3,8

Trabajo 1

Estudiantes

Juan Pablo Diaz
Andres Camilo Sanchez
Edwin Javier Rocha
Santiago Gamboa

Docente

Francisco Javier Rodriguez Cortes

Asignatura

Estadística II

Sede Medellín

13 de Octubre de 2022

Índice

Índice de figuras

Índice de tablas

1. Pregunta 1

17 pt

Estime un modelo de regresión lineal múltiple que explique el riesgo de infección en términos de las variables restantes (actuando como predictoras) Analice la significancia de la regresión y de los parámetros individuales. Interprete los parámetros estimados. Calcule e interprete el coeficiente de determinación múltiple R2

Teniendo en cuenta la base de datos asignada, la cual es **Equipo33.txt**, las covariables son:

Variable	Descripción
Y: Riesgo de infección(Rinf)	Probabilidad promedio estimada de adquirir infección en el hospital
X_1: Duración de la estadía(DE)	Duración promedio de la estadía de todos los pacientes en el hospital
X_2: Rutina de cultivos(RC)	Razón del número de cultivos realizados en pacientes sin síntomas de infección hospitalaria, por cada 100 pacientes
X_3: Número de camas(NC)	Promedio de camas en el hospital durante el periodo del estudio
X_4: Censo promedio diario	Número promedio de pacientes en el hospital por día durante el periodo del estudio
X_5: Número de enfermeras(Nenf)	Promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio

El modelo que se propone es:

$$Rinf_i = \beta_0 + \beta_1 DE_i + \beta_2 RC_i + \beta_3 NC_i + \beta_4 CPD_i + \beta_5 NEnf_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

$i = 1, 2, \ldots, 60$

1.1. Modelo de regresión

2 pt

Al ajustar el modelo de regresion para el riesgo de infeccion en un hospital, se obtienen los siguientes coeficientes:

Tabla 2: Tabla de valores de los coeficientes estimados

	Valor del parámetro
$\hat{\beta}_0$	-2.5355
$\hat{\beta}_1$	0.2430
$\hat{\beta}_2$	0.0394
$\hat{\beta}_3$	0.0741
$\hat{\beta}_4$	0.0117
$\hat{\beta}_5$	0.0018

No va en ec. ajustada

Por lo tanto, el modelo de regresión ajustado es:

$$\widehat{Y.Rinf}_i = -2.5355 + 0.243 DE_i + 0.0394 + 0.0741 NC_i + 0.0117 CPD_i + 0.0018 NEnf_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Donde $1 \leq i \leq 60$

1.2. Significancia de la regresión

4 pt

Se Plantea el siguiente Juego de Hipotesis

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_a : \text{Algún } \beta_j \neq 0 \ para \ j = 1, 2, 3, 4, 5 \end{cases}$$

Para la significancia de la regresión se hará uso de la siguiente tabla anova:

Tabla 3: Tabla anova significancia de la regresión

	Sumas de cuadrados	g.l	Cuadrado medio	F_0	Valor-P
Modelo de regresión	66.8182	5	13.363648	15.5352	1.87507e-09
Error	46.4516	54	0.860215		

Al observar los resultados de la Tabla Anova,La evidencia muestral nos dice que se rechaza la hipotesis nula, por tanto la evidencia muestral nos indica que la regresion es significativa

¿Cuál es esa evidencia? Mencionar Val-P

1.3. Significancia de los parámetros

6 pt

$$\begin{cases} \mathrm{H}_0 : \beta_j = 0 \\ \mathrm{H}_a : \text{Algún } \beta_j \neq 0 \ para \ \boxed{j = 1, 2, 3, 4, 5} \end{cases}$$

→ β_0?

En el siguiente tabla se presentara información los criterios de para determinar si los parametros son significativos individualmente

Tabla 4: Resumen de los coeficientes

	Estimación β_j	$se(\hat{\beta}_j)$	T_{0j}	Valor-P
β_0	-2.5355	1.7930	-1.4141	0.1631
β_1	0.2430	0.0999	2.4331	0.0183
β_2	0.0394	0.0332	1.1865	0.2406
β_3	0.0741	0.0179	4.1448	0.0001
β_4	0.0117	0.0072	1.6264	0.1097
β_5	0.0018	0.0007	2.4101	0.0194

Los resultados de las pruebas: valor del estadístico de prueba y el valor p para la prueba se obtiene en las dos últimas columnas de la tabla de los parámetros estimados.

Con un nivel de significancia$\alpha = 0.05$ se concluye que los parámetros individuales β_1,β_3,β_5 son significativos cada uno en presencia de los demás parametros Por el contrario los parametros β_0,β_2,β_4 individualmente no son significativos en presencia de los demas parametros

1.4. Interpretación de los parámetros

2 pt

- $\hat{\beta}_0$= -2.5355: El parámetro $\hat{\beta}_0$ no es interpretable porque no es significativo, ademas ninguna variable predictora lo contiene.

