Trabajo 1

3,7

Estudiantes

Camilo Andres Granada Mejia
Jose Manuel Carmona Estrada
Jhon Stiven Cifuentes Gomes
David Gil Rua

Equipo 51

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

16 pt

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

Donde:

- 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 obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-0.9267
β_1	0.2175
β_2	0.0144
β_3	0.0459
β_4	0.0134
β_5	0.0017

3 pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.9267 + 0.2175 X_{1i} + 0.0144 X_{2i} + 0.0459 X_{3i} + 0.0134 X_{4i} + 0.0017 X_{5i}\ 1 \leqslant i \leqslant 64$$

1.2. Significancia de la regresión

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

$$\begin{cases} H_0 : \cancel{\beta_0} = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1 : \text{Algún } \beta_j \text{ distinto de 0 para j=0, 1,..., 5} \end{cases}$$

Cuyo estadístico de prueba es:

2pt

$$F_0 = \frac{\overset{MSR}{\cancel{MST}}}{MSE} \overset{H_0}{\sim} f_{5,58} \tag{1}$$

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	78.8979	5	15.779588	17.4904	1.58805e-10
Error	52.3269	58	0.902188		

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 $\cancel{0} \leqslant j \leqslant 5$, 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

6pt

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-0.9267	1.4465	-0.6406	0.5243
β_1	0.2175	0.0754	2.8842	0.0055
β_2	0.0144	0.0268	0.5368	0.5935
β_3	0.0459	0.0121	3.7828	0.0004
β_4	0.0134	0.0068	1.9604	0.0548
β_5	0.0017	0.0006	2.6271	0.0110

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.

1.4. Interpretación de los parámetros

3pt

$\hat{\beta}_1$: El riesgo de infeccion aumenta significativamente en 0.2175 por cada dia de de estadia cuando las demas variables predictoras se mantienen fijas

$\hat{\beta}_3$: El riesgo de infeccion aumenta significativamente en 0.0459 por cada cama en el hospital durante el periodo de estudio cuando las demas variables predictoras se mantienen fijas

$\hat{\beta}_5$: El riesgo de infeccion aumenta significativamente en 0.0017 en relacion al numero promedio de enfermeras presentes equivalentes a tiempo completo, durante el periodo de estadia cuando las demas variables predictoras se mantienen fijas

1.5. Coeficiente de determinación múltiple R^2

2pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.6012423$, lo que significa que aproximadamente el 60.12423 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

¿cómo se calcula?

2. Pregunta 2

4pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más bajo 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	52.327	X1 X2 X3 X4 X5
Modelo reducido	96.374	X2 X4

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

$$Y_i = \beta_0 + \beta_2 X_{2i} + \beta_4 X_{4i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

2pt

$$
\begin{aligned}
F_0 &= \frac{(SSE(\beta_0,\beta_2,\beta_4) - SSE(\beta_0,\ \cdots,\ \beta_5))/3}{MSE(\beta_0,\ \cdots,\ \beta_5)} \overset{H_0}{\sim} f_{3,58} \\
&= \cancel{\frac{96.374 - 52.327}{0.902188}} \\
&= \cancel{48.82242}
\end{aligned}
\tag{2}
$$

Ahora, comparando el F_0 con $f_{0.95,3,58} = 2.7636$, se puede ver que $F_0 > f_{0.95,1,45}$, por tanto se rechaza la hipotesis nula, teniendo que al ser esto asi, no es posible descartar las variables del conjunto 2pt

3. Pregunta 3 4pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si la duracion de la estadia por dia es 2 veces el numero promedio de enfermeras en tiempo completo durante el periodo del estudio y si el numero de camas promedio durante el periodo del estudio es 3 veces el numero promedio de pacientes por dia durante el periodo del estudio. Por consiguiente se plantea la siguiente prueba de hipótesis:

$$
\begin{cases}
\mathrm{H}_0 : \beta_1 = 2\beta_5;\ \ \beta_3 = 3\beta_4 \\
\mathrm{H}_1 : \text{Alguna de las igualdades no se cumple}
\end{cases}
$$

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

$$
L = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & -2 \\ 0 & 0 & 0 & 1 & -3 & 0 \end{bmatrix}
$$

