Trabajo 1 Grupo 61

3,8

Esteban Cifuentes Arias-1036678145

Evelyn Tatiana Mina Valverde - 1192909536

Jaider Arley Silva Hurtado - 1004612145

Docente

Javier Armando Lozano

Asignatura

Estadística II

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medellín
05 de octubre de 2023

Índice general

Índice de figuras

Índice de cuadros

18pt

0.1. Pregunta 1

Teniendo en cuenta la base de datos 32, en la cual hay 5 variables regresoras denominadas por:
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.
Entonces se plantea el siguiente modelo de regresión para estimar el riesgo de infección:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2),\ 1 \leq i \leq 45$$

0.1.1. Modelo de regresión

Al ajustar el modelo de regresión propuesto, se obtienen los siguientes coeficientes:

Parámetro	Valor estimado
$\hat{\beta}_0$	-2.5926
$\hat{\beta}_1$	0.2302
$\hat{\beta}_2$	0.0614
$\hat{\beta}_3$	0.0617
$\hat{\beta}_4$	0.0014
$\hat{\beta}_5$	0.0022

Cuadro 1: Tabla coeficientes del modelo

3pt

De la tabla 1, se tiene que el modelo de regresión ajustado está dado por:

$$\hat{Y}_i = -2.5926 + 0.2302 X_{1i} + 0.0614 X_{2i} + 0.0617 X_{3i} + 0.0014 X_{4i} + 0.0022 X_{5i}$$

0.1.2. Significancia de la regresión

Se plantea el siguiente juego de hipótesis con el fin de analizar la significancia de la regresión.

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

Para esta prueba de hipótesis se tiene el siguiente estadístico de prueba:

$$F_0 = \frac{MSR}{MSE} \overset{bajo H_0}{\sim} f_{5,6} = \frac{57.6743/5}{33.8981/39} = 13.2709$$

Para analizar la significancia de la regresión, también se usará la siguiente tabla ANOVA:

	Suma de cuadrados	g.l.	Cuadrado medio	F_0	Valor-P
Regresión	57.6743	5	11.534864	13.2709	1.42292e-07
Error	33.8981	39	0.869183		

Cuadro 2: Tabla ANOVA para el modelo

De la tabla 2 se observa que el valor-p =$1.42292e^{-07} < \alpha = 0.05$, luego, con un nivel de significancia de $\alpha = 0.05$, se tiene que hay evidencia estadística suficiente para rechazar H_0 y se concluye que el modelo RLM es significativo. Esto quiere decir que el riesgo de infección depende significativamente de al menos una de las variables predictoras del modelo.

0.1.3. Significancia de los parámetros

Se plantea el siguiente juego de hipótesis con el fin de analizar la significancia individual de cada uno de los parámetros del modelo.

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

En la siguiente tabla se presenta la información resumida de cada uno de los parámetros del modelo, la cual permitirá analizar la significancia de cada uno de ellos.

	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	T_{0j}	Valor-P
β_0	-2.5926	1.6572	-1.5645	0.1258
β_1	0.2302	0.1149	2.0042	0.0520
β_2	0.0614	0.0303	2.0230	0.0499
β_3	0.0617	0.0162	3.8052	0.0005
β_4	0.0014	0.0079	0.1797	0.8583
β_5	0.0022	0.0009	2.4554	0.0186

Cuadro 3: Resumen de los coeficientes del modelo

Con un nivel de significancia de $\alpha = 0.05$, se puede concluir de la tabla 3 que los parámetros individuales β_2, β_3 *y* β_5 son significativos, ya que el valor-P de cada uno de ellos es menor a 0.05,

entonces se tiene evidencia estadística suficiente para rechazar la hipótesis nula y se concluye que estos parámetros son significativos en presencia de los demás parámetros.

Por otro lado, se tiene que los valores-P de los parámetros $\beta_0, \beta_1, \ y \ \beta_4$ son mayores a $\alpha = 0.05$, entonces no se tiene evidencia estadística suficiente para rechazar la hipótesis nula y se concluye que estos parámetros no son significativos en presencia de los demás parámetros.

0.1.4. Interpretación de los parámetros

$\hat{\beta_2}$: Indica que por cada unidad que aumenta la rutina de cultivos durante el periodo de estudio, la tasa de Probabilidad promedio de adquirir la infección en el hospital aumenta 6.14 % cuando los demás paramétros se mantienen fijos.