- $\hat{\beta}_1$= 0.243: Indica que por cada dia que aumente la Duracion de la estadia en el hospital(Dias), el promedio del porcentaje del Riesgo de infeccion aumenta en 0.243 cuando las demas variables predictoras se mantiene fijas.
- $\hat{\beta}_2$= 0.0394: El parámetro $\hat{\beta}_2$ no es interpretable porque no es significativo.
- $\hat{\beta}_3$= 0.0741: Indica que por cada unidad que aumente el numero promedio de camas en el hospital en el periodo de estudio, el promedio del Riesgo de infeccion aumenta en ~~0.0741~~% cuando las demas variables predictoras se mantiene fijas. → porcentaje promedio 7,41%
- $\hat{\beta}_4$= 0.0117: El parámetro $\hat{\beta}_2$ no es interpretable porque no es significativo.
- $\hat{\beta}_5$= 0.0018: Por cada unidad que aumenta el numero promedio de enfermeras en el hospital, el riesgo de infeccion aumenta promedio en ~~0.0018~~% cuando las demás variables predictoras se mantienen fijas. → porcentaje promedio

0,18%

1.5. Coeficiente de determinación múltiple R^2 3 pt

El modelo tiene un R^2 de 0.5899 lo cual significa que aproximadamente el \$ 58.99 %\$ de la variabilidad total en el porcentaje de Riesgo de infeccion es explicado por el modelo RLM

¿Cómo se calcula?

1.6. Comentarios

En el modelo, se observa que las variables que tienen una contribución significativa en la regresión son Duración de la estadía (DE), Número de camas (NC), y Número de enfermeras (NEnf) ~~según la metrias del R2 y~~ la significancia de los parametros.

2. Pregunta 2 2,5 pt

Use la tabla de todas las regresiones posibles, para probar la significancia simultánea del subconjunto de tres variables con los valores p más grandes del punto anterior. Según el resultado de la prueba es posible descartar del modelo las variables del subconjunto? Explique su respuesta.

2.1. Planteamiento prueba de hipotesis y modelo reducido

Los parametros cuyos valores P fueron los más altos corresponden a β_2 con VP=0.2406, β_4 con VP= 0.1097, β_5 con VP= 0.0194. Por tanto, se plantea la siguiente prueba de hipótesis:

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

El modelo completo es el definido en la sección 1.1, y el modelo reducido es:

$$\text{MR: } Rinf_i = \beta_0 + \beta_1 DE_i + \beta_3 NC_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Se presenta la siguiente tabla con el resumen de todas las regresiones para plantear el estadístico de prueba:

Tabla 5: Resumen de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	46.452	X1 X2 X3 X4 X5
Modelo reducido	53.204	X1 X3

→ Así no se llaman sus var's

$F_0 = \frac{SSR(\beta_2, \beta_4, \beta_5 \mid \beta_0, \ldots, \beta_5)/3}{MSE(MF)} \sim F_{3,54}$

2.2. Estadístico de prueba y conclusiones

Se construye el estadístico de prueba como:

1,5 pt

$$F_0 = \frac{(SSR(\beta_0, \beta_1, \beta_3, |\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,54}$$

$$F_0 = \frac{(SSE(\beta_0, \beta_1, \beta_3) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5))/3}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)} \overset{H_0}{\sim} f_{2,54} \to 3$$

$$= \frac{(53.204 - 46.452)/2}{46.452/54} = 3.924567$$

1 pt

Ahora, comparando a un nivel de significancia $\alpha = 0.05$, F_0 con $f_{0.05,2,54} = 3.168246$. Con valor P=0.0256226

Note que F_0= 3.924567 es mayor al $F_{\alpha=0.05,2,54}$= 3.168246 de la distribución, y el valor P es pequeño. Por tanto, la evidencia apunta a rechazar H_0, entonces podemos deducir atraves de la evidencia muestral nos indica que existe al menos un parametro que si es significativo en el sub conjunto

¿se pueden descartar?

3. Pregunta 3

3,5 pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \text{H}_0 : \beta_1 = \beta_2, \beta_3 = \beta_4 \\ \text{H}_a : \text{Alguna de las desigualdades no se cumple} \end{cases}$$

Reescribendo matricialmente:

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

2 pt

Donde $\mathbf{L}$ está dada por:

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

Donde el modelo reducido está dado por:

0 pt

$$Rinf = \beta_0 + \beta_1(DE_i + CDP_i) + \beta_3(NC_i + CDP_i) + \beta_5(Ht + Wt)\varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Ht? Wt? usaron el trabajo del semestre pasado? ojo con que era sea plagio.

