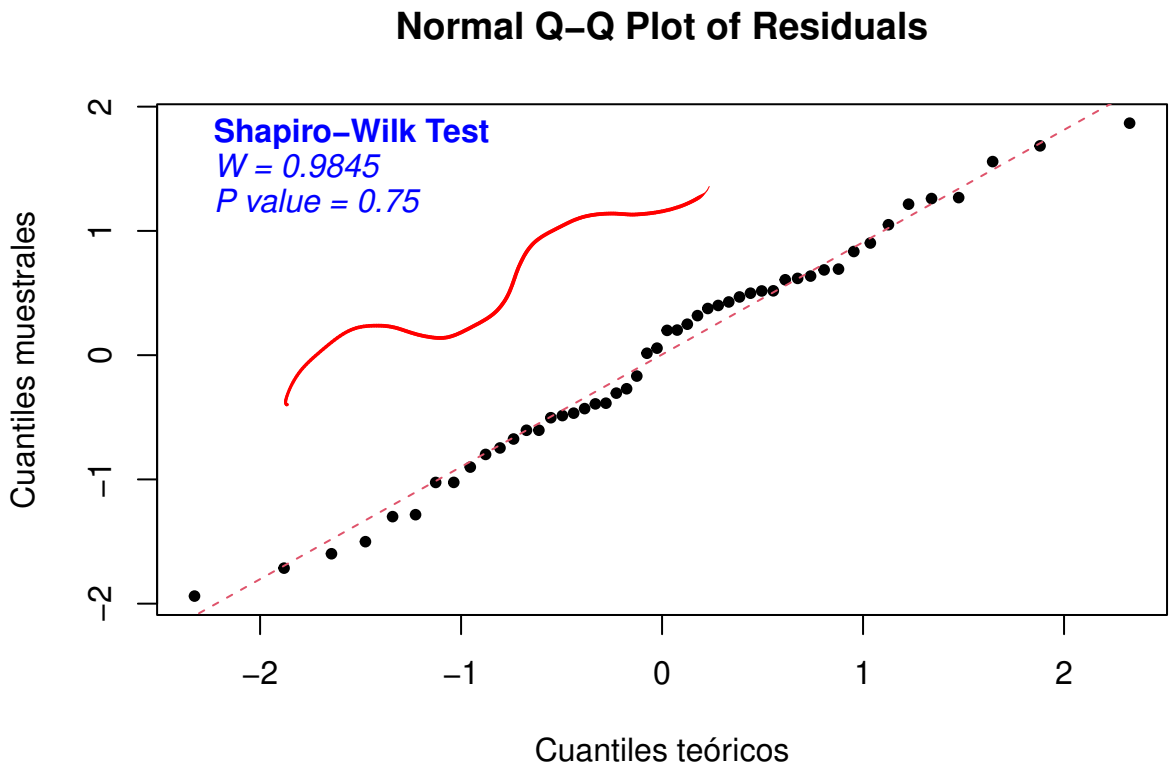

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

En la regresión lineal, el Valor-p se utiliza para determinar si la hipótesis nula de que no hay relación significativa entre las variables independientes y la variable dependiente puede ser rechazada o no. Si el Valor-p es menor que el nivel de significancia α (en este caso, 0.05), se

rechaza la hipótesis nula y se concluye que hay una relación significativa entre las variables. Por otro lado, si el Valor-p es mayor que α, se acepta la hipótesis nula y se concluye que no hay una relación significativa entre las variables.

En este caso, el Valor-p es aproximadamente igual a 0.75, lo que significa que es mucho mayor que el nivel de significancia α. Por lo tanto, no se puede rechazar la hipótesis nula de que los datos distribuyen normal con media μ y varianza σ^2. Sin embargo, la gráfica de comparación de cuantiles muestra colas más pesadas y patrones irregulares que sugieren que este supuesto puede no ser válido. Como la gráfica tiene más poder que la prueba estadística para detectar desviaciones de la normalidad, se decide rechazar la hipótesis nula y concluir que los datos no siguen una distribución normal.

Por último, se realizará una prueba para validar si la varianza cumple con el supuesto de ser constante en todos los niveles de la variable independiente. Si se encuentra evidencia en contra de este supuesto, se podría necesitar una transformación en los datos o una modificación en el modelo para corregir este problema.

