Trabajo 1

4,7

Estudiantes

Alejandro Feria Gonzalez
Juan Fernando Quintero Perez
Stiven Julio Doval
Santiago Lopez Mejia

Equipo 23

Docente

Mateo Ochoa Medina

Asignatura

Estadística II

Sede Medellín
Octubre 5 de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

Teniendo la base de datos asignada, observamos que se tienen cinco covariables regresoras por lo que para establecer un modelo de regresión lineal múltiple debemos formular un modelo teórico con sus respectivos supuestos así:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i,\ 1 \leqslant i \leqslant 69,$$

$$\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \forall_i = 1, 2, 3, ..., 69;$$

En este modelo cada covariable tiene un significado contemplado en las siguiente lista:

- 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

Con ayuda del software R y teniendo en cuenta la base de datos dada, obtenemos la siguiente información acerca de los parámetros que nos ayudarán a delimitar un modelo ajustado de la regresión:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	0.2902
β_1	0.1462
β_2	-0.0004
β_3	0.0378
β_4	0.0199
β_5	0.0015

Por lo que nuestro modelo ajustado queda establecido de la siguente manera, explicando la variable Y (riesgo de infección) en términos de las variables restantes que actúan como predictorias:

$$\hat{Y}_i = 0.2902 + 0.1462 X_{i1} - 4 \times 10^{-4} X_{i2} + 0.0378 X_{i3} + 0.0199 X_{i4} + 0.0015 X_{i5},\ 1 \leqslant i \leqslant 69$$

1.2. Significancia de la regresión

Ahora, para analizar la significancia de la regresión, se plantea un juego de hipótesis que ayudará a rechazar o aceptar la idea en la que todos los parámetros son iguales a cero.

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1 : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2,..., 5} \end{cases}$$

El estadístico de la prueba queda establecido como:

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

Y los resultados arrojados por el programa computacional del análisis de varianza del modelo son:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	60.9329	5	12.186570	14.1033	2.99179e-09
Error	54.4379	63	0.864093		

La tabla arroja dos datos importantes para el desarrollo de la prueba, partiendo del hecho de que establecemos una significancia $\alpha = 0.05$, el $valor - p$ es menor al valor de la significancia α $(2.99179e - 09 < 0.05)$, y el valor del estadístico F_0 es mayor al cuantil de f $f_{1-0.05,5,63}$ $(14.1033 > 2.360684)$, por lo que rechazamos la hipótesis nula y concluimos de momento que exíste al menos un parámetro de la regresión ajustada que es distinto de cero y por tanto la regresión es significativa, es decir que la probabilidad de riesgo de infección es explicada por al menos una de las variables del modelo.

1.3. Significancia de los parámetros

Siguiendo el desarrollo del análisis de significancia, hacemos uso de la siguiente tabla que nos ofrece un poco más de información individual de los parámetros, con su error estándar y los estadísticos de prueba que nos ayudarán con una nueva prueba de hipótesis a tener un poco más de conocimiento acerca de su comportamiento estadístico:

$$\begin{cases} H_0 : \beta_j = 0, \\ H_1 : \beta_j \neq 0, \end{cases} j = 0, 1, 2, ..., 5$$

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	0.2902	1.5414	0.1882	0.8513
β_1	0.1462	0.0851	1.7183	0.0907
β_2	-0.0004	0.0294	-0.0133	0.9894
β_3	0.0378	0.0120	3.1411	0.0026
β_4	0.0199	0.0069	2.8899	0.0053
β_5	0.0015	0.0007	2.1416	0.0361

Una vez más, partiendo de una significacia establecida de $\alpha = 0.05$, la tabla nos permite concluir que solo podemos rechazar la hipótesis nula en el caso de los parámetros β_3, β_4 y β_5, lo que significa que estos tres parámetros son significativos pues sus valores-p $(0.0026, 0.0053, 0.0361)$ son todos menores a 0.05.

1.4. Interpretación de los parámetros

Sabiendo que los parámetros significativos en nuestro modelo son β_3, β_4 y β_5, podemos ahora asignarle una interpretación numérica al valor de cada parámetro:

$\hat{\beta}_3$**:** Por cada unidad que aumente el promedio de camas en el hospital (X_{i3}), habrá un aumento del promedio en la probabilidad de riesgro de infección del 0.0378, cuando las demás variables se mantienen constantes.

$\hat{\beta}_4$**:** Si el censo promedio diario (X_{i4}), incrementa en una unidad y las demás covariables permanecen fijas; habrá un incremento en el promedio del porcentaje de riesgo de infección en 0.0199 unidades.

