Trabajo 1

4,6

Estudiante

Andrés Herrera Correa

Equipo 21

Docente

Julieth Verónica Guarín

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:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 59$$

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	0.8207
β_1	0.2146
β_2	0.0024
β_3	0.0416
β_4	0.0082
β_5	0.0010

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 0.8207 + 0.2146 X_{1i} + 0.0024 X_{2i} + 0.0416 X_{3i} + 0.0082 X_{4i} + 0.001 X_{5i};\ 1 \leqslant i \leqslant 59$$

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:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,53} \tag{1}$$

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	45.4197	5	9.08395	9.03019	2.95776e-06
Error	53.3155	53	1.00595		

De la tabla Anova, se observa un valor P aproximadamente igual a 2.95776e-06, por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $0 \leqslant j \leqslant 5$ a una significancia dada $\alpha = 0.05$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$, por lo tanto la regresión es significativa.Además, se rechaza H_0 también por el lado del estadístico ya que $F_0 = (9.03019) > f_{0.05,5,53} = (2.389)$. Al rechazar, se prueba que existe una relación de regresión.

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	0.8207	1.6488	0.4978	0.6207
β_1	0.2146	0.0820	2.6170	0.0115
β_2	0.0024	0.0311	0.0765	0.9393
β_3	0.0416	0.0134	3.1046	0.0031
β_4	0.0082	0.0075	1.0927	0.2794
β_5	0.0010	0.0007	1.4025	0.1666

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.

1.3. Interpretación de los parámetros

Solo los parámetros β_1 y β_3 son significativos, por lo que son los únicos que se pueden interpretar.

$\hat{\beta}_1$**:** Por cada unidad que aumenta la duración de la estadía (día), el riesgo de infección aumenta en probabilidad 0.2146 cuando las demás variables se mantienen constantes

$\hat{\beta}_3$**:** Por cada unidad que aumenta el número de camas, el riesgo de infección aumenta en probabilidad 0.0416 cuando las demás variables se mantienen constantes

1.4. Coeficiente de determinación múltiple R^2

2 pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.46$, lo que significa que aproximadamente el 46 % 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

9 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con los P-valor más pequeños 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	53.316	X1 X2 X3 X4 X5
Modelo reducido	80.850	X2 X4

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

$$Y_i = \beta_0 + \beta_2 X_{2i} + \beta_4 X_{4i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 59$$

2.2. Estadístico de prueba y conclusión

3 pt

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_4) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,53} \\ &= \frac{27.534}{1.00595} \\ &= 27.3711417 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,53} = 2.7791$, se puede ver que $F_0 > f_{0.95,1,45}$ y por tanto se rechaza la hipótesis nula. Esto lleva a concluir que el modelo reducido no es viable y que al menos algún β_j distinto de 0 para j=1, 3, 5 es significativo, que como se pudo ver en la tabla de valor-p β_1 y β_3 son significativos.

10 pt

¿se descartan o no?

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hacen las siguientes preguntas, ¿el efecto de duración de la estadía (días) sobre el riesgo de infección es igual al efecto del número de camas sobre el riesgo de infección? y ¿el efecto del censo promedio diario sobre el riesgo de infección es igual a dos veces el efecto del número de enfermeras sobre el riesgo de infección? . Por consiguiente se plantea la siguiente prueba de hipótesis para responder a estas preguntas:

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

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

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_2 X_{2i} + \beta_3 X^*_{3i} + \beta_5 X^*_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 59$$

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

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,53} \tag{3}$$

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,53} = \frac{(SSE(MR) - 53.3155)/2}{1.00595} \overset{H_0}{\sim} f_{2,53} \tag{4}$$

Al plantear una prueba de hipótesis lineal general y el estadístico de prueba, se posibilitaría responder acerca de la veracidad de las preguntas con el valor del estadístico.