4.1.2. Varianza constante

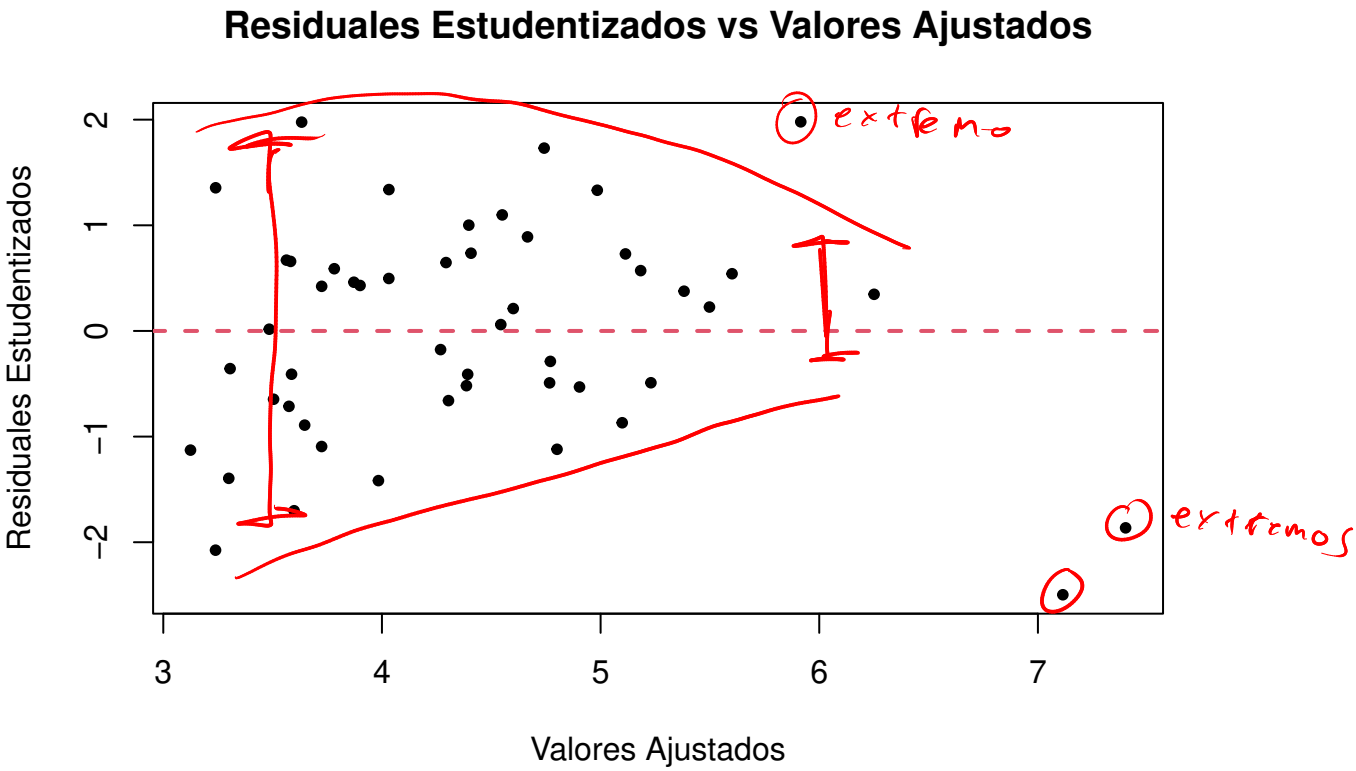

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que la varianza no es constante ya que hay una forma clara de embudo, lo cual indica que el modelo no está capturando adecuadamente la variación en los datos.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

