4,2

Trabajo 1

Estudiantes

Dioselin Esteban Brito Peñaloza
Juan David Cortés Amador
Sofía Hincapié Ibargüen
Alejandro Noriega Soto

Equipo 06

Docente

Julieth Veronica Guarin Escudero

Asignatura

Estadística II

Sede Medellín

5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

18 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

3 pt

Al ajustar el modelo, se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-0.1287
β_1	0.1790
β_2	0.0135
β_3	0.0609
β_4	0.0093
β_5	0.0011

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.1287 + 0.179 X_{1i} + 0.0135 X_{2i} + 0.0609 X_{3i} + 0.0093 X_{4i} + 0.0011 X_{5i}$$

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:

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

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	72.0119	5	14.40237	14.3156	4.02049e-09
Error	58.3517	58	1.00606		

De la tabla Anova, se observa un valor P aproximadamente igual a 0, teniendo en cuenta un $\alpha = 0.05$ 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. 4 pt

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	-0.1287	1.5338	-0.0839	0.9334
β_1	0.1790	0.0750	2.3862	0.0203
β_2	0.0135	0.0289	0.4663	0.6427
β_3	0.0609	0.0157	3.8893	0.0003
β_4	0.0093	0.0072	1.2966	0.1999
β_5	0.0011	0.0007	1.5790	0.1198

6 pt

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_1 y β_3 son significativos, pues sus P-valores son menores a α.

1.4. Interpretación de los parámetros

3 pt

Se interpretaran las siguientes variables que son significativas:

$\hat{\beta}_1$**:** En promedio, por cada unidad de aumento en la duración de la estadía, el riesgo de infección aumenta en 0.1790 unidades, cuando las demás variables se mantienen fijas.

$\hat{\beta}_3$**:** En promedio, por cada unidad de aumento en el número de camas, el riesgo de infección aumenta en 0.0609 unidades, cuando las demás variables se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2

2 pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.552$, lo que significa que aproximadamente el 55.2 % de la variabilidad total observada en el riesgo de infección es explicada por el modelo de regresión propuesto en el presente informe.

¿Cómo se calcula?

2. Pregunta 2

1 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

son 3!! → ¿?

Las covariable con el P-valor más pequeñas en el modelo fueron $X_1, X_3,$ por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

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

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	58.352	X1 X2 X3 X4 X5
Modelo reducido	62.364	X1 X3

0 pt

X2 X4 X5 → y eso que siendo congruente con el error

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

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

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

→ β_0?

$$
\begin{aligned}
F_0 &= \frac{(SSE(\beta_1, \beta_3) - SSE(\beta_0,\ \cdots,\ \beta_5))/2}{MSE(\beta_0,\ \cdots,\ \beta_5)} \overset{H_0}{\sim} f_{2,58} \\
&= \frac{2.006}{1.006} \\
&= 1.994
\end{aligned} \tag{2}
$$

Ahora, comparando el F_0 con $f_{0.05,2,58} = 3.1559$, se puede ver que $F_0 < f_{0.05,2,58}$ y por tanto no se rechaza H_0, y se concluye que el conjunto de predictoras individualmente no significativas en presencia de los demás parámetros, se pueden descartar del modelo.

1 pt

3. Pregunta 3

4,5 pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se plantea la siguiente pregunta, ¿Existe una relación entre la duración de la estadía y el número de camas, y que el censo promedio diario es 2 veces al número de enfermeras? por consiguiente se plantea la siguiente prueba de hipótesis:

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

reescribiendo matricialmente:

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

$$
\begin{aligned}
\beta_1 - \beta_3 &= 0 \\
2\beta_4 - \beta_5 &= 0
\end{aligned}
$$

1,5 pt

Con **L** dada por

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

El modelo reducido está dado por:

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

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

1 pt

3.2. Estadístico de prueba

2pt

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - 58.3517)/2}{1.00606} \overset{H_0}{\sim} f_{2,58} \quad (3)$$

4. Pregunta 4

18pt

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:

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

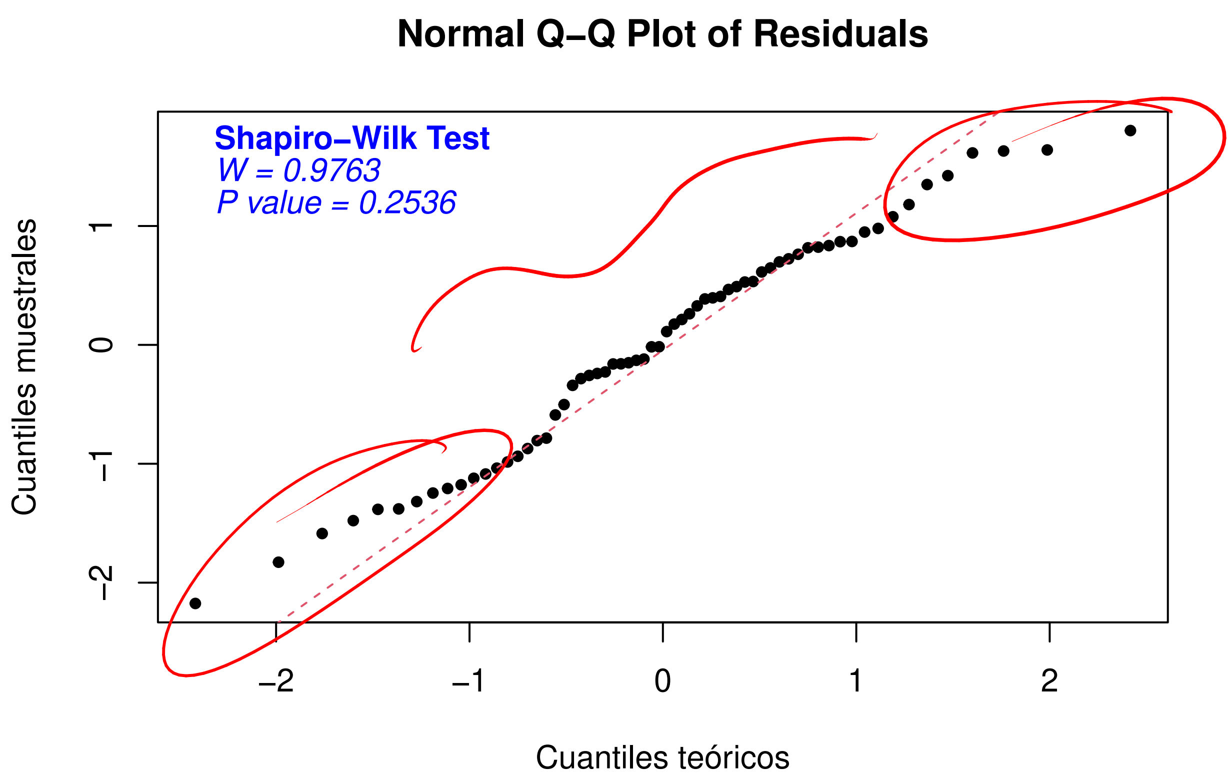

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

4pt

Al ser el P-valor aproximadamente igual a 0.2536 y teniendo en cuenta que el nivel de significancia $\alpha = 0.05$, el P-valor es mucho mayor y por lo tanto, no hay evidencia suficiente para rechazar la hipótesis nula, es decir que los datos distribuyen normal con media μ, 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