$\hat{\beta_3}$: Indica que por cada unidad que aumenta el número de camas en el hospital durante el periodo del estudio, la tasa de Probabilidad promedio de adquirir la infección en el hospital aumenta 6.17 % cuando los demás paramétros se mantienen fijos.

$\hat{\beta_5}$: Indica que por cada unidad que aumenta el número de enfermeras en el hospital, la tasa de Probabilidad promedio de adquirir la infección en el hospital aumenta 0.22 % cuando los demás paramétros se mantienen fijos.

0.1.5. Coeficiente de determinación múltiple de R^2

El coeficiente de determinación múltiple R^2 está dado por:

$$R^2 = \frac{SSR}{SST} = \frac{57.6743}{57.6743 + 33.8981} = 0.6298$$

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.6298$, esto quiere decir que aproximadamente el 62.98 % de la variabilidad total observada en el riesgo de infección es explicada por el modelo de regresión propuesto.

0.2. Pregunta 2

0.2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables cuyo valor-p son los más altos corresponden a; $X_1, X_2 \ y \ X_4$.
Por lo tanto, usando la tabla de todas las posibles regresiones, para probar la significancia del subconjuto de las tres covariables mensionadas anteriormente, se plantea la siguiente prueba de hipótesis:

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

El modelo completo es el definido en la sección 0.1.1 y el modelo reducido para la prueba de significancia del subconjuto es:

$$MR : Y = \beta_o + \beta_3 X_{3i} + \beta_5 X_{5i} + \epsilon_i, \ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \ 1 \leq i \leq 45$$

	SSE	Covariables en el modelo
Modelo completo	33.898	X_1 X_2 X_3 X_4 X_5
Modelo reducido	44.428	X_3 X_5

Cuadro 4: Resumen de tabla con todas las regresiones

0.2.2. Estadístico de prueba y conclusión

$$F_0 = \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5))/3}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)} \overset{bajo H_0}{\sim} f_{3,39} = \frac{(44.428 - 33.898)/3}{33.898/39} = 4.0383$$

Con un nivel se significancia de $\alpha = 0.05$ se tiene que $F_0 = 4.0383 > f_{3,39} = 2.8451$, entonces se rechaza la hipótesis nula y se concluye que el riesgo de infección depende de al menos una de las variables del subconjunto, por lo que no se pueden descartar del modelo.

0.3. Pregunta 3

Se quiere saber si el efecto sobre el riesgo de infección de la rutina de cultivos es igual al efecto de número de camas del hospital y si el efecto del número de enfermeras es dos veces el efecto del censo promedio diario.

0.3.1. Prueba de hipótesis y prueba de hipótesis matricial

$$\begin{cases} H_0 : \beta_2 = \beta_3; \beta_5 = 2\beta_4 \\ H_a : \text{Alguna de las desigualdades no se cumple} \end{cases}$$

Esto se puede reescribir matricialmente como:

$$\begin{cases} H_0 : L\underline{\beta} = \underline{0} \\ H_a : L\underline{\beta} \neq \underline{0} \end{cases}$$

Con **L** dada por:

$$\begin{bmatrix} 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2 & 1 \end{bmatrix} \quad (1)$$

El modelo reducido está dado por.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X^*_{2i} + \beta_4 X^*_{4i} + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2),\ 1 \leq i \leq 45$$

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

0.3.2. Estadístico de prueba

$$F_0 = \frac{(SSE(\beta_0, \beta_1, \beta_2, \beta_4) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5))/2}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)} \overset{bajo H_0}{\sim} f_{2,39}$$

$$F_0 = \frac{(SSE(\beta_0, \beta_1, \beta_2, \beta_4) - 33.898)/2}{33.898/39} \overset{bajo H_0}{\sim} f_{2,39}$$

0.4. Pregunta 4

0.4.1. Supuestos del modelo

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:

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

Se tiene un juego de hipótesis en el que se aceptará H_0 si se tiene un p-valor > 0.05, se realizó un gráfico en el que se analizará la normalidad de los datos a través de los residuales