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Para la validación de este supuesto, se planteará una prueba de hipótesis y una prueba analítica de normalidad (Shapiro-Wilk), acompañada de un gráfico cuantil-cuantil:

$$\begin{cases} H_0 : \varepsilon_i \sim \text{Normal} \\ H_1 : \varepsilon_i \nsim \text{Normal} \end{cases}$$

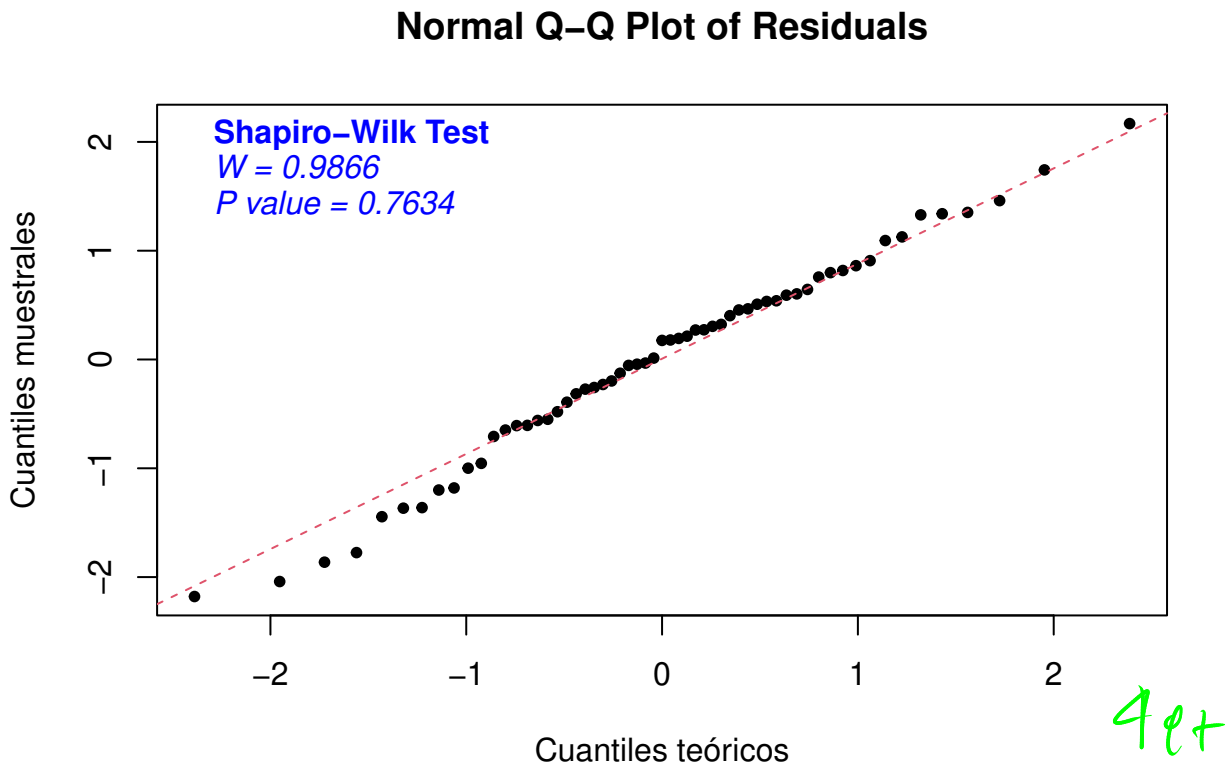

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.7634 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 una colas pesada a la izquierda 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