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,54}$$

1,5 pt

¿?

$$F_0 = \frac{(SSE(MR) -)/2}{0.860215} \overset{H_0}{\sim} f_{2,54}$$

Obteniendo esto podemos definir la region de rechazo de la hipotesis nula como $F_0 > F_{0.05,2,54} = 3.168246$ y con valor p:P($F_{2,54} > |F_0|$)

4. Pregunta 4 15 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales 2 pt

Para la validación de este supuesto, se plantea la siguiente prueba de hipótesis (~~shapiro wilk~~)

$$\begin{cases} \text{H}_0: & \varepsilon_i \sim N(\mu, \sigma^2) \\ \text{H}_a: & \varepsilon_i \nsim N(\mu, \sigma^2) \end{cases}$$

No están probando media constante μ ni var cte σ^2 con esta prueba.

acompañado de un grafico cuantil-cuantil:

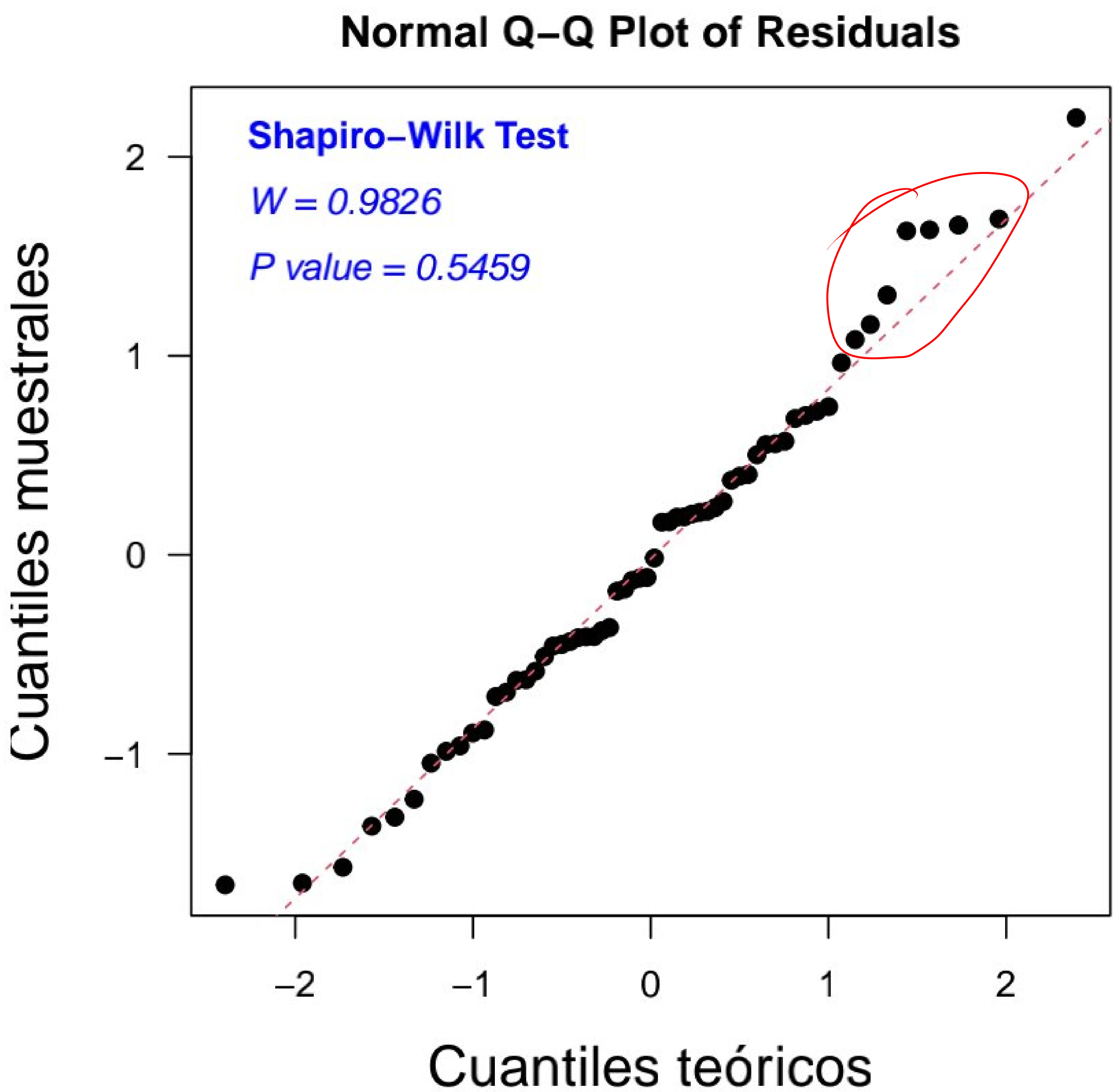

Figura 1: Gráfico cuantil-cuantil y normalidad de los residuales

Como el valor P es es grande, por tanto la evidencia muestra apunta no rechazar H_0, luego el modelo comple con el supuesto de Normalidad de los residuales

No hicieron análisis gráfico que es más importante.