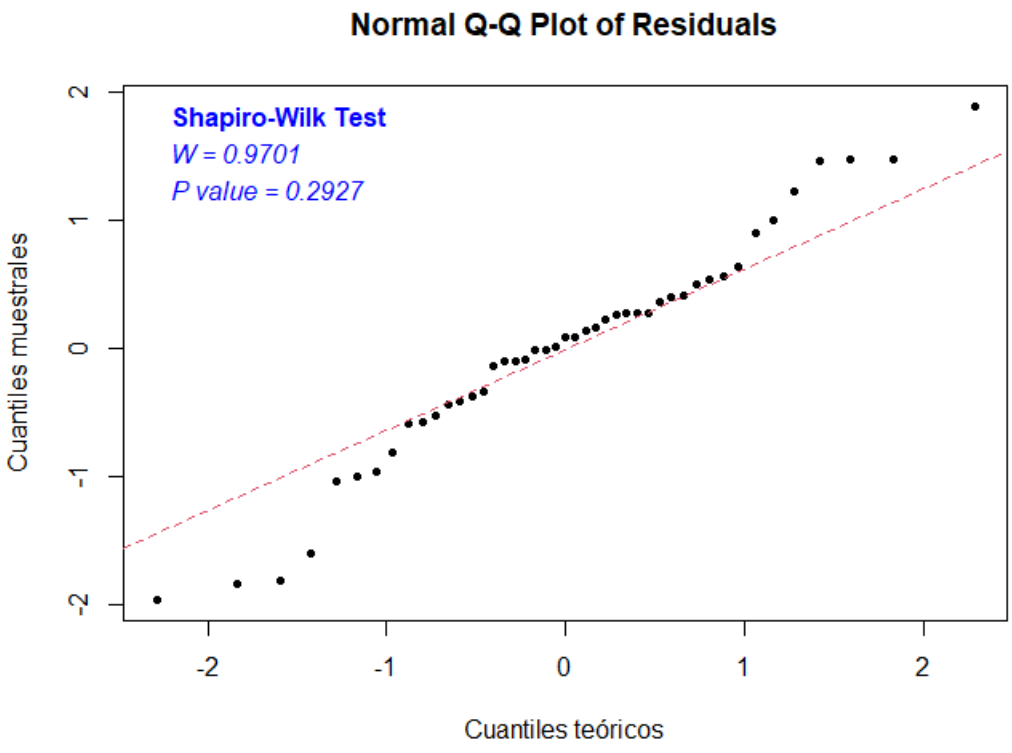

Figura 1: Q-Q plot para análisis de normalidad en los residuales.

Más análisis, si rechazaban

De la Figura 1 se puede concluir que no hay evidencia estadística suficiente para rechazar H_0,por tanto, dado un p-valor de $0.2927 > 0.05$ los datos presentan normalidad con media μ y varianza σ^2 ; aunque todos los puntos no tengan un ajuste perfecto a la recta esto no afecta el supuesto de normalidad y se analizará posteriormente si estos puntos son datos atípicos.

Varianza constante

Se realizará un gráfico de residuales estudentizados, los cuales no son necesariamente independientes ni tienen varianza constante, en comparación con los valores ajustados ; esto se realiza con el objetivo de analizar si la varianza es constante.