2pt

El modelo reducido está dado por:

$$
Y_i = \beta_o + \beta_1 X_{1,5i} + \beta_2 X_{2i} + \beta_3 X_{3,4i} + \varepsilon_i,\ \ \varepsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2);\ \ 1 \leqslant i \leqslant 64
$$

0pt

Donde $X_{1,5i} = X_{1i} + 2X_{5i}$ y $X_{3,4i} = X_{3i} + 3X_{4i}$

$X_{1,5} = 2X_{1i} + X_{5i} \ \wedge \ X_{3,4} = 3X_{3i} + X_{4i}$

3.2. Estadístico de prueba

2 pt

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,58} = F_0 = \frac{(SSE(MR) - 52.327)/2}{0.902188} \overset{H_0}{\sim} f_{2,58} \quad (3)$$

4. Pregunta 4

13 pt

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:

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

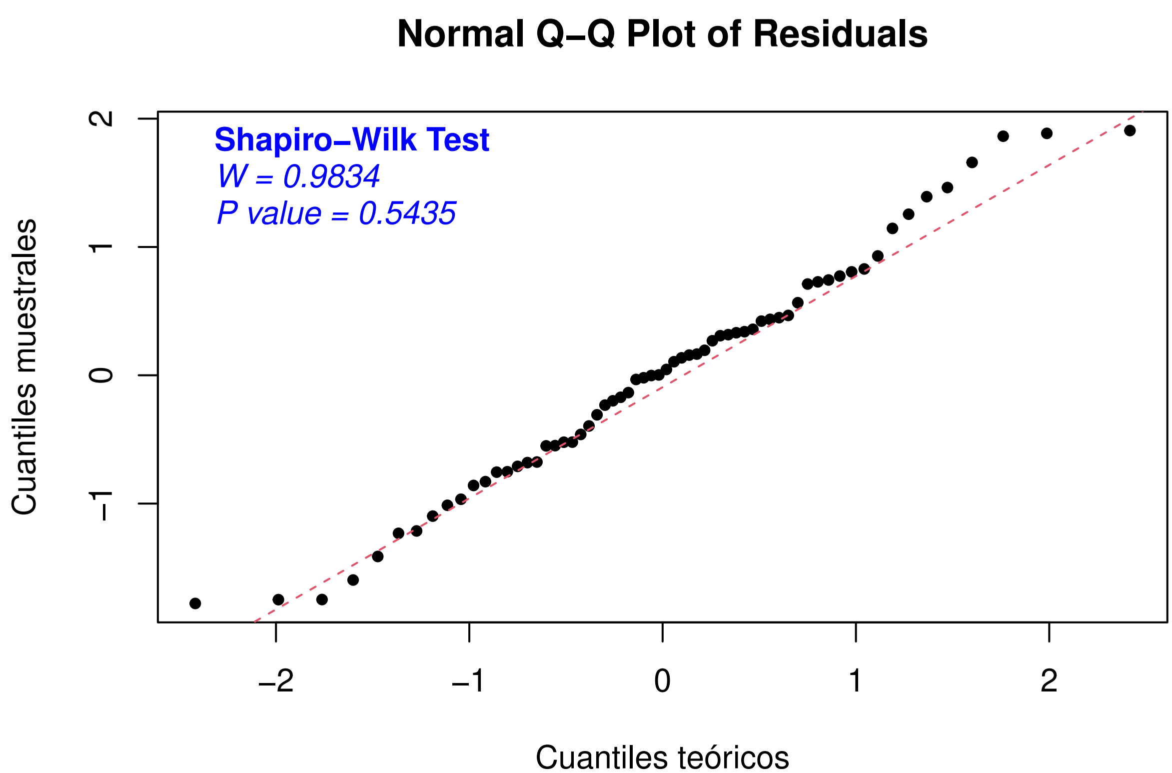
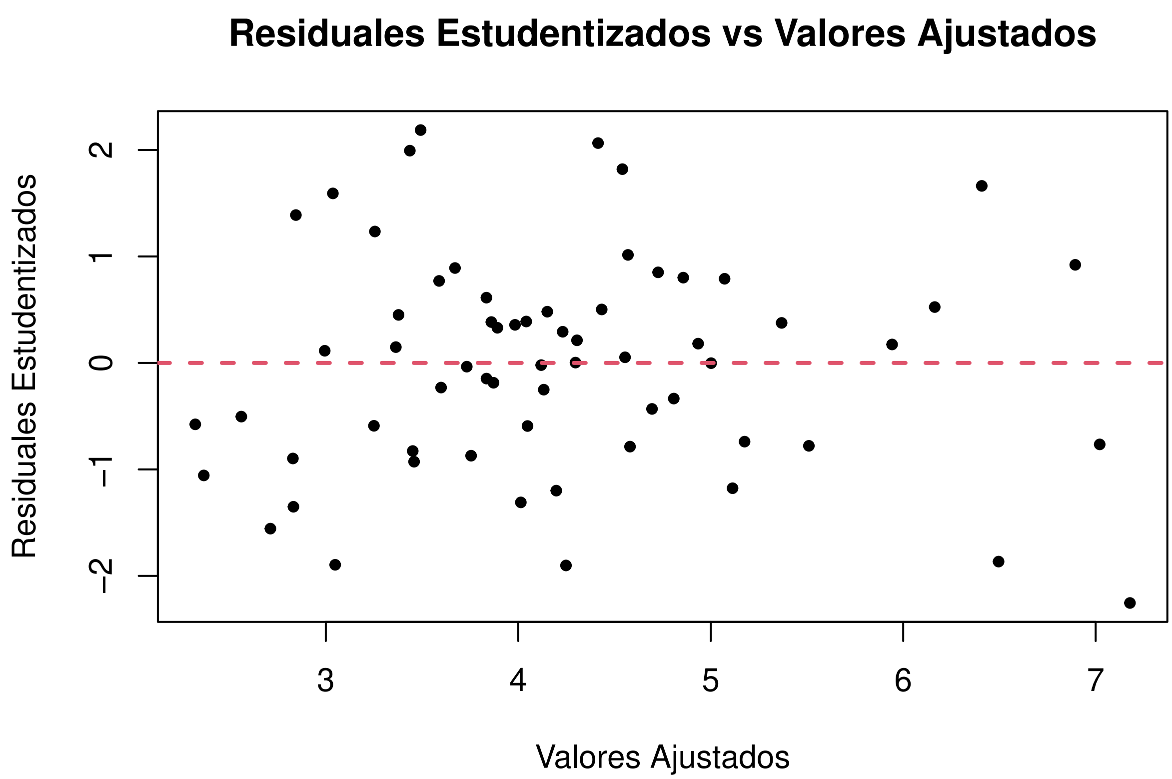