4.1.2. Media 0 y Varianza constante 1,5 pt

En esta prueba se quiere probar

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

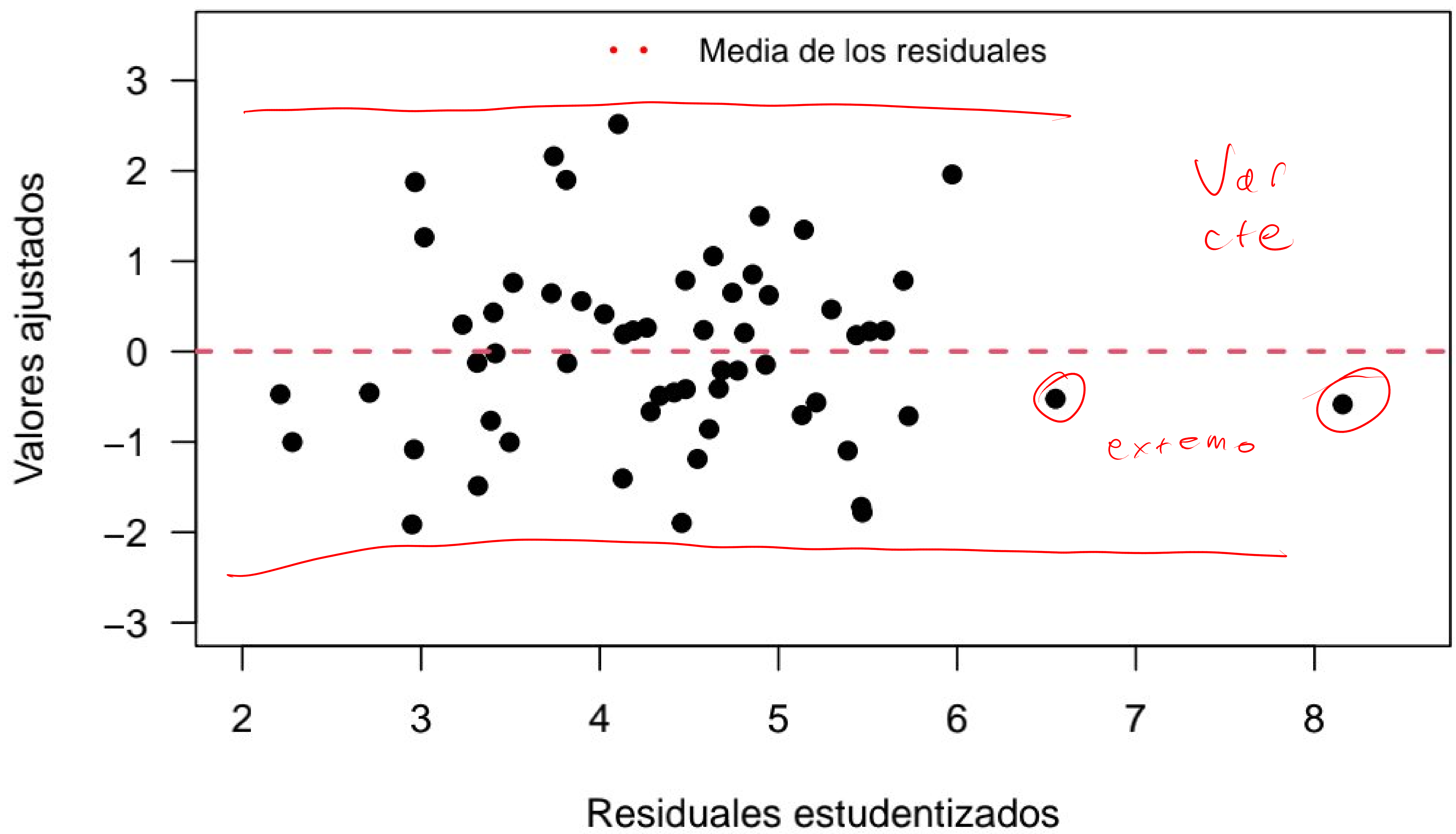

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Notese que la linea punteada roja, que representa la media de los errores, esta en cero o muy cercana a este por lo que podemos concluir que los errores tiene media cero × Res. estud siempre la tienen, eso se ve con los res. crudos
Tambien notemos que no se obseva ningun patron en los residuales luego podemos concluir que la varianza de estos es constate

Patrones de aumento o decrecimiento.
No es observar cualquier patrón
ya que puede haber no linealidad
y aún tener var cte