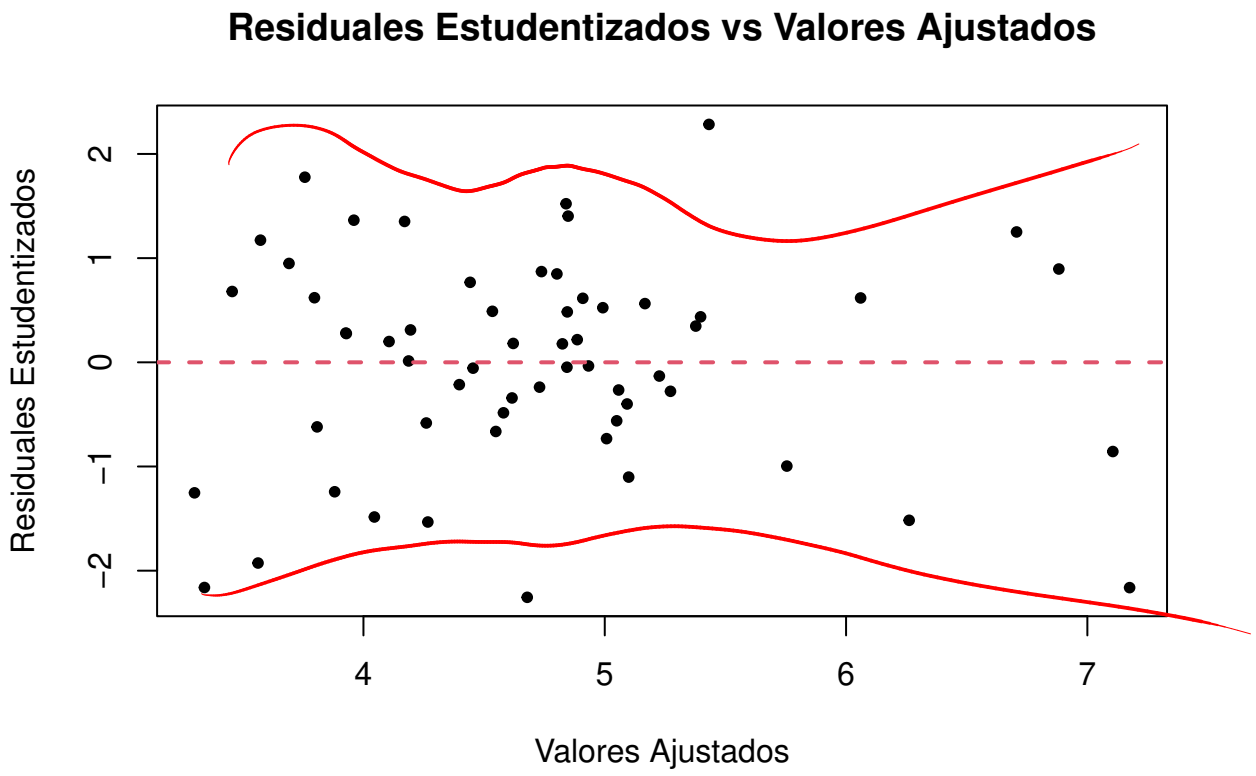

Figura 2: Gráfico residuales estudentizados vs valores ajustados

3pt

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.

Siempre sucede con residuales estudentizados.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