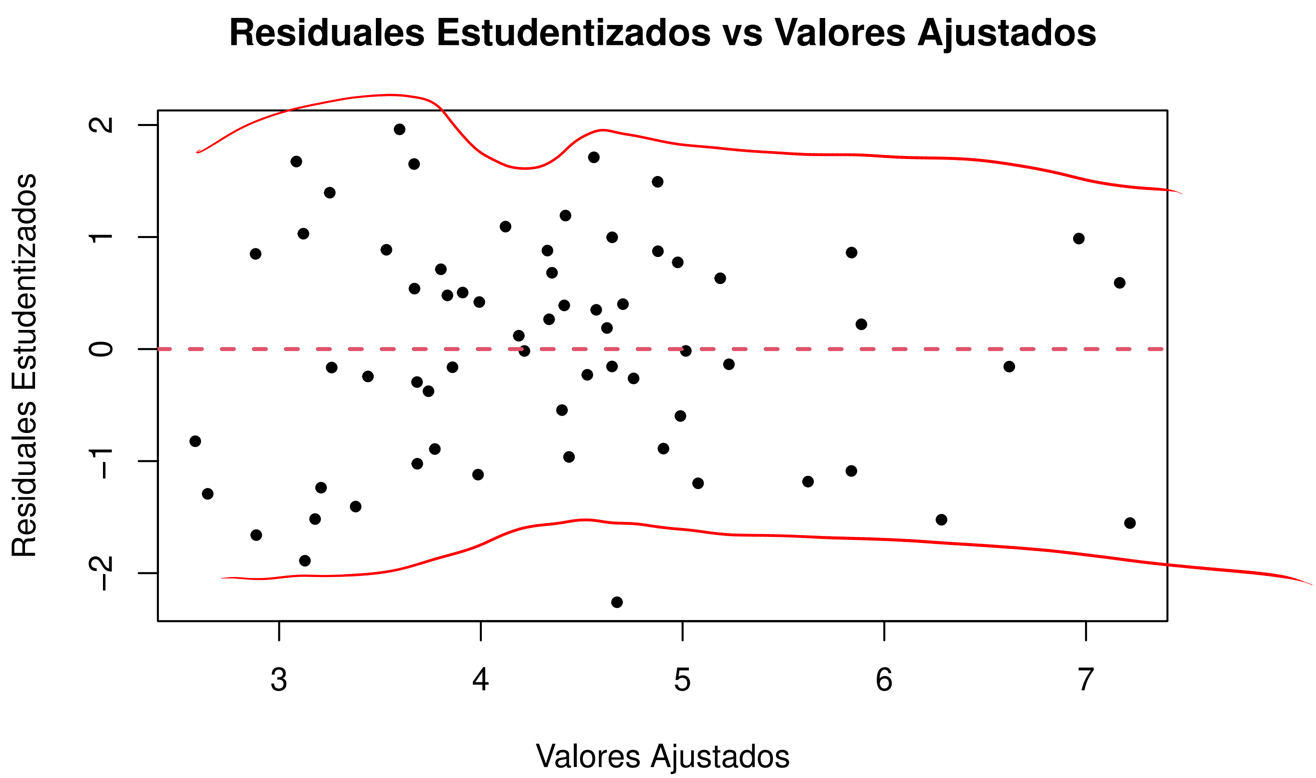

Figura 2: Gráfico residuales estudentizados vs valores ajustados

3 pt

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que hay una varianza constante, porque muchos de los puntos están ubicados entre -2 y 2, lo cual nos da una idea de que puede haber media cercana a 0, también se observa que son muy pocos los puntos que se pueden encontrar en los extremos. Finalmente no se observan patrones de varianza no constante como el embudo, por lo tanto al no tener evidencia suficiente para ir contra el supuesto de que es varianza constante, se acepta.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