4.2. Observaciones extremas

4.2.1. Datos atípicos

3 pt

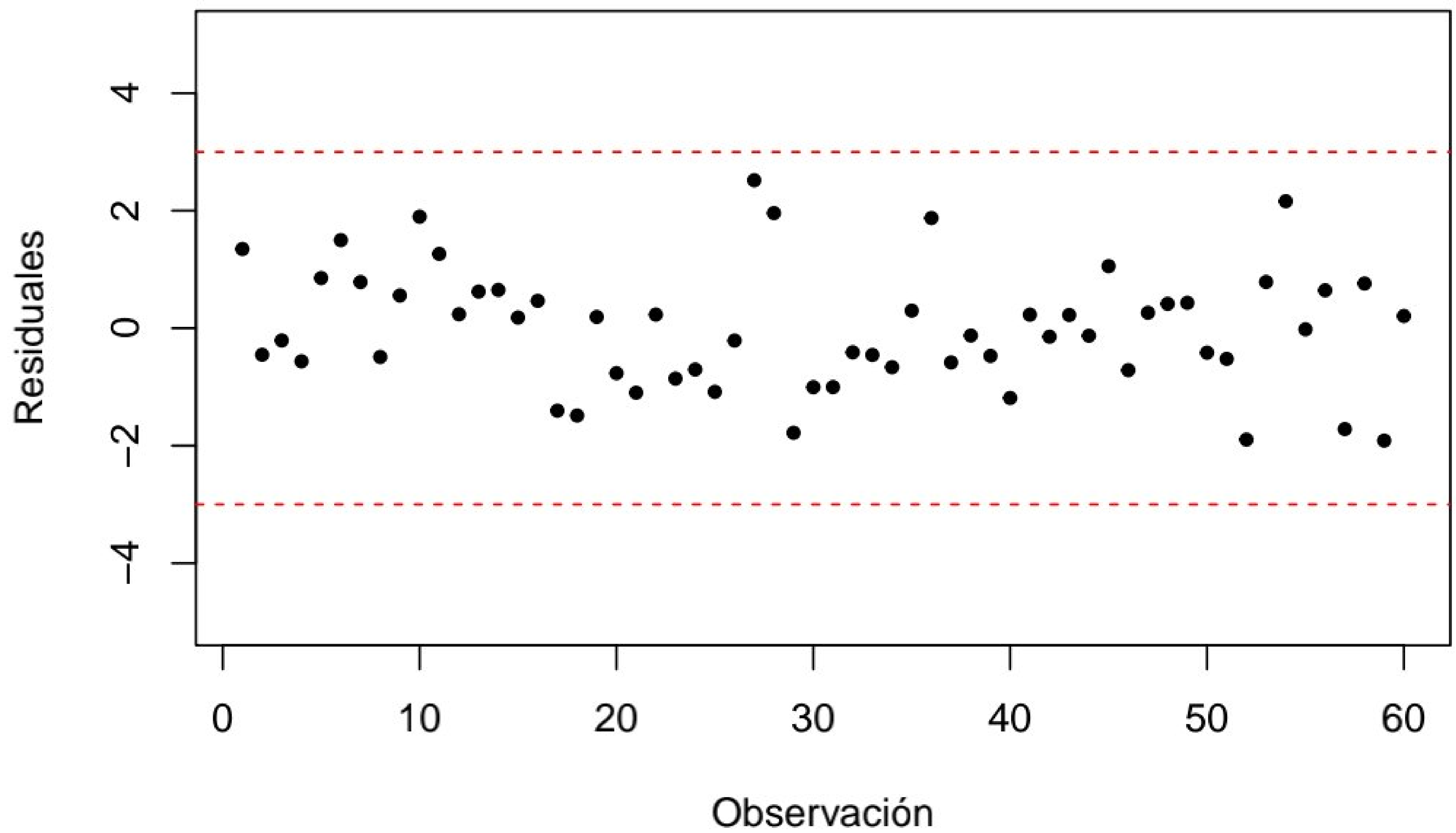

Figura 3: Identificación de datos atípicos

Segun la Figura 3, no hay datos atipicos bajo el criterio de los residuales estudenrizados ya que ningun $|r_i|>3$.

