Trabajo 1

4,6

Estudiantes

José Miguel Arroyave Rojas
Katlyn Vanessa Espinosa Lara
Valentina Domínguez Herrera
Juliana Zapata Cardona

Equipo 07

Docente

Julieth Verónica Guarín Escudero

Asignatura

Estadística II

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medellín
20 de septiembre de 2023

Índice

Índice de figura

Índice de cuadros

1. Pregunta 1

Teniendo en cuenta la base de datos Equipo 7, 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

Por lo tanto, se plantea el modelo de regresión lineal múltiple:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i; \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2); \ 1 \leq i \leq 74$$

1.1. Modelo de regresión

Al ajustar el modelo, se estiman los siguientes coeficientes:

Valor de los parámetros estimados

	Valor de parámetro
β_0	-0.14919
β_1	0.25323
β_2	0.01119
β_3	0.03927
β_4	0.00919
β_5	0.00061

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.14919 + 0.25323 X_{1i} - 0.01119 X_{2i} + 0.03927 X_{3i} + 0.00919 X_{4i} + 0.00061 X_{5i}$$

1.2. Significancia de la regresión

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

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

H_a : Algún β_j distinto de 0 para j=1, 2, ... ,5

Donde el estadístico de prueba corresponde a:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,68} \qquad (1)$$

Teniendo la información anterior, se presenta la tabla Anova correspondiente a este modelo:

Tabla ANOVA para el modelo

	Sumas de cuadrados	Grados de libertad	Cuadrado medio *F*	Estadí F	Valor-p
Regresión	53.9171	5	10.78343	10.82803	1.10194e-07
Error	67.7180	68	0.995853		

Como Valor-p < 0.05, se rechaza H_0 concluyendo que el modelo de RLM propuesto essignificativo. Esto quiere decir que, el riesgo de infección es afectado significativamente por almenos una de las predictoras consideradas.

1.3. Significancia de los parámetros

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

$$H_0 : \beta_j = 0 \text{ para j } = 0,1,2,\ldots 5$$
$$H_a : \beta_j \neq 0$$

Donde el estadístico de prueba corresponde a:

$$T_{i.0} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \overset{H_0}{\sim} t_{68}$$

Tabla de parámetros estimados

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	valor-P
β_0	-0.1491	1.6327	-0.0913	0.9274
β_1	0.2532	0.0890	2.8438	0.0058
β_2	0.0111	0.3111	0.3599	0.7199
β_3	0.0392	0.0144	2.7165	0.0083
β_4	0.0091	0.0068	1.3385	0.1851
β_5	0.0006	0.0006	0.9065	0.3678

De la tabla de parámetros estimados, a un nivel de significancia $\alpha = 0.05$, se concluye que los parámetros individuales β_1 y β_3 son significativos debido al valor-p que se evidenció.

Por otro lado, se encuentra que $\beta_0, \beta_2, \beta_4$ y β_5 son individualmente no significativos para el modelo.

Interpretación de los parámetros

$\hat{\beta}_1 =$ Por cada día en el aumento de la duración de la estadía de todos los pacientes en el hospital, aumenta en un 25.32% la probabilidad promedio de que el paciente adquiera una infección en el hospital, cuando las demás variables se mantienen constantes.

$\hat{\beta}_3 =$ Por cada cama que aumenta en el hospital, aumenta un 3.92% la probabilidad promedio de que el paciente adquiera una infección en el hospital, cuando las demás variables se mantienen constantes.

1.4. Coeficiente de determinación múltiple R^2

El Para hallar el coeficiente de determinación múltiple R^2 empleamos el SSR y SSE dados en la tabla Anova

$$R^2 = \frac{SSR}{SST} \rightarrow \frac{53.9171}{53.9171 + 67.7180} = 0.443269$$

El modelo tiene un coeficiente de determinación múltiple R2 = 0.4433, lo que significa que aproximadamente el 44.33 % de la variabilidad total observada en el riesgo de infección que es explicada por el modelo de regresión propuesto en el presente informe.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con el P-valor más pequeños en el modelo fueron X_1, X_3, X_4, por lo tanto, a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$H_0 : \beta_1 = \beta_3 = \beta_4 = 0$