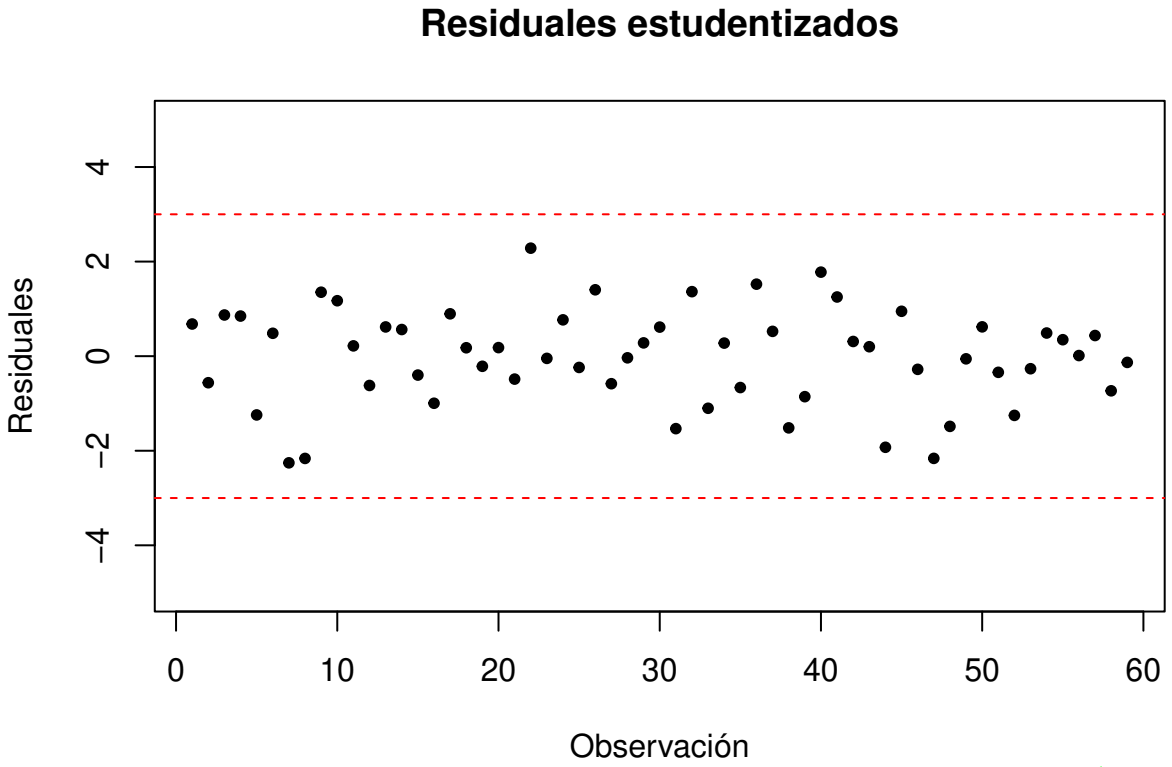

Figura 3: Identificación de datos atípicos

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 $|r_{estud}| > 3$. Esto muestra que ninguna observación está separada (en su valor de la respuesta Y) del resto de las observaciones por lo que no afecta los resultados del ajuste del modelo de regresión.