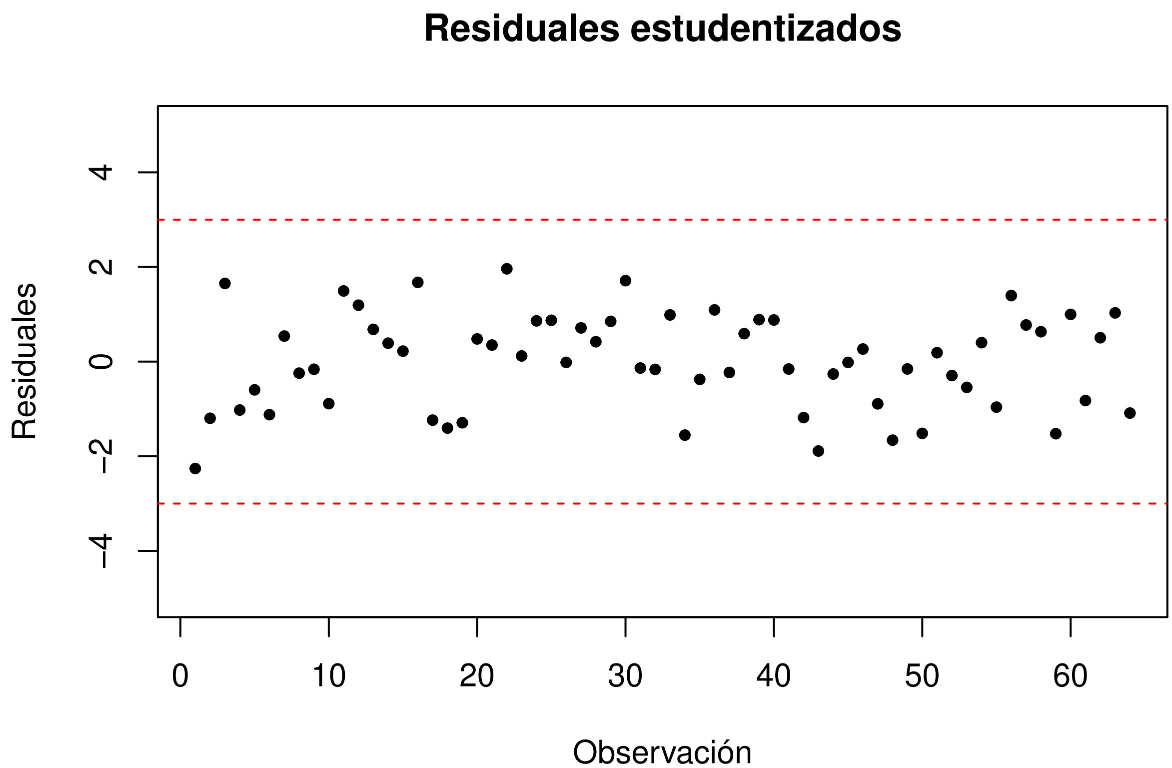

Figura 3: Identificación de datos atípicos

3 pt

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 $|r_{estud}| > 3$, se puede notar en la gráfica (Línea roja) que por encima de 3 y por debajo de -3, no hay puntos que estén afuera de esos limites.