H_1 : Algún β_j distinto de 0 para j= 1,3,4

Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	67.718	X1 X2 X3 X4 X5
Modelo reducido	106.876	X2 X5

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

$$Y_i = \beta_0 + \beta_2 X_{2i} + \beta_5 X_{5i} + \varepsilon_i ; \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2); \ 1 \leq i \leq 74$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$F0 = \frac{(SSE(\beta 0,\beta 2,\beta 5) - SSE(\beta 0,\ldots,\beta 5))/3}{MSE(\beta 0,\ldots,\beta 5)} \quad \text{Ho} \sim f_{3,68}$$

$$F0 = \frac{(106.876 - 67.718)/3}{0.995853} \quad \text{Ho} \sim f_{3,68}$$

$$F0 = 13{,}1070215$$

Comparando el F_0 obtenido con el cuantil $f_{0.95,3,68} = 2{,}735541$, se puede ver que $F_0 > f_{0.95,3,68}$ esta en la región de aceptación. por ende, se tiene suficiente evidencia estadística para rechazar la hipótesis nula H_0, por lo que se concluye que el subconjunto es significativo, es decir, depende de al menos una de las variables asociadas al modelo.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se desea analizar si el efecto de la duración de la estadía de los pacientes en el hospital es igual a el efecto de la rutina de cultivos realizados en pacientes sin síntomas de infección hospitalaria, por cada 100 pacientes. Asimismo, queremos investigar si el efecto medio de la disponibilidad de camas en el hospital durante el período del estudio es igual al efecto del promedio diario de pacientes en el hospital durante dicho período.

Para responder a la pregunta se plantea la siguiente prueba de hipótesis:

$$\mathbf{H}_0: \beta_1 = \beta_2; \beta_3 = \beta_4$$
$$\mathbf{H}_1: \beta_1 \neq \beta_2 \text{ ó } \beta_3 \neq \beta_4$$

Reescribiendo matricialmente:

$$H_0 : \mathbf{L}\beta = \mathbf{0}$$
$$H_1: \mathbf{L}\beta \neq \mathbf{0}$$

L esta dada por la siguiente matriz

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

Quedando como modelo reducido:

$$Y = \beta_0 + \beta_1(X_1 + X_2) + \beta_3(X_3 + X_4) + \beta_5 X_5 + \varepsilon_i \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$
$$= \beta_0 + \beta_1 X_{1,2} + \beta_3 X_{3,4} + \beta_5 X_5$$

Donde $X_{1,2} = X_1 + X_2$ y $X_{3,4} = X_3 + X_4$

3.2. Estadístico de prueba

El estadístico de prueba F0 está dado por:

$$F_0 = \frac{SSH/_{gl.ssh}}{MSE(MF)} = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)}$$

$$= \frac{\frac{SSE(MR) - 67.718}{2}}{0.995853} \quad \text{Ho} \sim f_{2,68}$$

Si $F_0 > f_{0.05,2,68,}$ entonces se rechaza la hipótesis nula y al menos uno de los supuestos no se cumple con una significancia del 95%

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

En esta parte se revisarán dos criterios, uno es la siguiente prueba de hipótesis de normalidad, la cual se sustentará con el grafico de Shapiro-Wilk y la otra es la prueba grafica cuantil-cuantil:

$H_0 : \varepsilon_i \sim$ Normal
$H_1 : \varepsilon_i \neq$ Normal

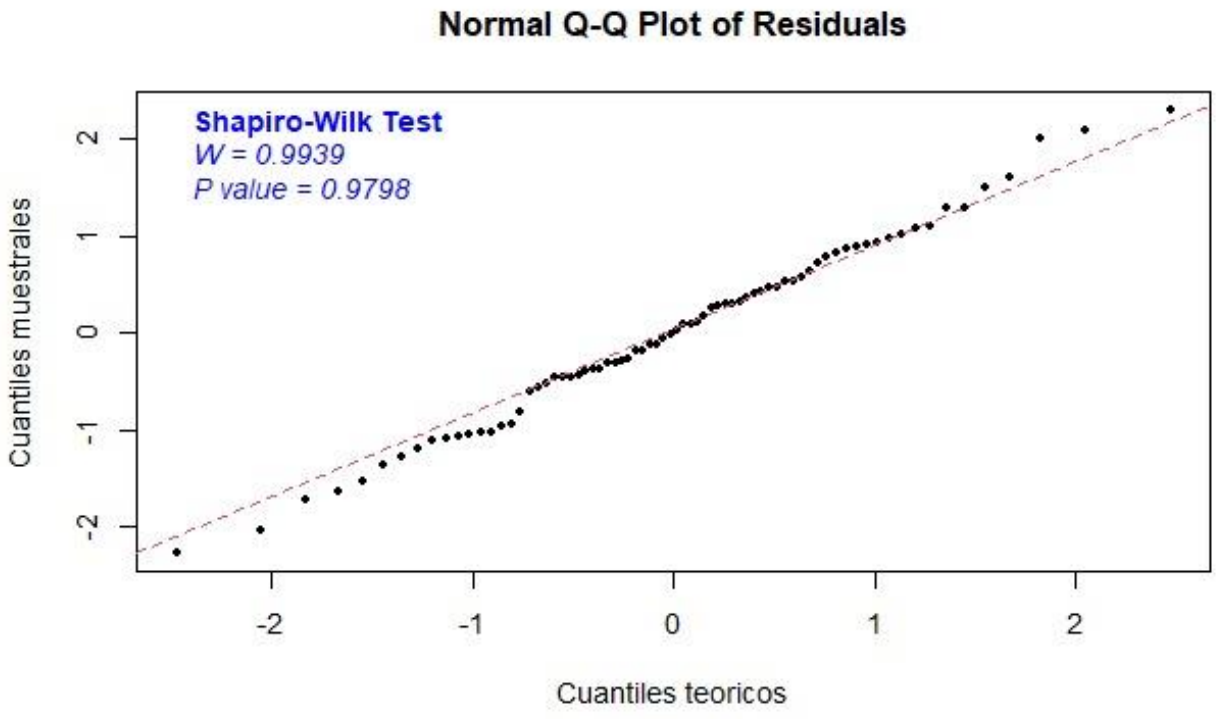

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Se puede notar que el patrón de puntos no sigue completamente la línea roja que representa el ajuste de la distribución de los residuales a una distribución normal, presentando partes con patrones irregulares, sin embargo estas observaciones presentes no generan demasiada alteración para el modelo y en consideraciones relevantes se obtiene un valor-P mayor a 0.05 y su valor es significativo, ya que es muy grande y en su mayoría los puntos están sobre la línea roja o muy cercanos a ella, por lo que debido a esto se concluye que, el supuesto de normalidad se cumple.

