Trabajo 1

2,8

Estudiantes

Jhon Jairo Parra Amaya
Angie Pahola Tobar Calpa
Maria Camila Diosa Oliveros

Equipo

Docente

Carlos Mario Lopera

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:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

Las variables

- Y: Probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje).
- X_1:Duración promedio de la estadía de todos los pacientes en el hospital (en días).
- X_2:Razón del número de cultivos realizados en pacientes sin síntomas de infección hospitalaria, por cada 100.
- X_3;Número promedio de camas en el hospital durante el periodo del estudio.
- X_4;Número promedio de pacientes en el hospital por día durante el periodo del estudio.
- X_5;Número promedio de enfermeras, equivalentes a tiempo completo,durante el periodo del estudio.

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	0.1456
β_1	0.1160
β_2	0.0072
β_3	0.0487
β_4	0.0162
β_5	0.0021

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 0.1456 + 0.116 X_{1i} + 0.0072 X_{2i} + 0.0487 X_{3i} + 0.0162 X_{4i} + 0.0021 X_{5i} + \varepsilon_i,\ 1 \leqslant i \leqslant 69$$

1.2. Significancia de la regresión

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

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2,..., 5} \end{cases}$$

Cuyo estadístico de prueba es:

$$F_0 = \frac{MST}{MSE} \overset{H_0}{\sim} f_{5,63} \tag{1}$$

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	62.1222	5	12.424435	14.9975	1.10369e-09
Error	52.1912	63	0.828431		

De la tabla Anova, se define el P-valor $< \alpha$ 1.10369e-09, osea muy cercano a cero por lo que 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, sin embargo esto no garantiza, por el momento, que la regresion sea significativa.

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	0.1456	1.4299	0.1018	0.9192
β_1	0.1160	0.0938	1.2363	0.2210
β_2	0.0072	0.0283	0.2537	0.8006
β_3	0.0487	0.0124	3.9136	0.0002
β_4	0.0162	0.0068	2.3852	0.0201
β_5	0.0021	0.0008	2.5070	0.0148

Con un nivel de significancia $\alpha = 0.05$, los parámetros β_3, β_4 y β_5 son significativos, pues sus P-valores son menores a α por lo tanto un valor de $t_n - p$ mayor que $t_0 j$ es evidencia en contra la hipotesis nula fuerte.

1.4. Interpretación de los parámetros

$\hat{\beta}_3$:Por cada unidad de aumento de camas, la probabilidad promedio de riesgo de infección aumentará en 0.0487 unidades cuando las demas predictoras se mantienen constantes.

$\hat{\beta}_4$**:Por cada unidad de aumento en el censo diario, la probabilidad promedio de riesgo de infección aumentará en 0.0162 unidades cuando las demas predictoras se mantienen constantes.**

$\hat{\beta}_5$**:Por cada unidad de aumento en el número de enfermeras, la probabilidad promedio de riesgo de infección aumentará en 0.0021 unidades cuando las demas predictoras se mantienen constantes.**

En este caso en particular β_0 no tiene interpretación, puesto que el cero no está dentro del rango de las variables en estudio.

1.5. Coeficiente de determinación múltiple R^2

con $R^2 = 0.5434 = SSR/SST$, por lo tanto el modelo de RLM planteado anteriormente, explica el 53.12 % de la variabilidad total de la variable respuesta, lo que sugiere que no es el mejor modelo para explicar el riesgo de infección, una posible solución seria el análisis de observaciones extremas, información redundante o la ampliación de la base de datos.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más alto en el modelo fueron X_1, X_2, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 2. \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	52.191	X1 X2 X3 X4 X5
Modelo reducido	79.700	X1 X2

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

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$
\begin{aligned}
F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_2) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,63} \\
&= \frac{(79.700 - 52.191/3)}{0.828431} \\
&= 11.061
\end{aligned} \tag{2}
$$

Ahora, comparando el cuartil F_0 con $f_{0.95,3,6363} = 2.7505$, se puede ver que $F_0 < f_{0.95,1,45}$ esta esta en la región de aceptación, por ende, no se tiene suficiente evidencia estadística para rechazar la hipótesis nula, por lo que se concluye que el subconjunto no es significativo, siendo posible descartar del modelo las variables del subconjunto.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

$$
\begin{cases}
\mathrm{H}_0 : \beta_2 = \beta_4;\ \beta_1 = \beta_0 \\
\mathrm{H}_1 : \text{Alguna de las igualdades no se cumple}
\end{cases}
$$