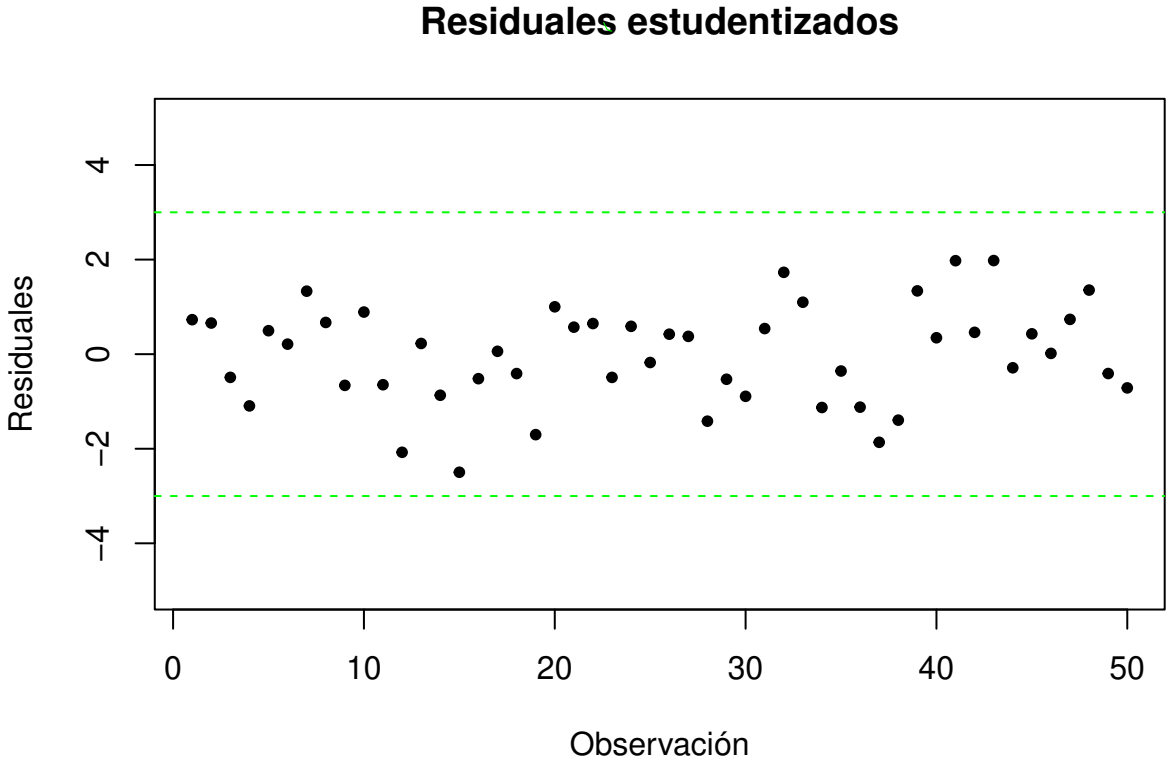

Figura 3: Identificación de datos atípicos

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

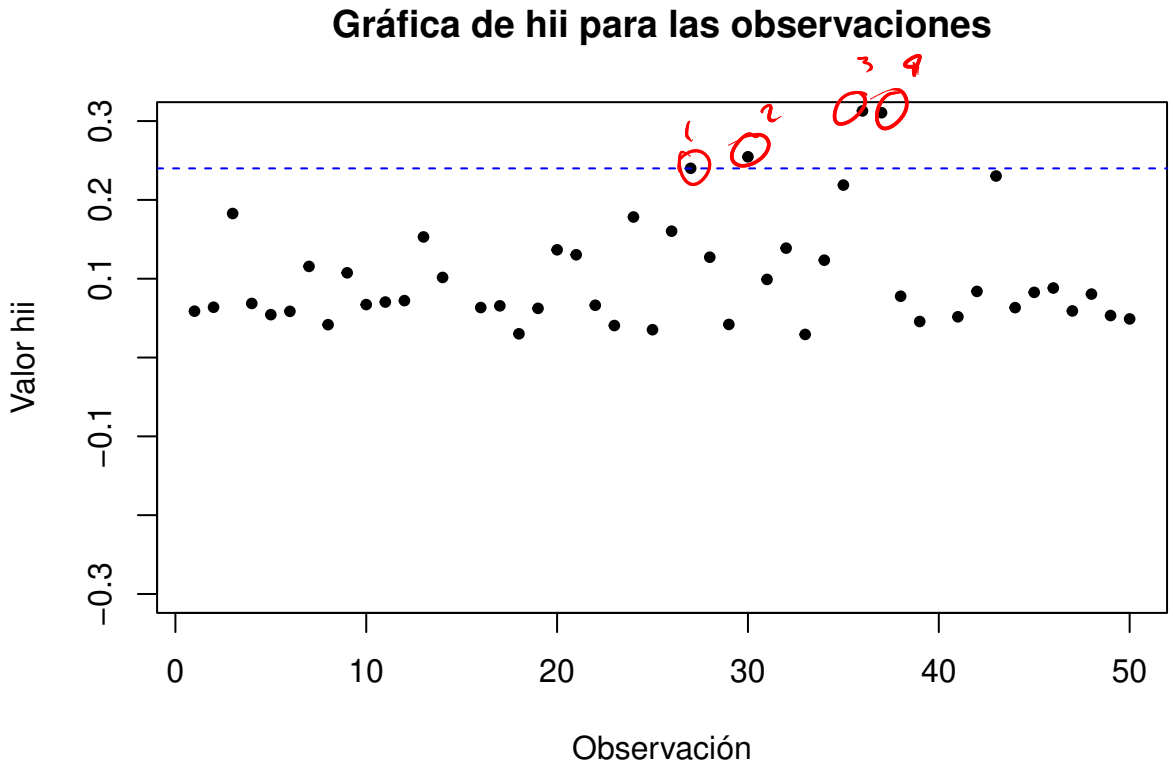

Figura 4: Identificación de puntos de balanceo

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 0.24$, se puede apreciar que existen 3 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 como los datos #15, #32 y #37 de la columna hii.value. Esto indica que las observaciones tienen valores hii muy altos y por lo tanto tienen una gran influencia en el ajuste del modelo, lo que sugiere que deban ser evaluados cuidadosamente para determinar si deben ser excluidos del analisis o si se deben considerar modelos alternativos.