4.2.2. Puntos de balanceo

3 pt

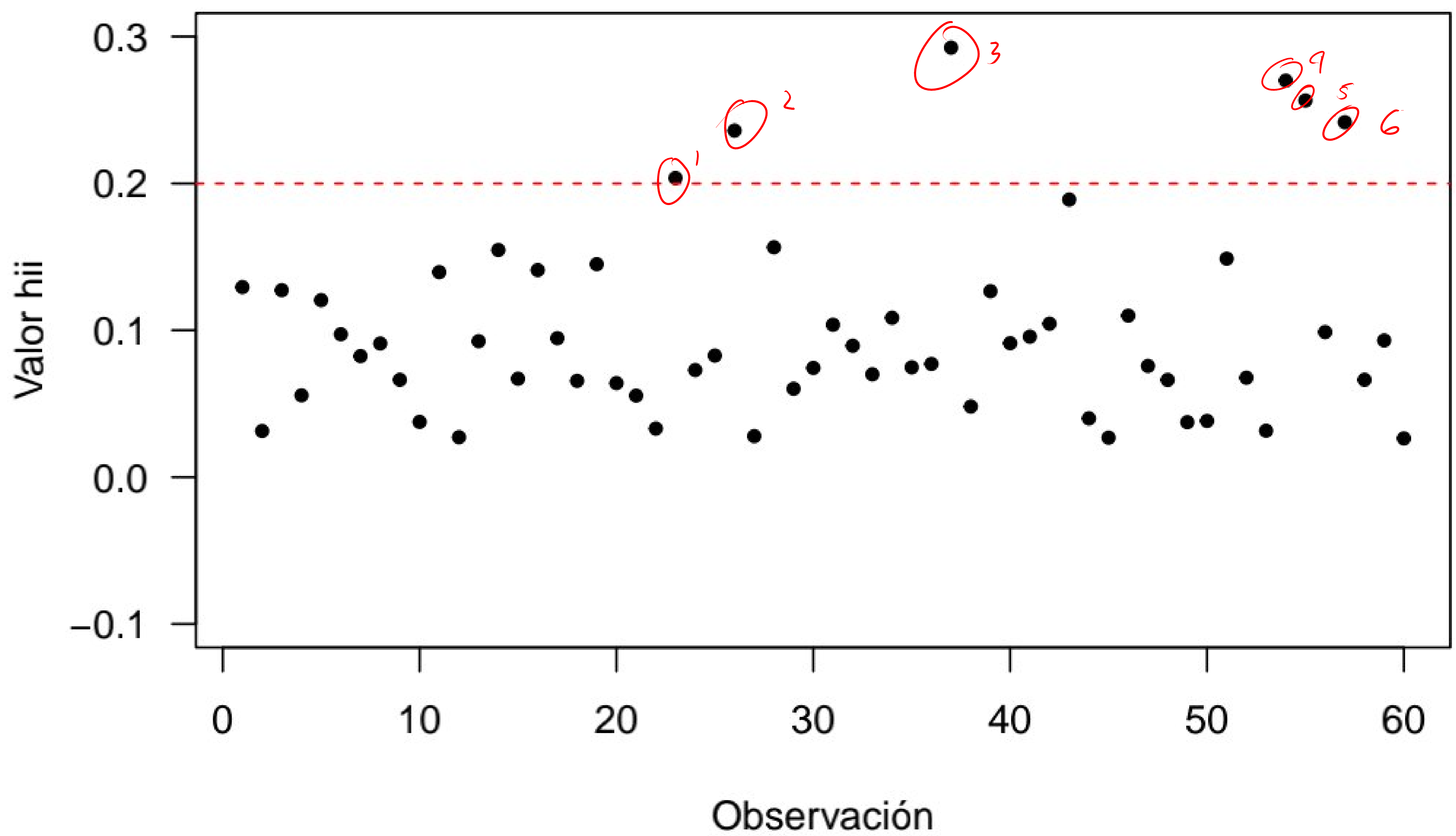

Figura 4: Identificación de puntos de balanceo

tienen 6

Analizando los elementos de la diagonal principal de la matriz Hat vs las observaciones, obtenemos el gráfico anterior. A partir de este gráfico podemos concluir que el modelo tiene 5 puntos de balanceo. Estos puntos de balanceo controlan ciertas propiedades del modelo, como el R^2 y los errores estandar de los coeficientes estimados. Es decir, estos puntos causan una sobreestimación en el R^2 y pueden afectar el supuesto de varianza, media y normalidad. ✓ Muy bien

Tabla 6: Tabla de puntos de Balanceo

	Errores Estudentizados	D.Cook	Valor hii	DFFITS
23	-0.8582	0.0315	0.2037	-0.4340
26	-0.2104	0.0023	0.2360	-0.1169
37	-0.5832	0.0237	0.2925	-0.3750
54	2.1610	0.2695	0.2700	1.3142
55	-0.0198	0.0000	0.2565	-0.0116
57	-1.7182	0.1513	0.2417	-0.9700

Notese que los datos de balanceo que deben ser investigados son los datos 23,26,37,54,55 y57, ya que estos cumplen con el criterio el siguiente criterio $h_{ii} > \frac{2p}{n}$

4.2.3. Puntos influenciales

Bajo el criterio de Cook, se hace la siguiente gráfica:

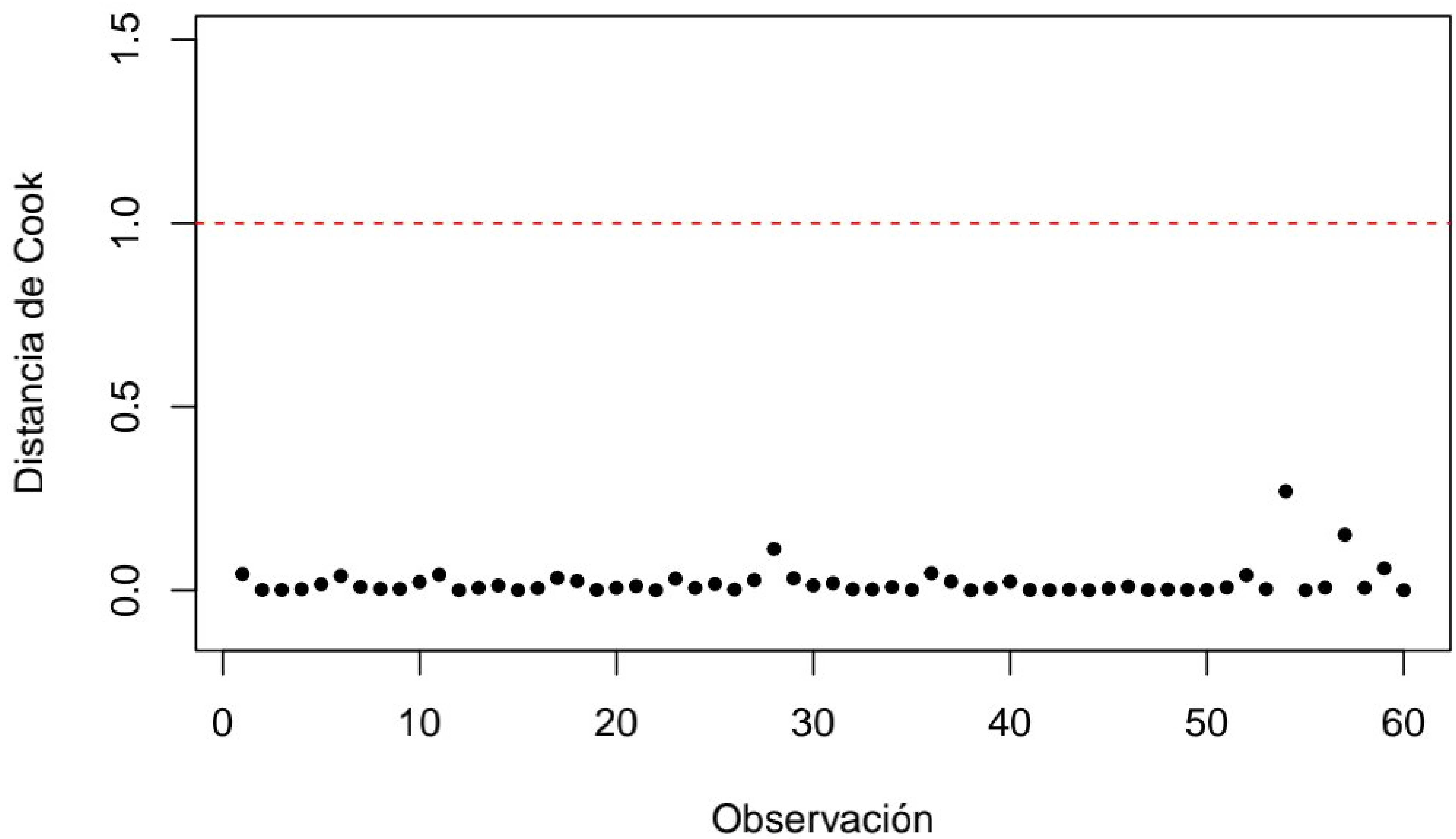

Figura 5: Criterio distancias de Cook para puntos influenciales

Bajo el criterio de cook, ~~se obtuvo la~~ anterior gráfica. A partir de la gráfica podemos concluir que no existen puntos influenciales bajo este criterio