4.1.2. Varianza constante

Residuales Estudentizados vs Valores Ajustados

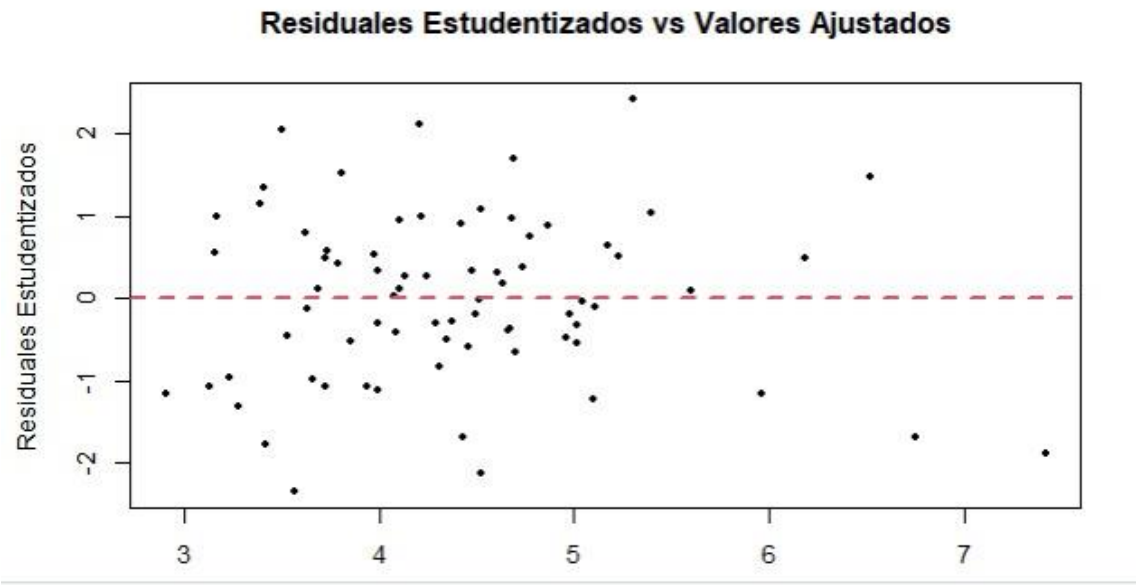

Figura 2: Gráfico residuales estudentizados vs valores ajustados

3pt

Del análisis de la gráfica realizando un trazo de líneas por encima y debajo de los puntos, se tiene que el patrón de los puntos indica un aumento de la dispersión hasta el centro de la gráfica y después hay un leve decrecimiento de la dispersión, sin embargo, la evidencia en estos no es tan fuerte, es decir no se observan patrones demasiado marcados que generen cambios bruscos, por lo que nos lleva concluir que el supuesto de varianza constante se cumple.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