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

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

El modelo reducido está dado por:

$$
Y_i = \beta_o + \beta_2 X^*_{X2i+X4i} + \beta_3 X^*_{3i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69
$$

Donde $X^*_{2i} = X_{2i} + X_{4i}$ y $X^*_{4i} = 0$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$
F_0 = \frac{(SSE(MR) - 52.1912)/2}{0.828431} \overset{H_0}{\sim} f_{100,63} \tag{3}
$$

4. Pregunta 4

12pt

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, acompañada de un gráfico cuantil-cuantil:

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

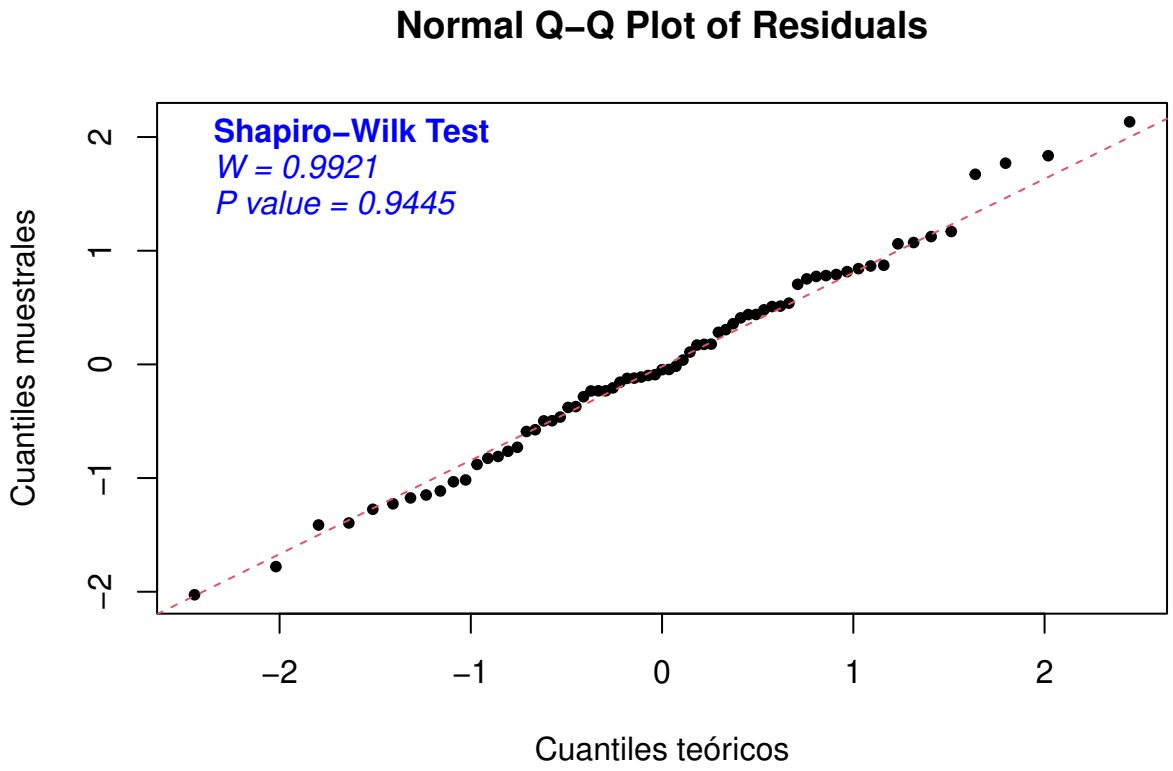

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

2pt

Al ser el P-valor aproximadamente igual a 0.9445 y teniendo en cuenta que el nivel de significancia $\alpha = 0.05$, el P-valor es mucho mayor y por lo tanto, se no se rechazaría la hipótesis nula, Ahora se validará si la varianza cumple con el supuesto de ser constante en la siguiente subseccion, por ultimo pero no menos importante, como el patron de los residuales no sigue la linea roja que representa el ajuste de la distribucion de los residuales a una normal, es probable que sea por observaciones influenciables por lo que es un motivo mas fuerte para rechazar H_0 que el P-valor.