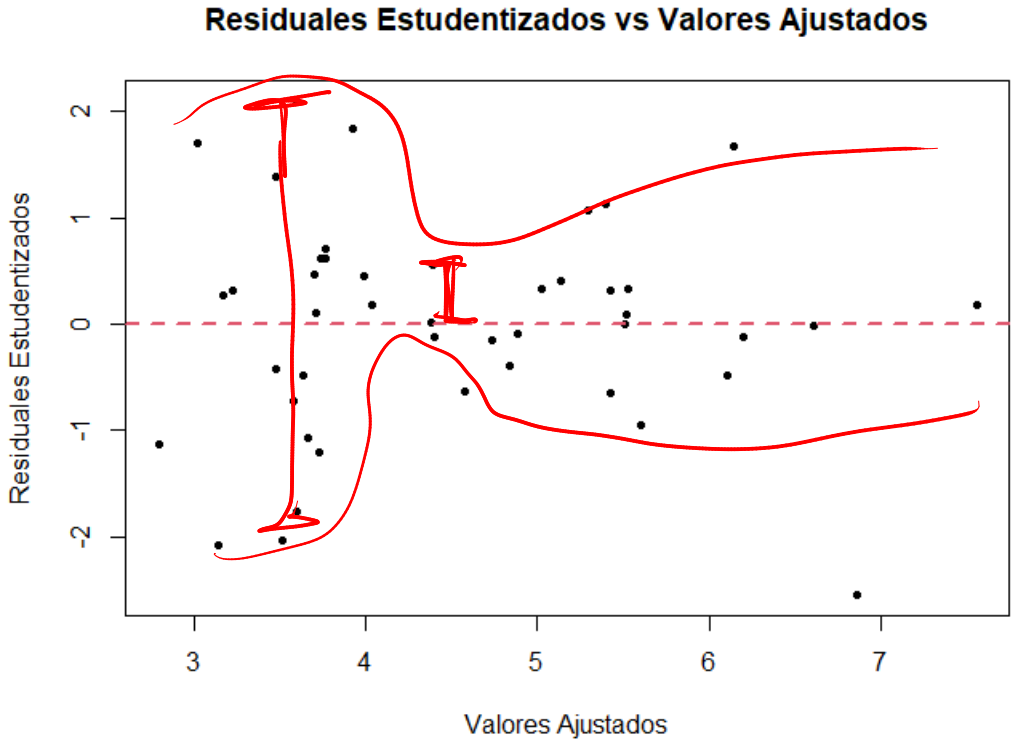

Figura 2: Gráfico residuales estudentizados vs valores ajustados

3 pt

De la figura 2 se analiza que no hay tendencias marcadas entre los puntos, ni comportamientos entre estos que permitan concluir que la varianza no es constante; por tanto, se concluye que la varianza del modelo es constante.

0.4.2. Verificación de las observaciones

Datos atípicos

Como se mencionó previamente, dado el análisis de la figura 1, se observaron datos que no se ajustaban de manera apropiada a la recta, por tanto, se analizará si son datos atípicos y su naturaleza.

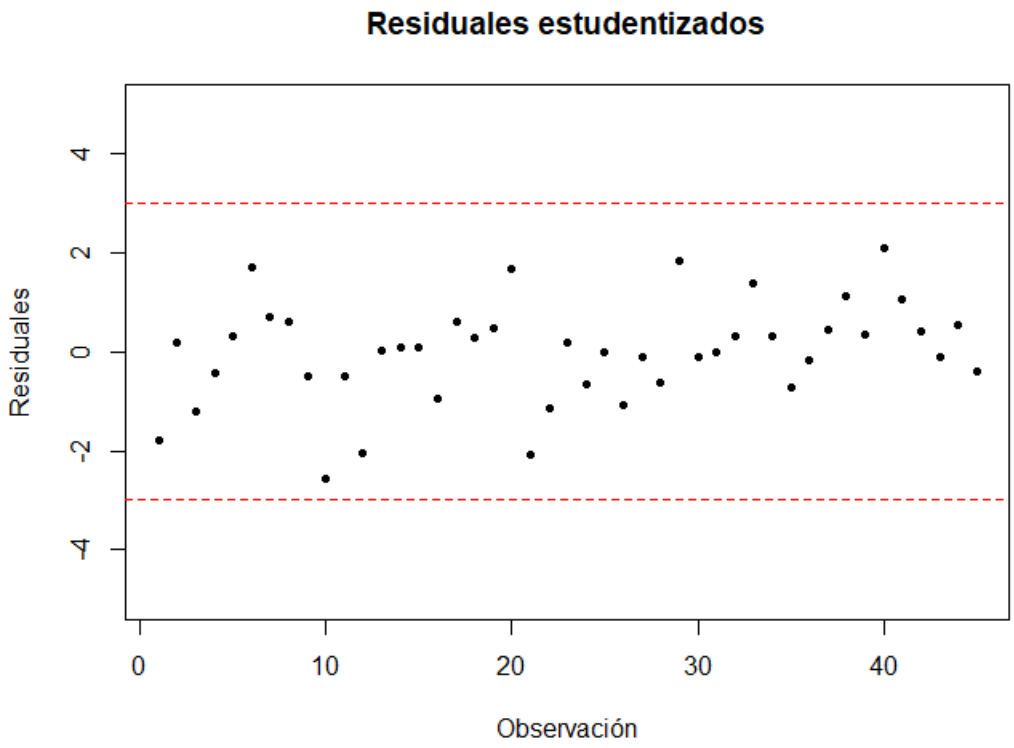

Figura 3: Identificación de outliers

Un dato se considerará atípico $|r_{estud}|_{>3}$,con base en la figura 3, se puede analizar que ninguno de los puntos se encuentra afuera de las bandas de confianza ,las cuales corresponden a los valores 3 y -3, esto índica que no hay datos atípicos.