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

4 pt

Al ser el P-valor aproximadamente igual a 0.5435 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 0 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



2pt

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Sí la hay

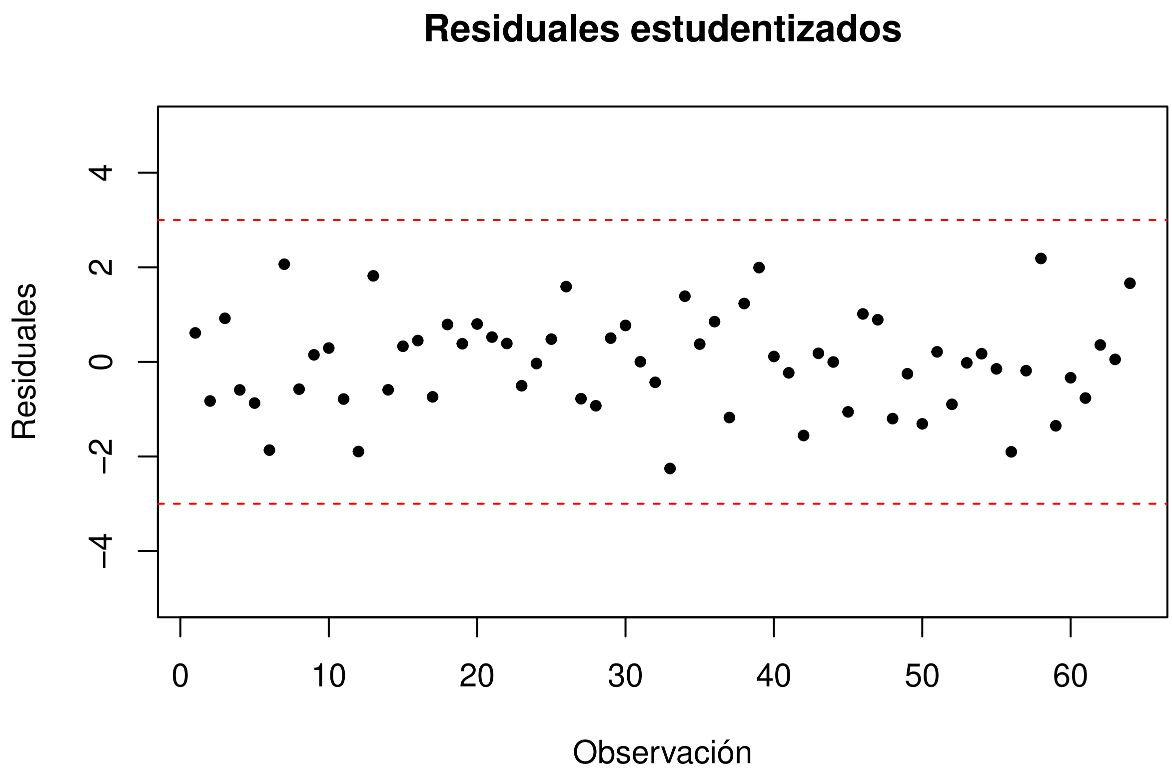
En el gráfico de residuales estudentizados vs valores ajustados se puede observar que no hay patrones en los que la varianza aumente, decrezca ni un comportamiento que permita descartar una varianza constante, al no haber evidencia suficiente en contra de este supuesto se acepta como cierto. Además es posible observar media 0. → se ve sólo con los e_i

4.2. Verificación de las observaciones

−1pt

Tengan cuidado acá, modifiquen los límites de las gráficas para que tenga sentido con lo que observan en la tabla diagnóstica. También, consideren que en aquellos puntos extremos que identifiquen deben explicar el qué causan los mismos en el modelo.