→ No! se parece demasiado a una normal

4.1.2. Varianza constante

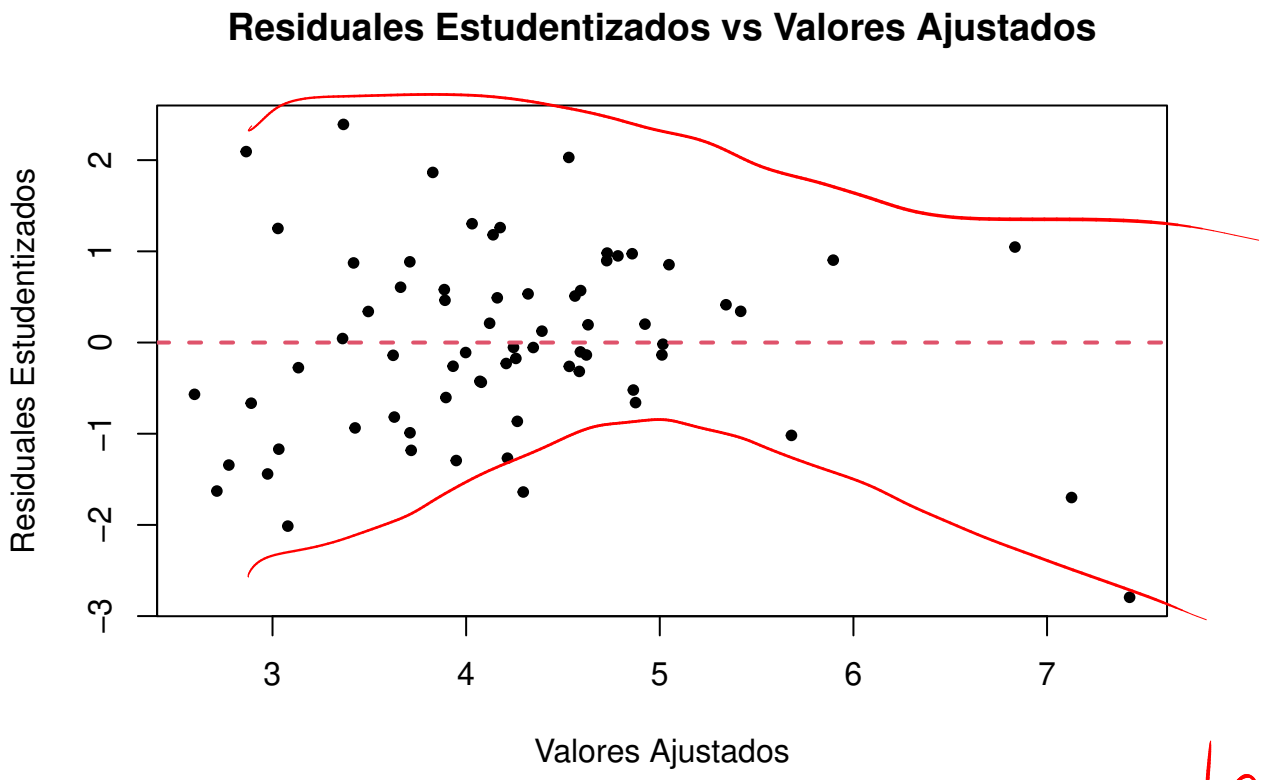

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que hay patrones en los que la varianza aumenta, decrece y tiene un comportamiento que permite descartar una varianza no constante, al haber evidencia suficiente en contra de este supuesto se rechaza. Además es posible observar media 0.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