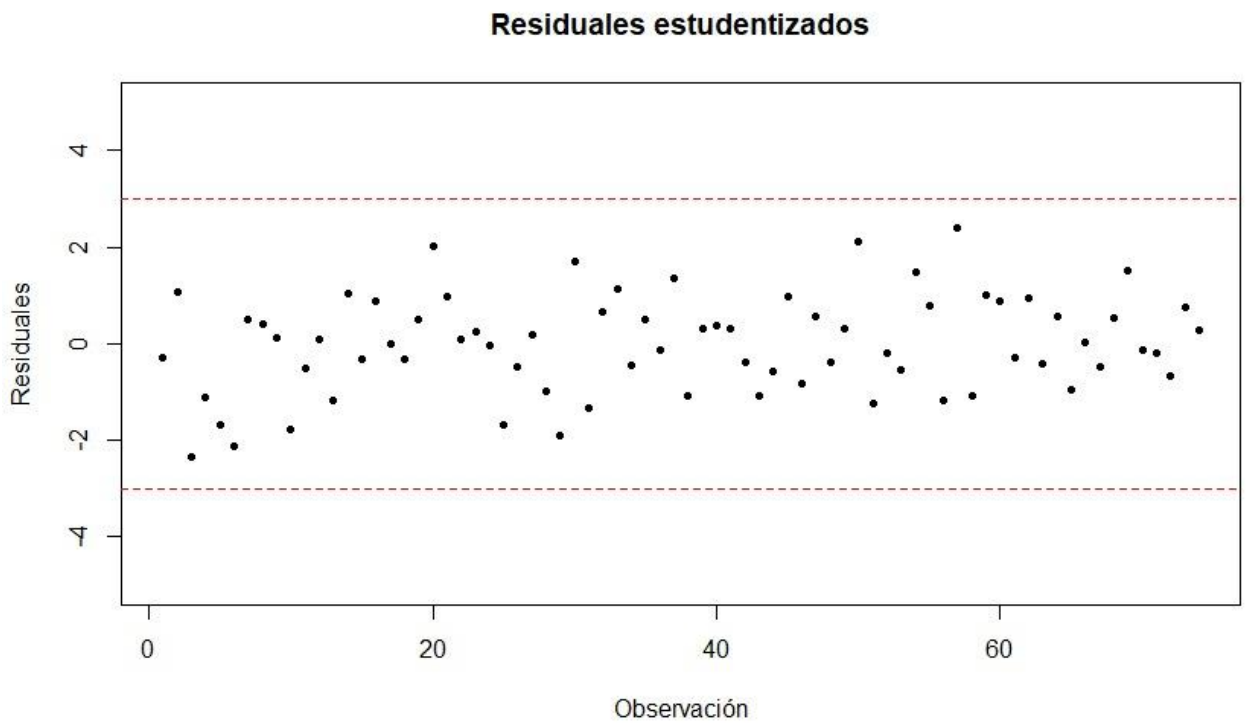

Figura 3: Identificación de datos atípicos

Al analizar la gráfica anterior podemos ver que ninguno de los residuales estudentizados cumple con el criterio de $|r_{estud}| > 3$, por lo tanto, podemos decir que no se registran datos atípicos en la muestra estudiada.

4.2.2. Puntos de balanceo

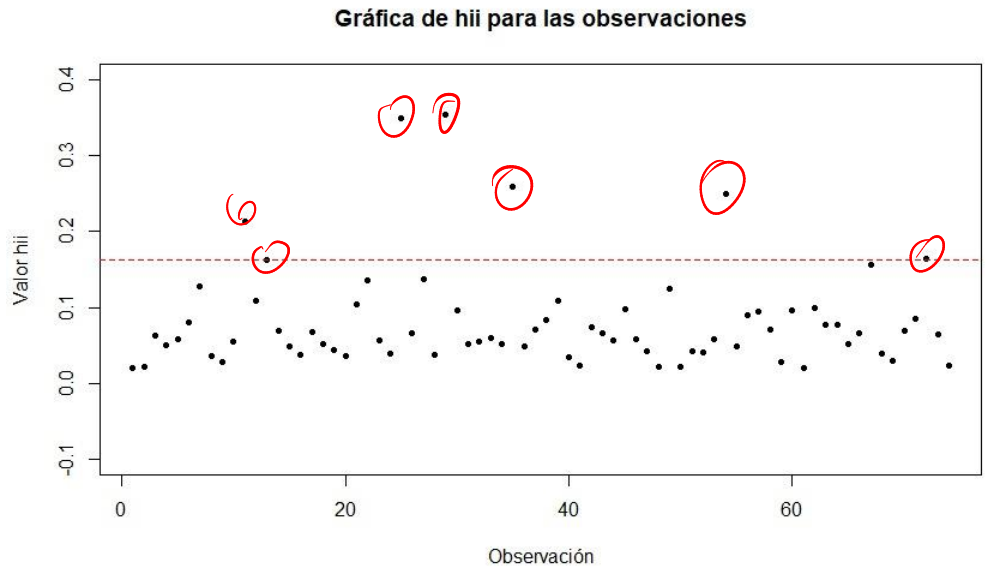

Figura 4: Identificación de puntos de balance

Cuadro 5: Tabla de puntos de balanceo

	res.stud	Cooks.D	hii.value	Dffits
11	-0.5177	0.0121	0.2135	-0.2683
13	-1.1646	0.0441	0.1632	-0.5156
25	-1.6788	0.2512	0.3484	-1.2446
29	-1.8911	0.3260	0.3536	-1.4264
35	0.4877	0.0138	0.2586	0.2864
54	1.4798	0.1209	0.2488	0.8594
72	-0.6588	0.0142	0.1639	-0.2904

Al analizar la gráfica de observaciones vs valores h_{ii} se puede apreciar que hay seis observaciones que están cumpliendo con el criterio definido en los puntos de balanceo, el cual es $h_{ii} > 2\frac{p}{n}$

Donde $h_{ii} > 2\frac{6}{76} = 0{,}158$, y al revisar en la tabla podemos ver que dichas observaciones son, la observación número once con un valor de $h_{ii} = 0.2135$, la observación número trece con un valor de $h_{ii} = 0.1632$, la observación número veinticinco con un valor de $h_{ii} = 0.3484$, la observación numero veintinueve con un valor de $h_{ii} = 0.3536$, la observación numero treinta y cinco con un valor de $h_{ii} = 0.2586$, la observación numero cincuenta y cuatro con un valor de $h_{ii} = 0.2488$ y la observación numero setenta y dos con un valor de $h_{ii} = 0.1639$, es de suma importancia identificar estos puntos porque pueden afectar las estadísticas resumen del modelo como el R^2 y los errores estándar de los coeficientes estimados, lo cual se puede traducir en una mala interpretación del modelo y de la significancia de los parámetros estimados.