	hii.value
15	0.4993
27	0.2401
30	0.2547
36	0.3129
37	0.3106
40	0.4525

4.2.3. Puntos influenciales

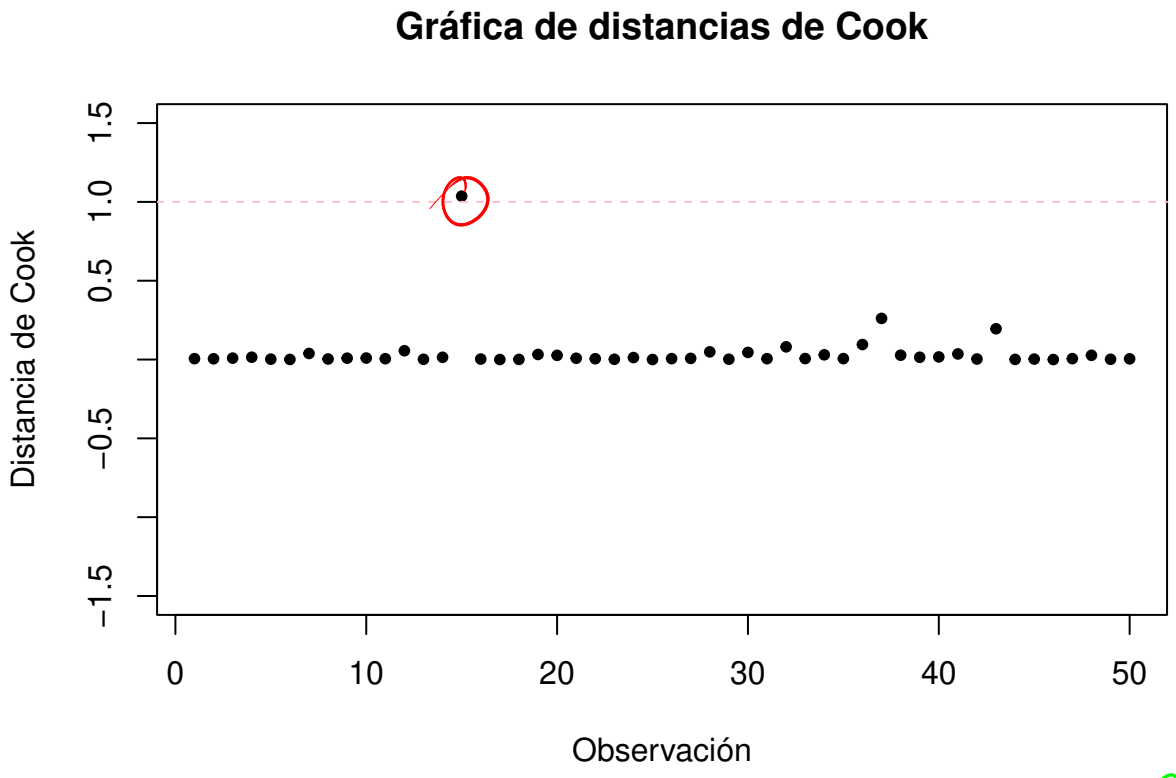

Figura 5: Criterio distancias de Cook para puntos influenciales

Podemos observar que hay un valor alto del estadistico de Cook, lo que nos indica que este punto tiene gran influencia en los parámetros del modelo y por lo tanto en las predicciones basadas en el modelo. Esto sugiere que el punto #15 debe ser evaluado cuidadosamente y considerado para ser excluido del modelo si tiene un impacto significativo en las estimaciones de los parámetros.