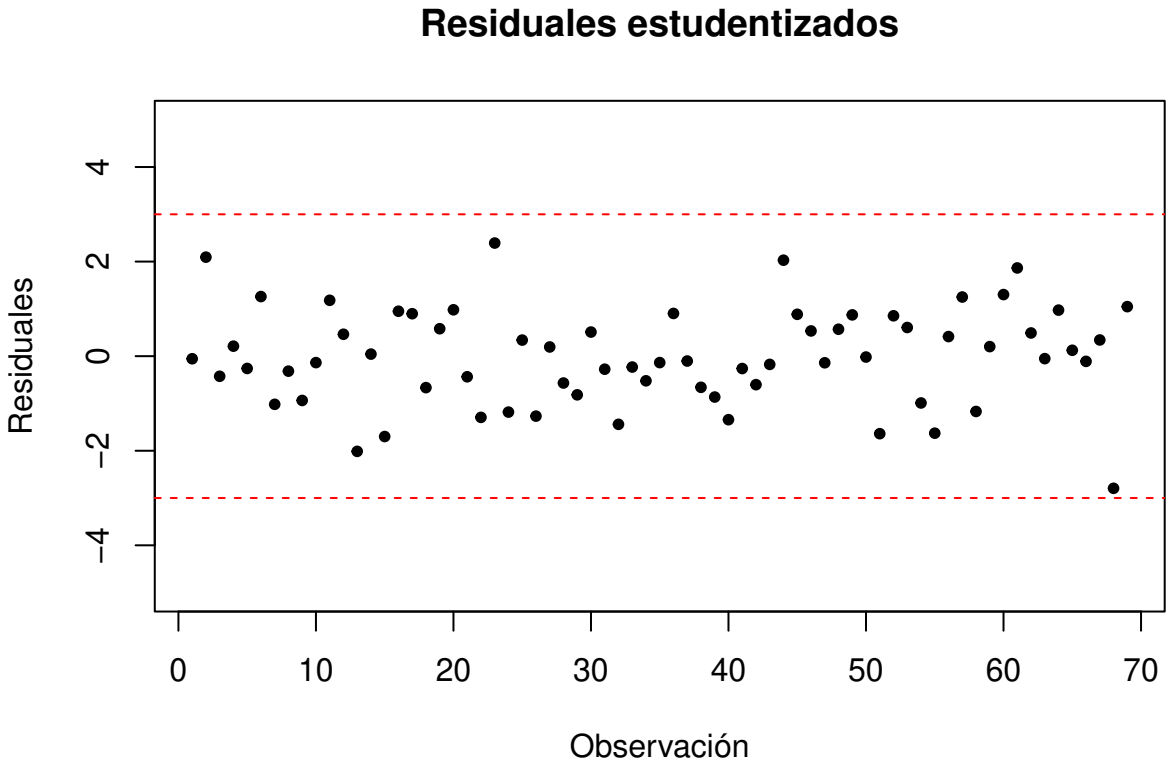

Figura 3: Identificación de datos atípicos

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 $|r_{estud}| > 3$.