2 pt

muy redundantes

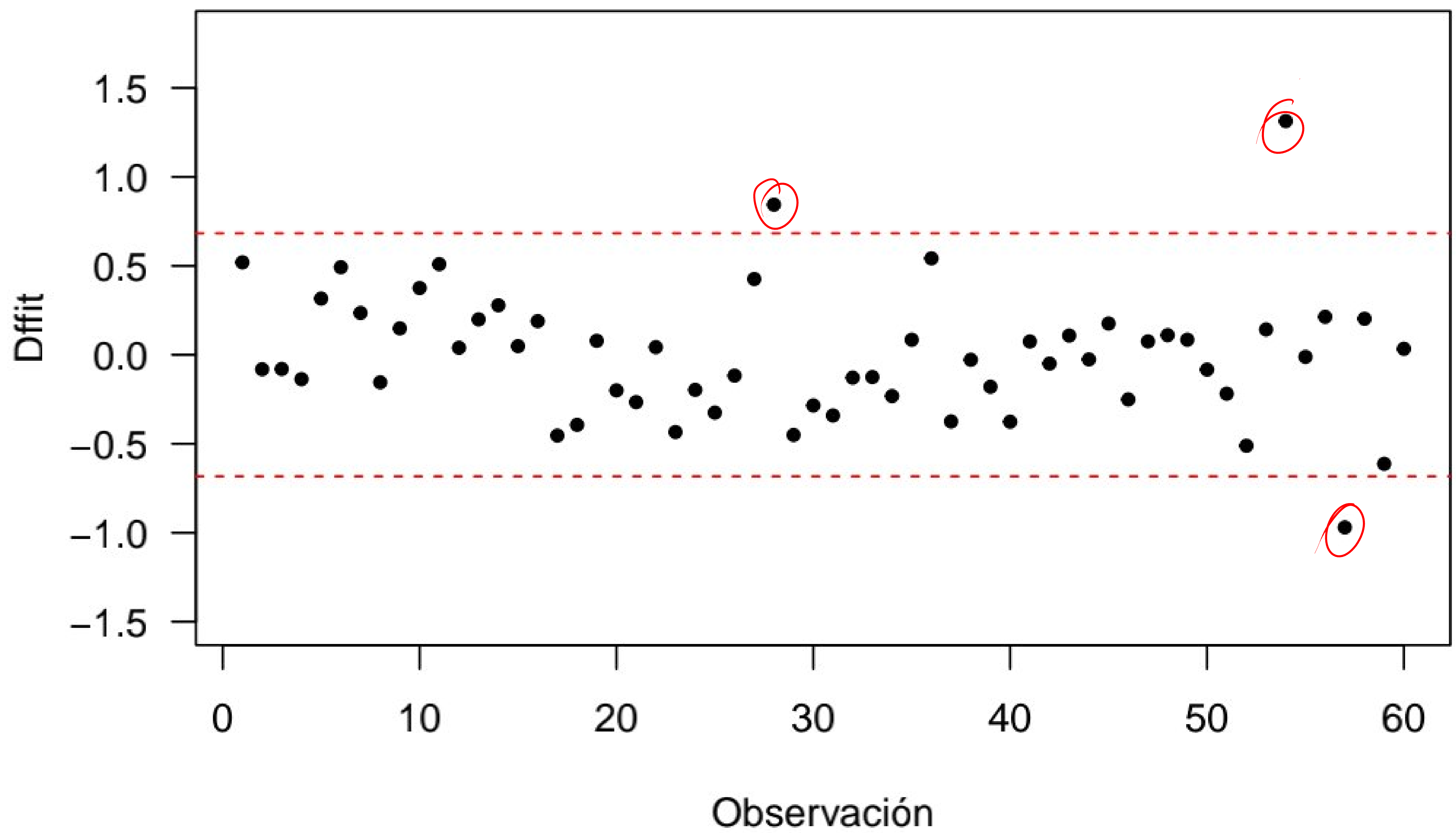

Figura 6: Criterio Dffits para puntos influenciales

1,5 pt

Tabla 7: Tabla del criterio DFFITS para encontrar puntos influenciales

	Errores Estudentizados	D.Cook	Valor hii	DFFITS
28	1.9602	0.1128	0.1565	0.8442
54	2.1610	0.2695	0.2700	1.3142
57	-1.7182	0.1513	0.2417	-0.9700

según este criterio es sobre !

Bajo el criterio de Dffits, se obtuvo la anterior gráfica. A partir de la gráfica podemos concluir que existen varios valores influenciales en el modelo ($\cancel{Y_i}$, i = 28, 54, 57). Como los coeficientes de Dffits son mayores a $2 * \sqrt{(\frac{7}{70})} = 0.6325$, podemos afirmar que los Y_i, i = 28, 54, 57 tienen influencia sobre los $\boxed{\beta_i}$, i = 0,. . . ,6. Por ende, tales datos deben ser investigados para determinar su influencia sobre el modelo de regresion

4.3. Conclusiones

2 pt

entonces es válido o no ?

El modelo de regresión parece ajustarse adecuadamente a los supuestos de normalidad y homocedasticidad. Sin embargo, se observan algunos puntos influyentes que parecen tener un impacto significativo en los resultados del modelo. Por lo tanto, es importante considerar estos puntos y evaluar su posible eliminación o ajuste en futuros análisis para mejorar la calidad de las predicciones del modelo