Puntos de balanceo

Un punto se considera de balanceo si $h_{ii} > 2\frac{p}{n}$, en el caso de los datos con los que se está trabajando $h_{ii} > 0.267$.

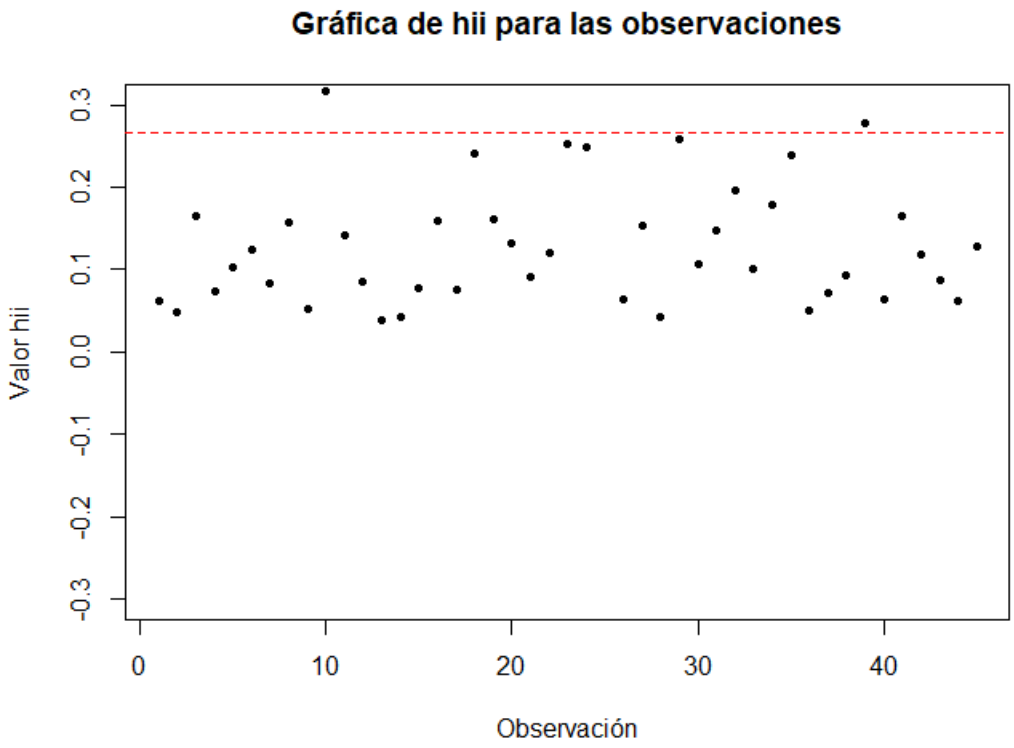

Figura 4: Identificación de puntos de balanceo

Como se puede observar en la figura 4, se traza una línea horizontal en el valor 0.267, los puntos que se encuentren por encima de esta serán considerados puntos de balanceo. Se observa que hay 2 puntos de balanceo, los cuales son los presentados en la siguiente tabla.

	h_{ii}
10	0.3173
39	0.2775

Cuadro 5: Puntos de balanceo

Puntos influenciales

Un punto influencial causa cambios importantes en la ecuación de regresión ajustada y se usarán 2 criterios para definir si los puntos son de esta naturaleza ,**Distancia de Cook y el criterio Dffits**