4.2.2. Puntos de balanceo

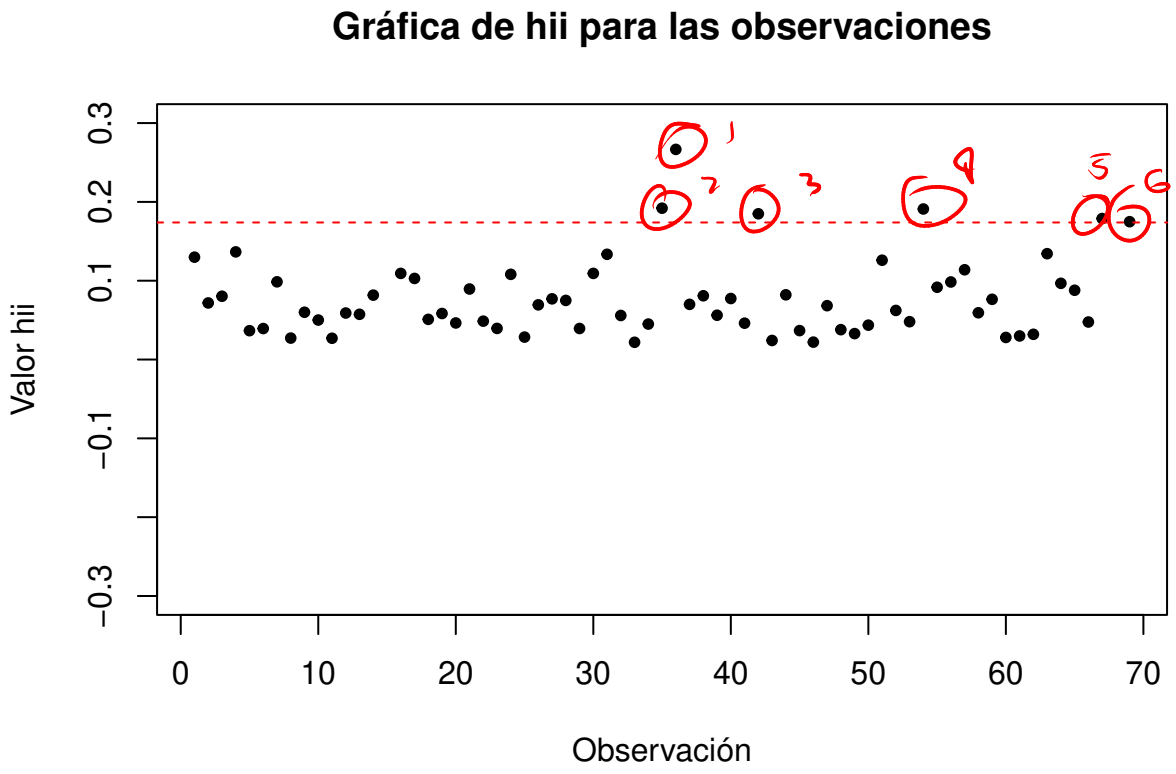

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 15  -1.6999  0.2849    0.3717 -1.3278
## 35  -0.1360  0.0007    0.1920 -0.0658
## 36   0.9035  0.0495    0.2666  0.5439
## 42  -0.6031  0.0138    0.1849 -0.2858
## 54  -0.9897  0.0385    0.1910 -0.4808
## 67   0.3421  0.0042    0.1789  0.1585
## 68  -2.7940  0.7480    0.3650 -2.2453
## 69   1.0471  0.0387    0.1749  0.4825
```

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 $h_{ii} = 2\frac{p}{n} = (0.14492)$, a continuacion presentaremos los 8 h_{ii} las obsevaciones numero 15, 35, 36, 42, 54, 67, 68, 69; Es importante identificar estos puntos porque pueden afectar las estadísticas resumen del modelo como el R2 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.

4.2.3. Puntos influenciales

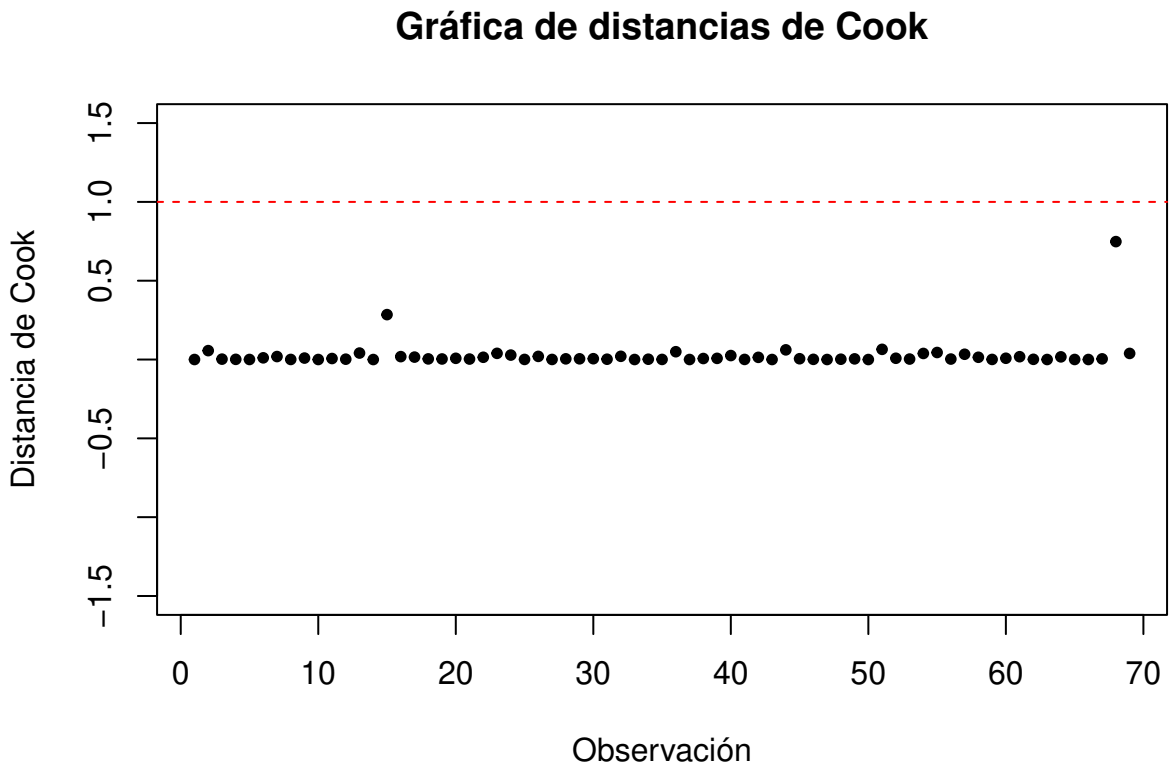

Figura 5: Criterio distancias de Cook para puntos influenciales

Al hacer la evaluación de puntos influénciales con la distancia de cook se obtiene que ninguna de las observaciones cumple con el criterio de Di > 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 vector de parámetros no es suficiente para considerarlo un punto influencial.

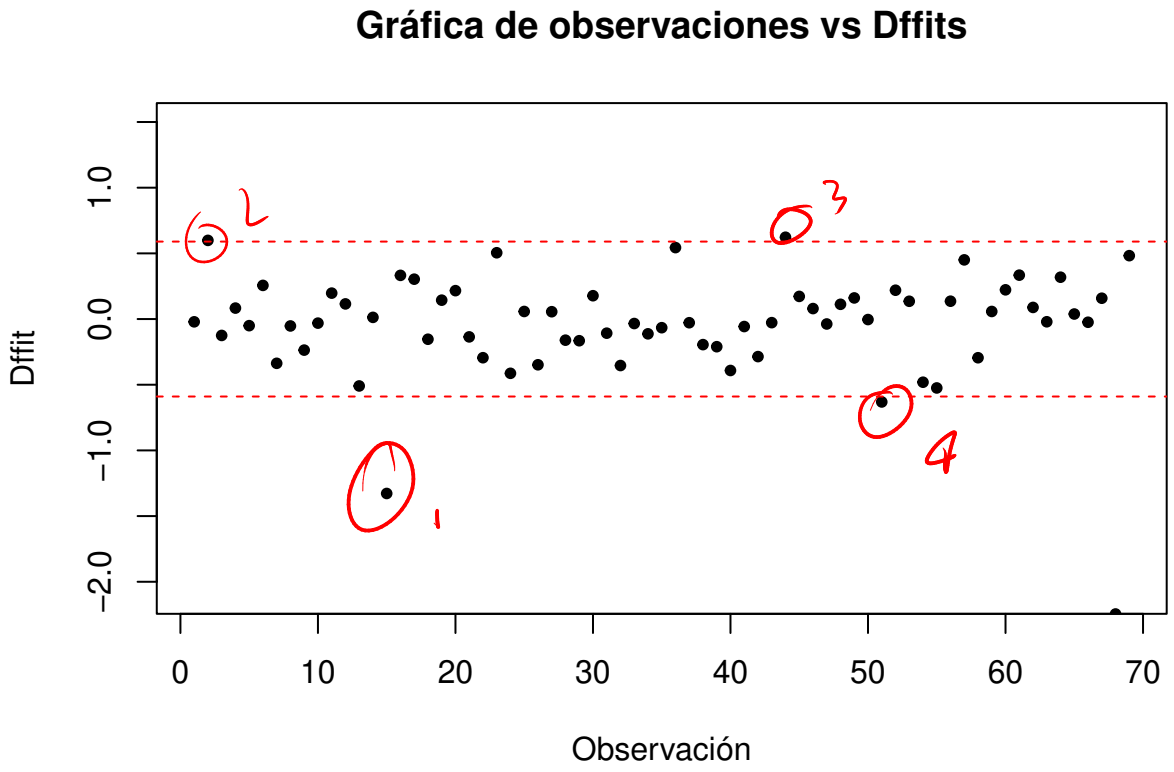

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 2    2.0936  0.0565    0.0718  0.5990
## 15  -1.6999  0.2849    0.3717 -1.3278
## 44   2.0296  0.0614    0.0821  0.6229
## 51  -1.6394  0.0646    0.1260 -0.6311
## 68  -2.7940  0.7480    0.3650 -2.2453
```

Podemos apreciar que las observaciones # están cumpliendo con el criterio de la prueba Dffits, $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 0.38$

lo que significa que dichas observaciones son influénciales, esta vez se usa el vector con los valores ajustados para ver si hay una diferencia significativa al no incluir la observación, para estos caso se obtiene gracias a la evaluación Dffits que 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 de regresión 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

Acerca de la validez de la regresión se podría concluir que el modelo no está ajustando los datos de la manera más óptima, por lo que se ve en la prueba grafica no se está cumpliendo con el supuesto de normalidad y esto de entrada hace que el ajuste pierda un importante fundamento teórico para su verificación, esto probablemente se deba a los puntos influénciales

que observamos anteriormente, estas observaciones desajustan el modelo ya que lo halan en direcciones en las cuales no se logra acomodar a la mayoría de los datos y por lo tanto da estimaciones de la variable respuesta alejadas de los valores reales o esperados.

Confunde ajuste con validez