4.2.2. Puntos influenciales

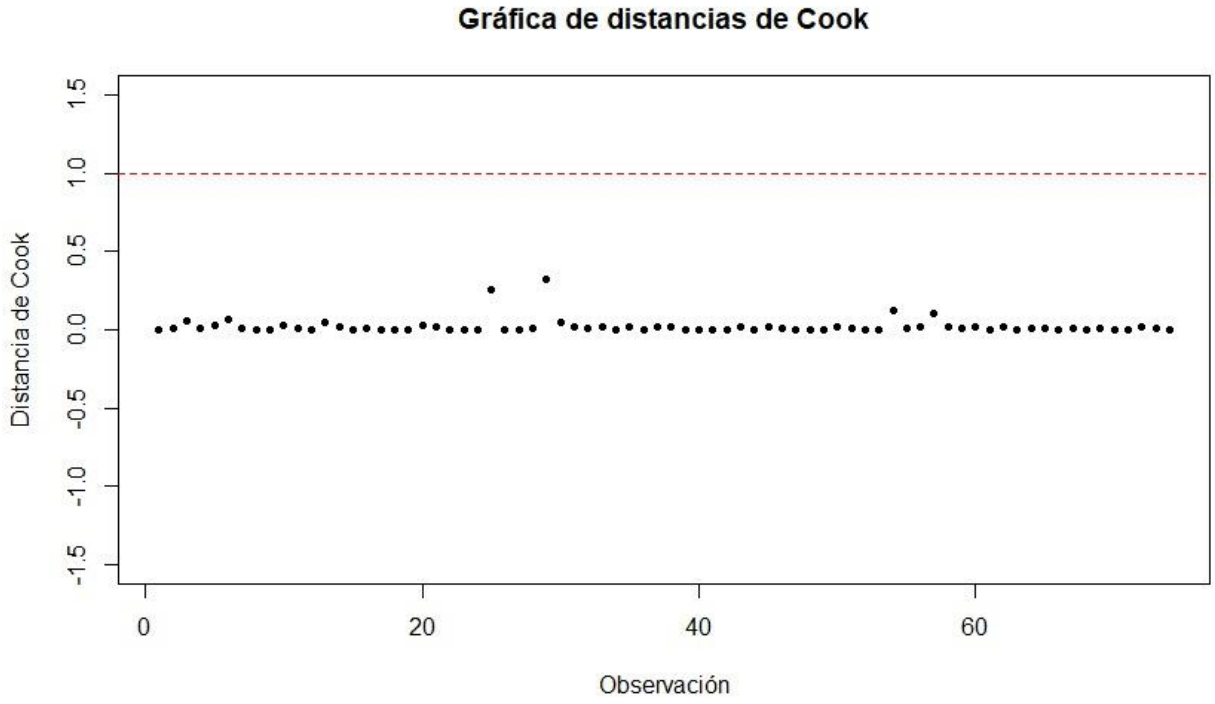

Figura 5: Criterio distancias de Cook para puntos influenciales

Realizando la valoración de puntos influénciales con la distancia de cook se obtiene que ningunade las observaciones cumple con el criterio de $D_i > 1$, es decir que en cada una de las observaciones no se obtuvo una gran diferencia de los estimadores por mínimos cuadrados sin incluir esa i-ésima observación, o visto de otra forma la influencia del punto sobre el vectorde parámetros no es suficiente para considerarlo un punto influencial.