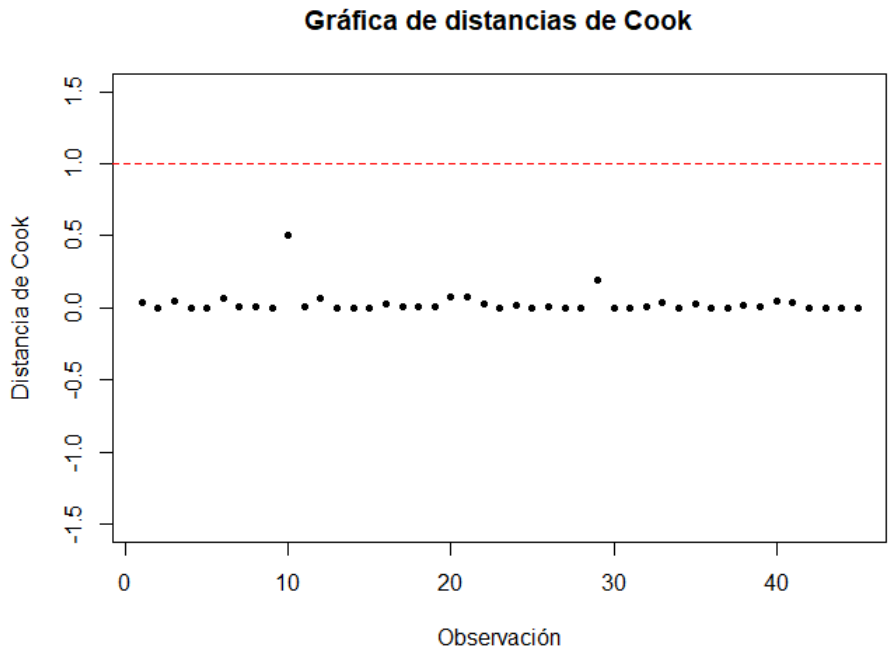

Figura 5: Criterio distancia de Cook para puntos Influenciales

Según la distancia de Cook un punto es influencial si $D_i > 1$ y de la figura 5 se concluye que ningún punto es inlfuencial bajo este criterio.

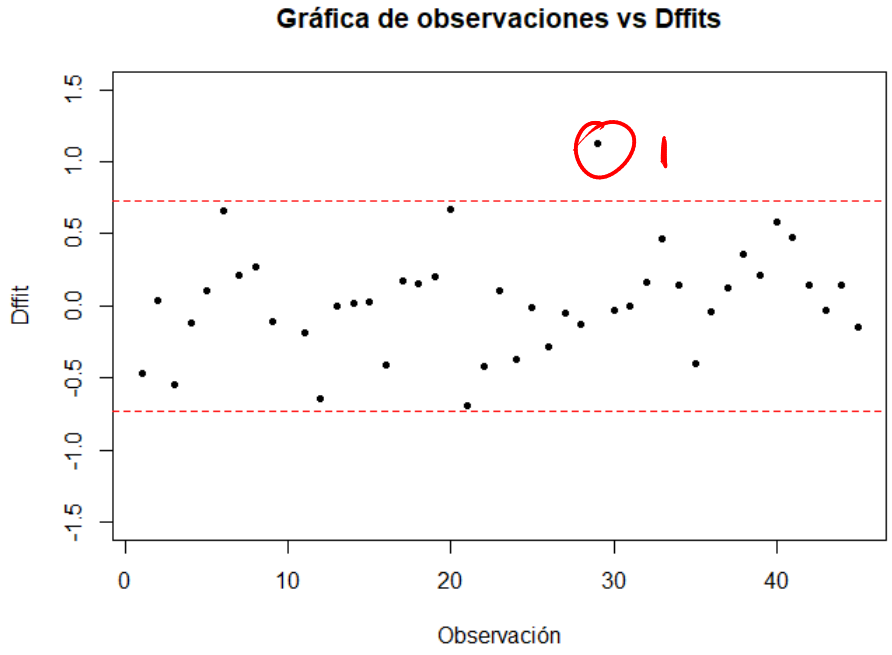

Figura 6: Criterio Dffits para puntos Influenciales

Se utilizará el criterio Dffits con el objetivo de confirmar o descartar lo analizado visualmente en

la Figura 5;para que un punto sea considerado influencial debe cumplir con que $|DFITTS| > 2\sqrt{\frac{p}{n}}$,para el caso de los datos con los que se está trabajando $|DFITTS| > 0.73$,de la Figura 6 se puede analizar de manera visual que hay dos puntos influenciales pues estos se ubican por fuera de las bandas de confianza del criterio Dffits.

	res.stud	Cooks.D	hii.value	Dffits
10	-2.5501	0.5038	0.3173	-1.8802
29	1.8425	0.1970	0.2538	1.1232

Cuadro 6: Puntos influenciales

Con esta tabla se confirma que bajo el criterio Dffits hay 2 puntos influenciales que se corresponden a $i = 10$ y $i = 29$, y bajo el criterio de distancias de Cook, ningún punto es influencial.

0.4.3. Conclusión

A partir de los resultados obtenidos en la validación en los supuestos de los errores, y los resultados en las observaciones extremas en las que en resumen tenemos dos observaciones de balanceo, dos observaciones influenciales y cero valores atípicos podemos concluir que el modelo es válido.