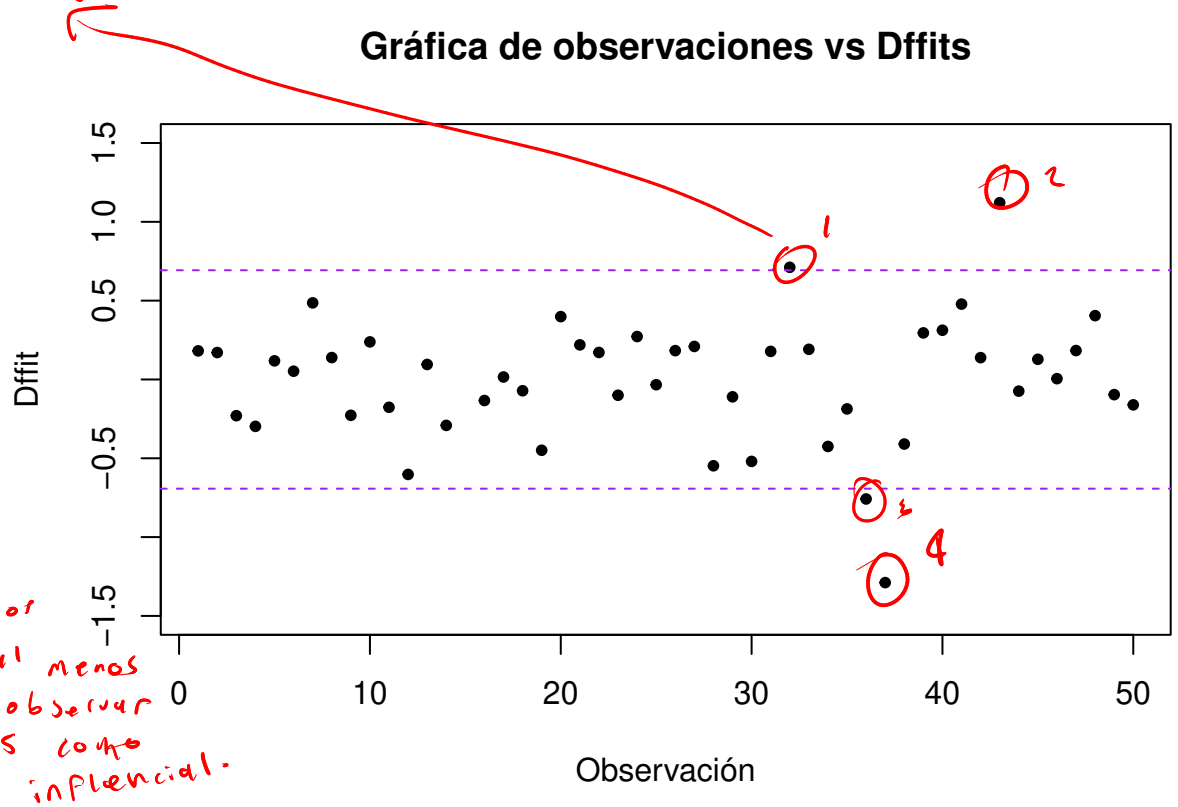

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 15  -2.4971  1.0363    0.4993 -2.6608
## 32   1.7307  0.0805    0.1388  0.7116
## 36  -1.1203  0.0952    0.3129 -0.7582
## 37  -1.8638  0.2609    0.3106 -1.2888
## 43   1.9793  0.1953    0.2303  1.1213
```

Como se puede ver,las observaciones de los puntos #15, #37 y #43 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. También para cualquier punto cuya $D_i > 1$, es un punto influencial. Las tres observaciones pueden estar afectando significativamente las estimaciones de los parámetros y las predicciones basadas en el modelo.

4.3. Conclusión

Según los resultados y las observaciones, el modelo de regresión no cumple con los supuestos de normalidad, la varianza no es constante, además presenta valores atípicos. Esto nos dice que los residuos del modelo no se distribuyen normalmente y las inferencias basadas en el modelo como los intervalos de confianza y las pruebas de hipótesis son incorrectas, además que la varianza de los residuos no es constante en todo el rango de los valores ajustados. Finalmente los valores atípicos tienen una gran influencia en el ajuste del modelo, lo cual afecta las inferencias basadas en el modelo.

Se puede analizar que de las variables predictoras X_1,X_2 y X_3 no nos permiten extraer datos relevantes para el estudio del modelo, las cuales no inciden en el riesgo de infección en el hospital.

No dicen si es válido o no