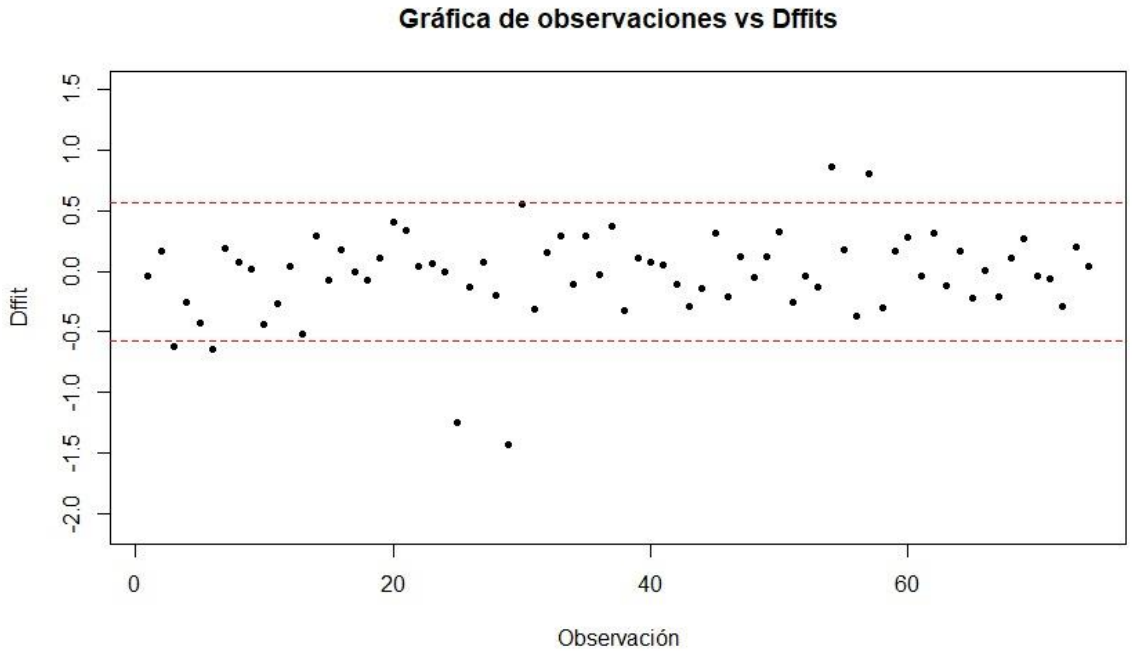

Figura 6: Criterio Dffits para puntos influenciales

Cuadro 6: Tabla de puntos influenciales

	res.stud	Cooks.D	hii.value	Dffits
3	-2.3441	0.0606	0.0621	-0.6244
6	-2.1116	0.0644	0.0798	-0.6384
25	-1.6788	0.2512	0.3484	-1.2446
29	-1.8911	0.3260	0.3536	-1.4264
54	1.4798	0.1209	0.2488	0.8594
57	2.4151	0.1005	0.0937	0.8061

Podemos ver que las observaciones 3,6,25,29,54 y 57 cumplen con el criterio de la prueba Dffits

$|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, donde $2\sqrt{\frac{6}{76}} = 0.5619$

lo que significa que dichas observaciones son influénciales, esta vez se usa el vector con los valores ajustados, afectando las estimaciones de $\hat{\boldsymbol{Y}}_{,}$ para ver si hay una diferencia significativa al no incluir la observación, para este caso se obtiene por la evaluación Dffits, en la cual si hay diferencias importantes en el modelo al no incluir los puntos enunciados por la prueba. Estos puntos influénciales tienen un impacto muy importante en el modelo ya que lo halan en su dirección , haciendo que no sea el que mejor se ajuste a la mayoría de datos proporcionados.

4.3. Conclusión

En cuanto a la validez de la regresión, se puede concluir que el modelo es adecuado para comparar las variables predictoras y respuesta, aunque hay algunos puntos influénciales y muchos de balanceo que pueden estar interfiriendo en el ajuste del modelo, es importante identificar estos puntos porque se pueden traducir en una mala interpretación del modelo y de la significancia de los parámetros estimados. Estas observaciones perturban el modelo al tirarlo en direcciones que no pueden acomodar la mayoría de los datos, lo que a su vez resulta en estimaciones de la variable de respuesta que estén alejadas de los valores reales o esperados. Tal vez aplicando una transformación al modelo actual se pueda convertir en uno apropiado para hacer dichas estimaciones.