$\hat{\beta}_5$**:** Debido a que las enfermeras son seres humanos que también pueden ayudar a esparcir un poco el virus, si aumenta en una unidad (X_{i5}) la cantidad promedio de enfermeras a tiempo completo presentes en el hospital; la probabilidad de riesgo de infección del modelo se verá aumentada 0.0015 unidades, siempre y cuando las variables predictorias restantes permanezcan constantes.

1.5. Coeficiente de determinación múltiple R^2

De la teoría conocemos que el coeficiente de determinación puede ser calculado como:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{2}$$

Que reemplazando con los valores arrojados por la tabla ANOVA en el punto 1.2 nos da:

$$R^2 = \frac{60.9329}{54.4379 + 60.932} = 0.5281 \tag{3}$$

Lo que significa que el 52.81 % de la variabilidad total de la probabilidad de riesgo de infección es explicada por el modelo de regresión lineal múltiple propuesto.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Teniendo en cuenta la tabla del punto 1.3, los coeficientes de las covariables con los valores-p más pequeños son $\beta_3, \beta_4, \beta_5$, para las correspondientes covariables X_3, X_4, X_5, por lo tanto a través de la tabla de todas las regresiones posibles procedemos a hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_3 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 3, 4, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	54.438	X1 X2 X3 X4 X5
Modelo reducido	76.391	X1 X2

Por lo que un modelo reducido para la prueba de significancia del subconjunto de parámetros es:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

2.2. Estadístico de prueba y conclusión

El estadístico de prueba usado para continuar con el procedimiento será:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_2) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,63} \\ &= \frac{7.3176667}{0.864093} \\ &= 8.4686101 \end{aligned} \tag{4}$$

Como resultado tenemos un valor de estadístico de prueba $8.4686101 > 5.6754711$ siendo este último número el valor del cuantil de la distribución f siendo distribuida con los parámetros: $f_{0.95,3,63}$, por lo que podemos rechazar la hipótesis nula y afirmar que al menos un coeficiente β_j con $j = 3, 4, 5$ es significativo en presencia de los otros.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Formulamos las siguintes preguntas, ¿Es β_1 igual a β_2? y ¿Es β_3 igual a β_4?, para responder a estas preguntas, generamos el siguiente juego de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2;\ \beta_4 = \beta_3 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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

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

Ya que las filas de la matriz L no puede reescribirse como múltiplo escalar las unas de las otras, obtenemos que el número de filas linealmente independientes es $r = 2$ y el modelo reducido de la prueba está dado por:

$$Y_i = \beta_o + \beta_1 X^*_{1,2} + \beta_3 X^*_{3,4} + \beta_5 X_{i5} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

Donde $X^*_{1,2} = X_1 + X_2$ y $X^*_{3,4} = X_3 + X_4$

3.2. Estadístico de prueba

Ahora el estadístico de prueba se especifica como:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,63} \tag{5}$$

Y con ayuda de R generamos dos nuevas columnas en la base de datos que ayudarán al calculo del $SSE(MR)$ y usaremos el $MSE(MF)$ brindado en la tabla Anova anterior para continuar:

$$F_0 = \frac{(57.9763 - 54.4379)/2}{0.864093} \overset{H_0}{\sim} f_{2,63} \tag{6}$$

Dejandonos por último el resultado del estadístico $F_0 = 3.1363$

4. Pregunta 4

17,5 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Continuando con la validación de los supuestos del modelo, rectificaremos la normalidad de los residuales por medio de una prueba de hipótesis y usando a Shapiro Wilk como método para obtener un valor-p que nos ayude con la desición:

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

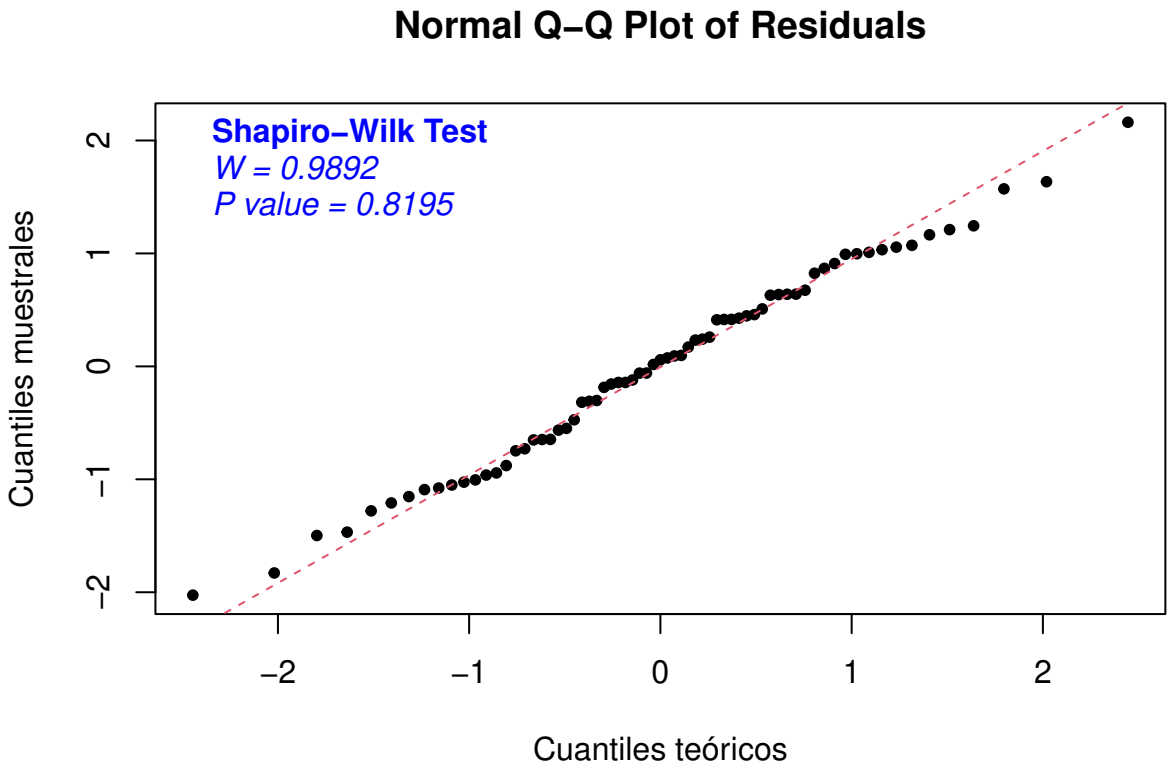

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

4 pt

Analizando gráficamente el diagrama expuesto, podemos observar que los datos están bastante alineados con la línea en el centro, pese que las colas de la gráfica parecen alejarse un poco de la línea central, son muy pocos datos los que lo hacen y no de una manera muy drástica. Adicionalmente, el valor-p que arroja el test de Shapiro es $0.8195 > 0.05$ siendo este último número el valor de la significancia que hemos decidido dar a la prueba. Por lo

tanto no rechamos la hipótesis nula y se acepta el supuesto de normalidad en los residuales del modelo.