4.2.1. Datos atípicos



3pt

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

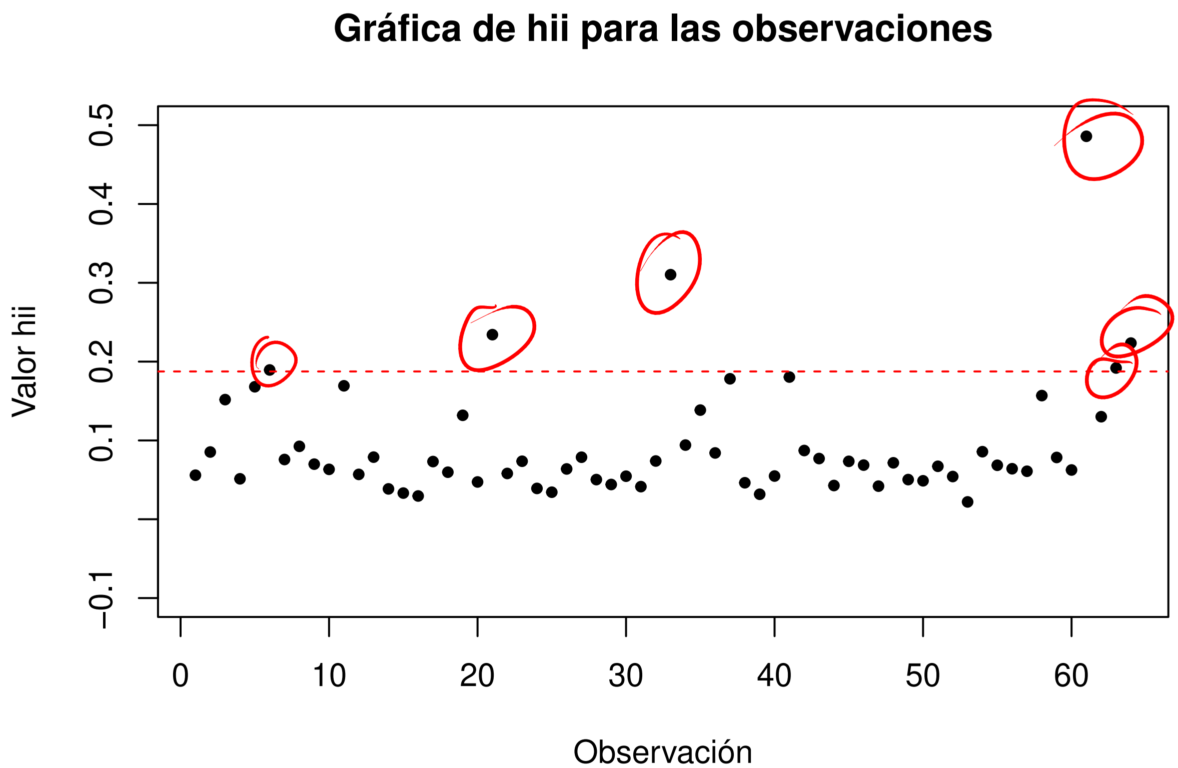


Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 6   -1.8658  0.1355    0.1894 -0.9221
## 21   0.5247  0.0140    0.2342  0.2883
## 33  -2.2544  0.3811    0.3103 -1.5694
## 61  -0.7651  0.0922    0.4859 -0.7411
## 63   0.0530  0.0001    0.1919  0.0256
## 64   1.6629  0.1326    0.2235  0.9062
```

2pt

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 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.

Causan...?

4.2.3. Puntos influenciales

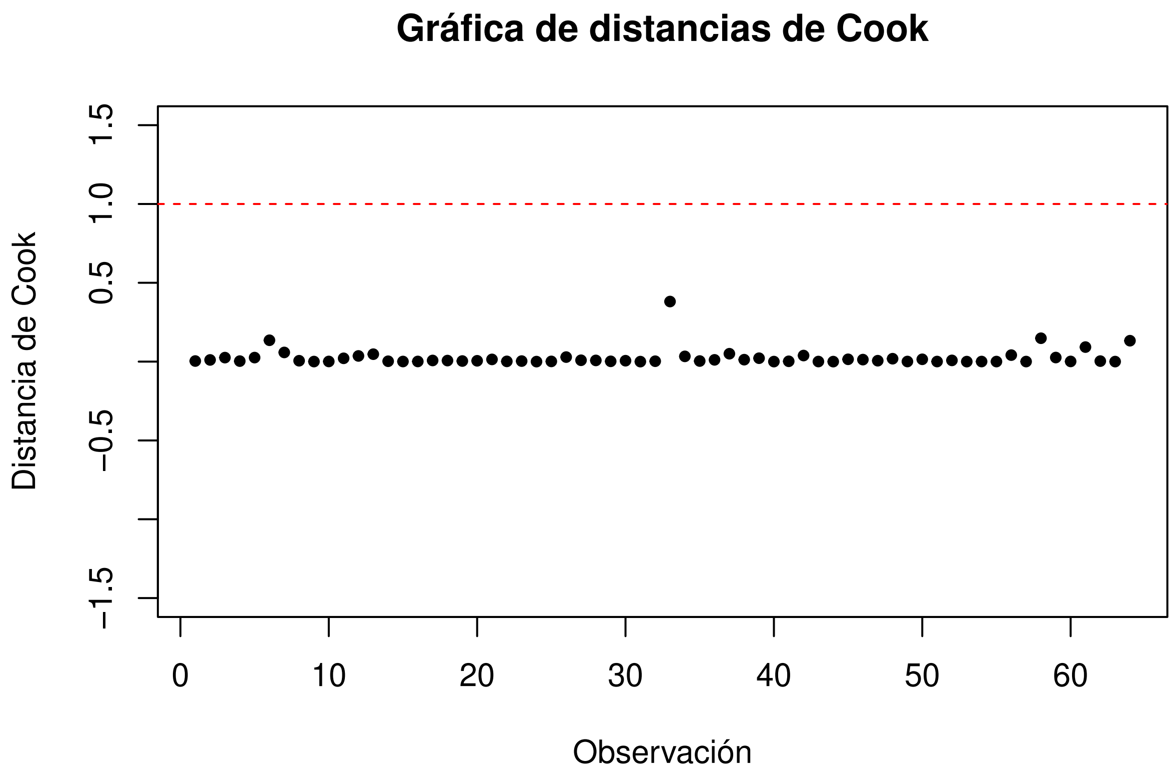


Figura 5: Criterio distancias de Cook para puntos influenciales

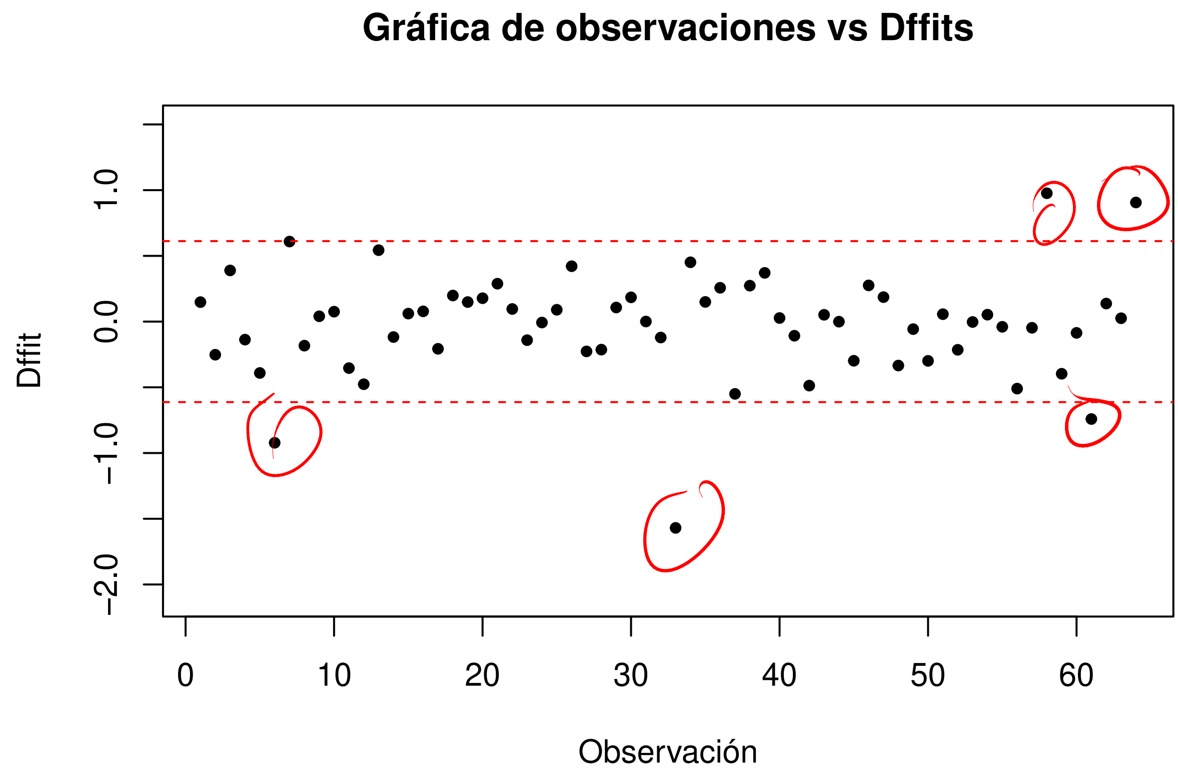


Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 6   -1.8658  0.1355    0.1894 -0.9221
## 33  -2.2544  0.3811    0.3103 -1.5694
## 58   2.1869  0.1484    0.1569  0.9765
## 61  -0.7651  0.0922    0.4859 -0.7411
## 64   1.6629  0.1326    0.2235  0.9062
```

Causan...?

5pt

Como se puede ver,las observaciones 6, 33, 58, 61, y 64 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

0pt

Dado que el P-Valor obtenido en la tabla ANOVA es de 1.58805e-10, podemos afirmar que el modelo funciona para cualquier alfa. → Esa conclusión qué...?

evaluando las observaciones de influencia con los diagnosdicadores Cook y DFFITS concluimos que ninguna de las observaciones son significativas.

→ ¿Qué dice que un modelo funcione? válido o no?