4.2.2. Puntos de balanceo

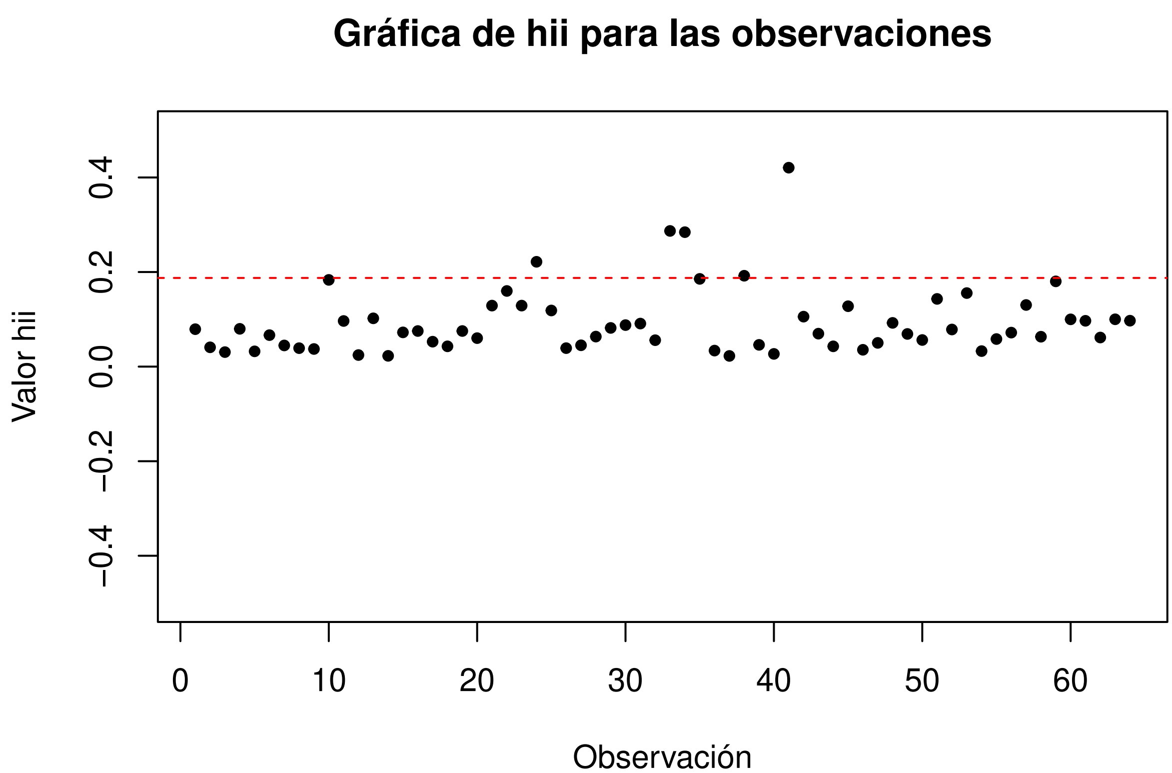

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 24   0.8612  0.0352    0.2217  0.4586
## 33   0.9864  0.0653    0.2869  0.6256
## 34  -1.5533  0.1597    0.2843 -0.9913
## 38   0.5912  0.0139    0.1922  0.2868
## 41  -0.1566  0.0030    0.4207 -0.1324
```

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 5 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.

2pt

¿Qué causan?

4.2.3. Puntos influenciales

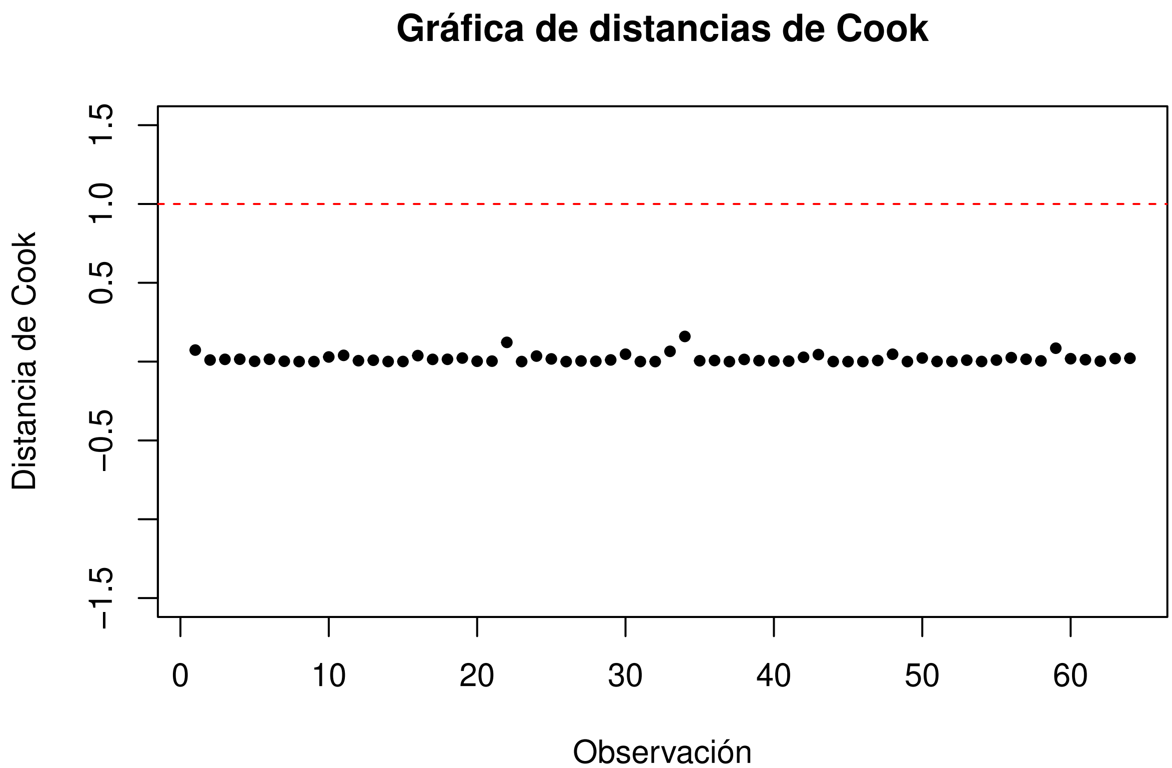

Figura 5: Criterio distancias de Cook para puntos influenciales

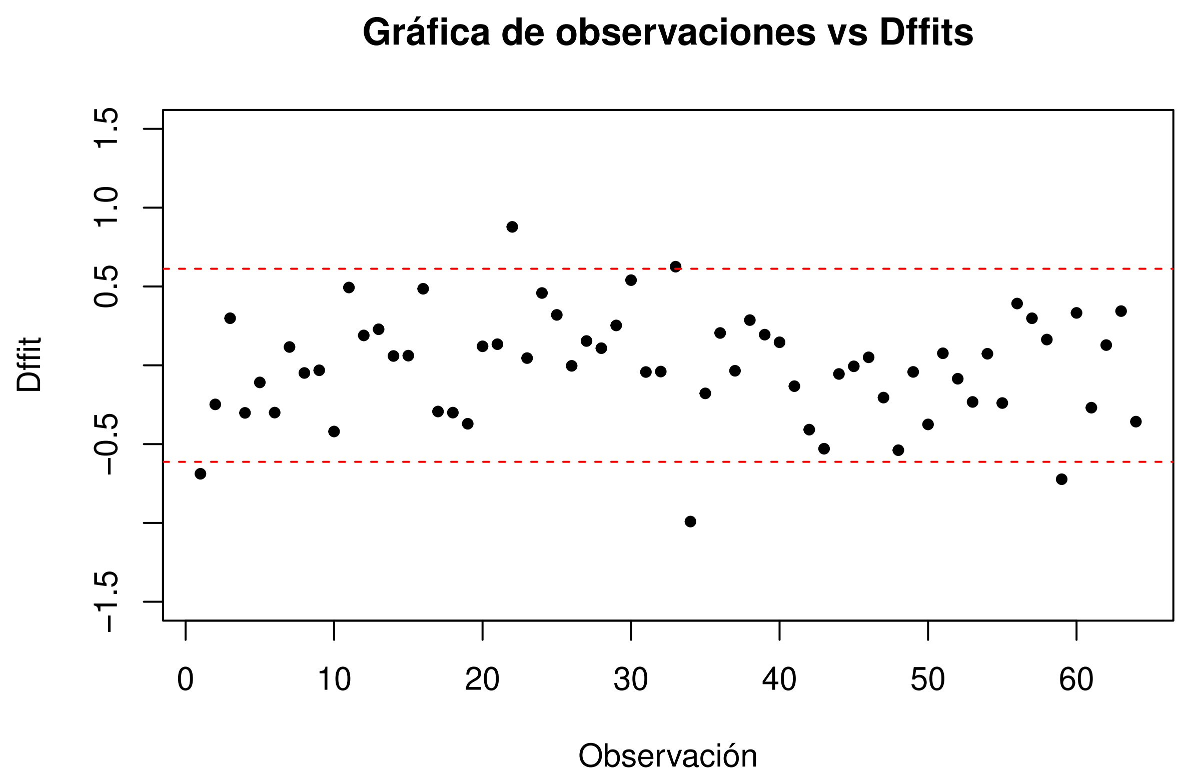

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 1   -2.2598  0.0733    0.0792 -0.6882
## 22   1.9611  0.1221    0.1600  0.8782
## 33   0.9864  0.0653    0.2869  0.6256
## 34  -1.5533  0.1597    0.2843 -0.9913
## 59  -1.5235  0.0851    0.1803 -0.7229
```

3pt

Como se puede ver,las observaciones 1, 22, 33, 34 y 59 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.

¿Qué causan?

4.3. Conclusión

En resumen tenemos para el analisis de las observaciones extremas:

- No se encontraron datos atípicos en el conjunto de datos, ya que ningún residual estudentizado sobrepasa el criterio de |r estud | > 3

3 pt

- Se tienen 5 puntos de balanceo: 24, 33, 34, 38 y 41.
- Se tienen 5 puntos influenciales: 1, 22,33, 34 y 59.
- En conclusión se pueden identificar valores extremos que afectan el modelo para la estimacion y/o prediccion de valores de respuesta.
- Finalmente, el modelo de regresión no es válido porque no cumple con uno de los supuestos que en este caso es que los residuales no se distribuyen de manera normal.