4.2.2. Puntos de balanceo

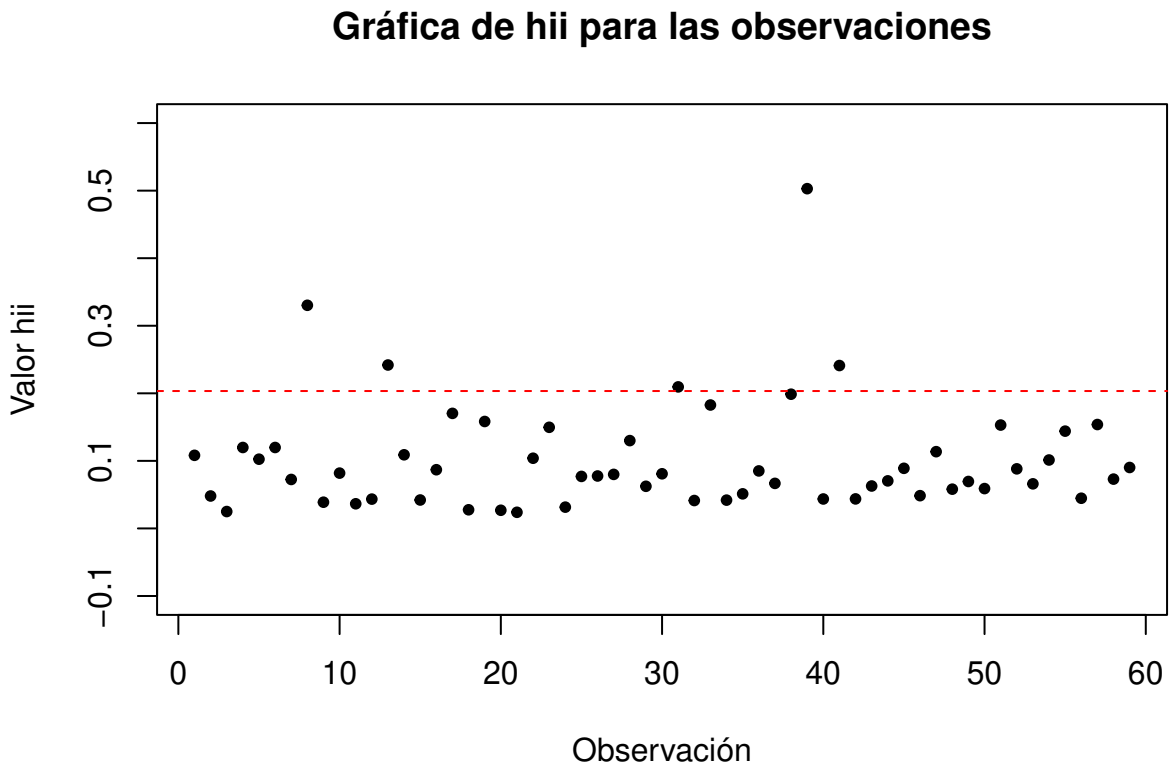

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 8   -2.1628  0.3845    0.3303 -1.5757
## 13   0.6179  0.0203    0.2419  0.3470
## 31  -1.5327  0.1038    0.2095 -0.7995
## 39  -0.8564  0.1237    0.5029 -0.8592
## 41   1.2513  0.0829    0.2411  0.7092
```

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. Estos cinco puntos de balanceo son observaciones en el espacio de las predictoras, alejadas del resto de la muestra y que pueden afectar estadística en el modelo como el R^2 y los errores estándar de los coeficientes estimados.

4.2.3. Puntos influenciales

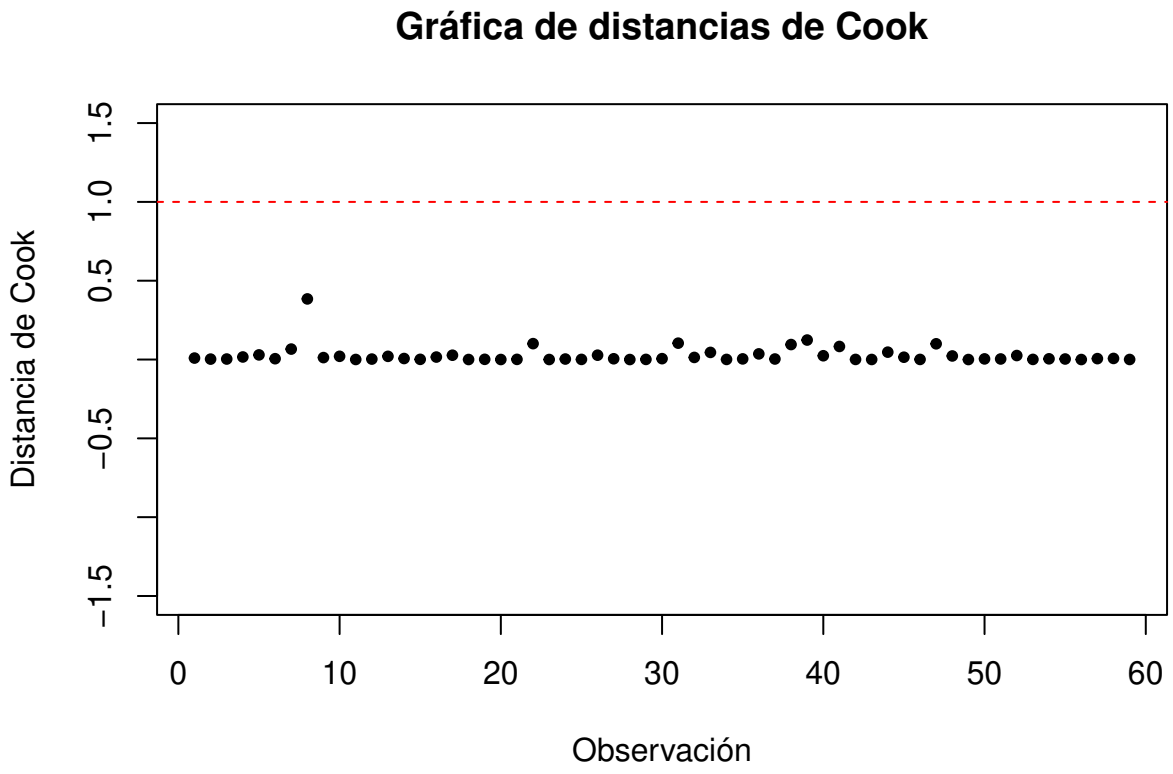

Figura 5: Criterio distancias de Cook para puntos influenciales

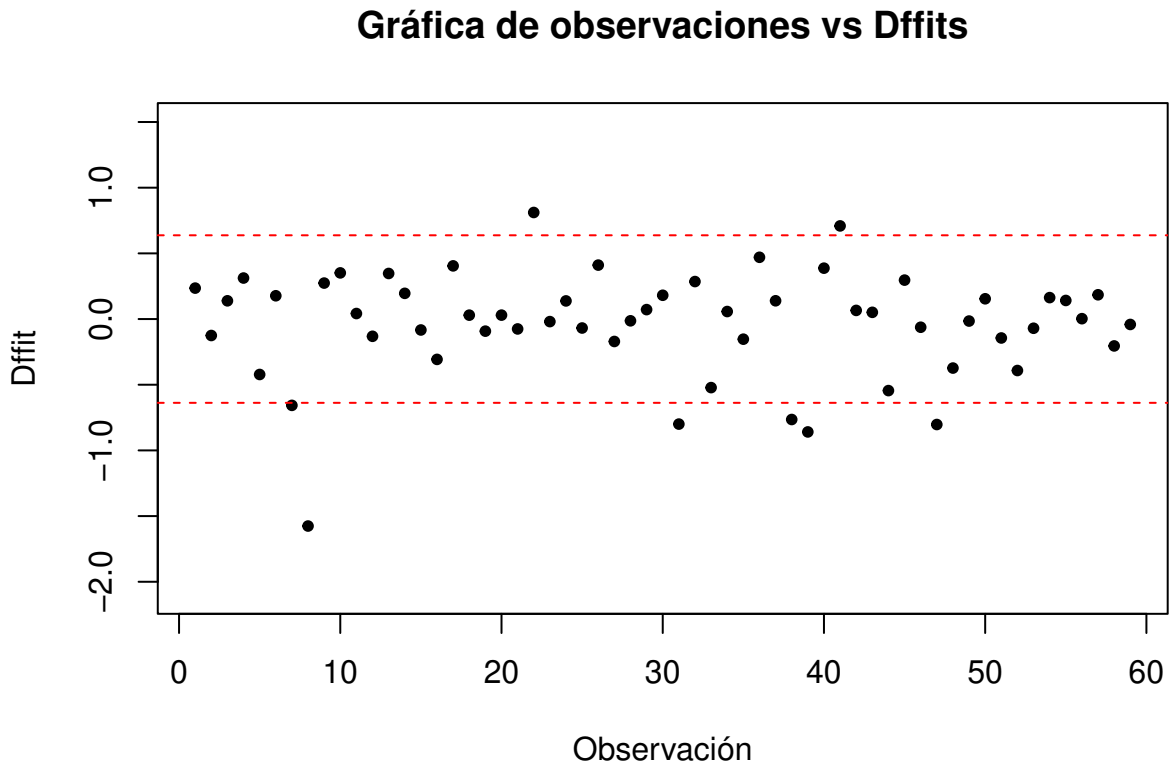

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 7   -2.2556  0.0663    0.0725 -0.6571
## 8   -2.1628  0.3845    0.3303 -1.5757
## 22   2.2836  0.1008    0.1039  0.8112
## 31  -1.5327  0.1038    0.2095 -0.7995
## 38  -1.5165  0.0950    0.1987 -0.7647
## 39  -0.8564  0.1237    0.5029 -0.8592
## 41   1.2513  0.0829    0.2411  0.7092
## 47  -2.1614  0.0998    0.1136 -0.8028
```

Como se puede ver,las ocho observaciones enlistadas 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. Las observaciones influenciales halan al modelo en su dirección y tienen un impacto notable sobre los coeficientes de regresión ajustados.

4.3. Conclusión

Para evaluar la validez del modelo además de cumplir los supuestos, entran a jugar también las observaciones extremas (observaciones atípicas, puntos de balanceo y puntos influenciables). En el caso de este modelo no hay observaciones atípicas, hay 5 puntos

de balanceo que afectan factores como el R^2 que mide la proporción de la variabilidad total observada en la respuesta que es explicada por el modelo propuesto, y hay 8 puntos influenciales; estos tienen un impacto alto sobre los coeficientes de regresión ajustados, su exclusión causa cambios puede tener importantes en la ecuación de regresión ajustada. Por otro lado el modelo no cumplió con el supuesto de normalidad, esto sumado a las observaciones extremas explicadas anteriormente dan a concluir que el modelo presentado no tiene validez.

→ Las observaciones extremas no necesariamente implican validez o no.