4.1.2. Varianza constante

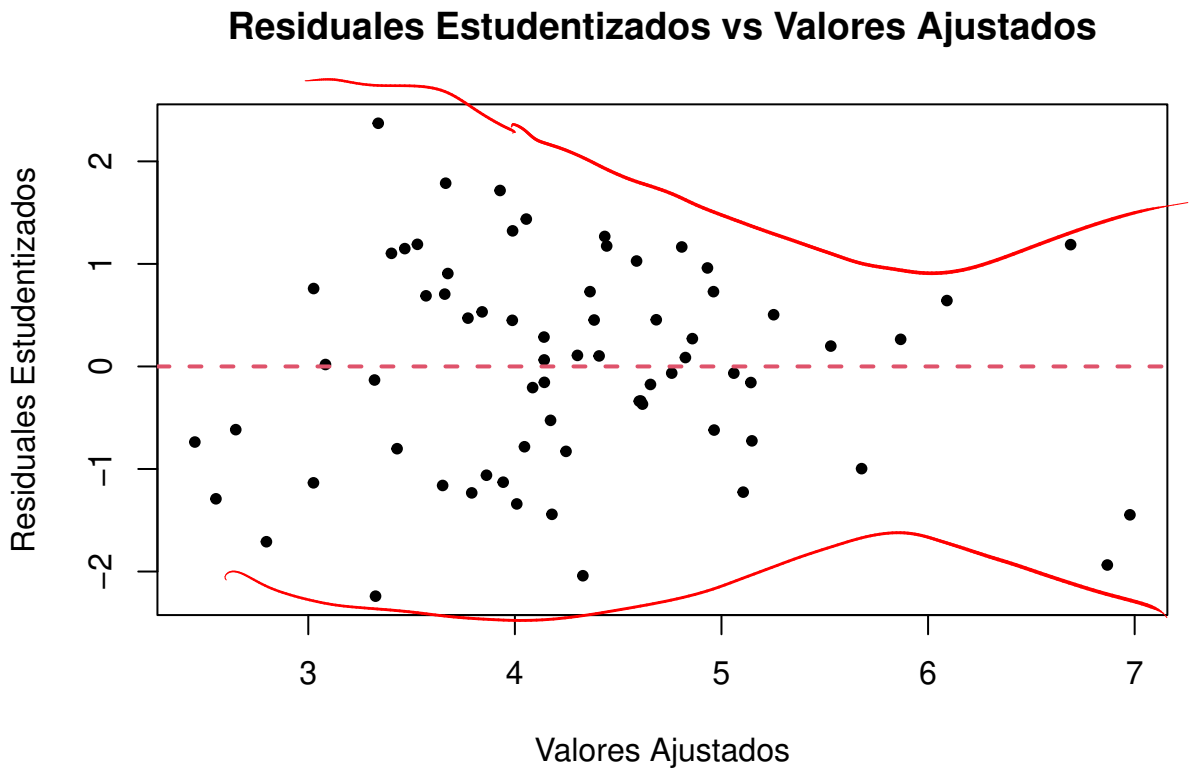

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el caso del análisis de la gráfica de residuales estudentizados, es un poco notoria la falta de constancia en la varianza pues, los puntos se encuentran dispersos al principio para luego acumularse y centralizarse a medida que se acerca a 4 y volviendo a dispersarse de 5 a 7. Esto nos da un criterio para poder rechazar gráficamente el cumplimiento de constancia de varianza en el modelo, desición que repercutirá más adelante en la toma de desiciones para la aceptación o rechazo del modelo.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

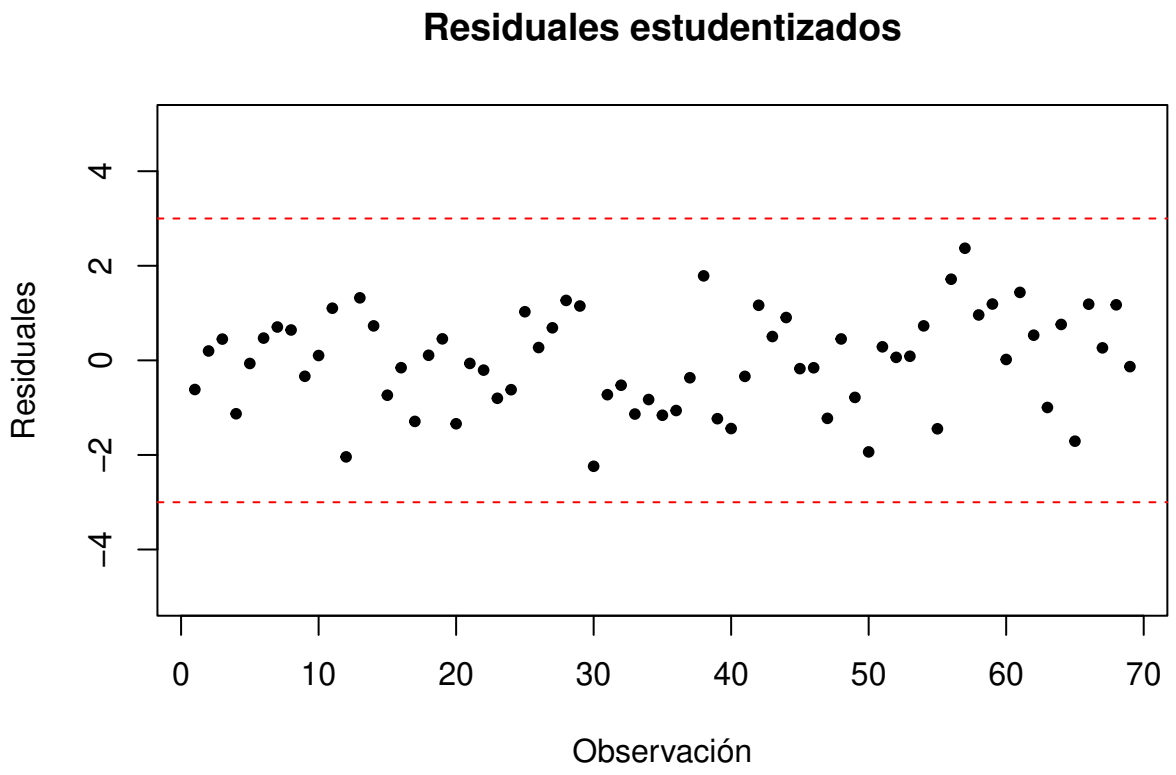

Figura 3: Identificación de datos atípicos

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

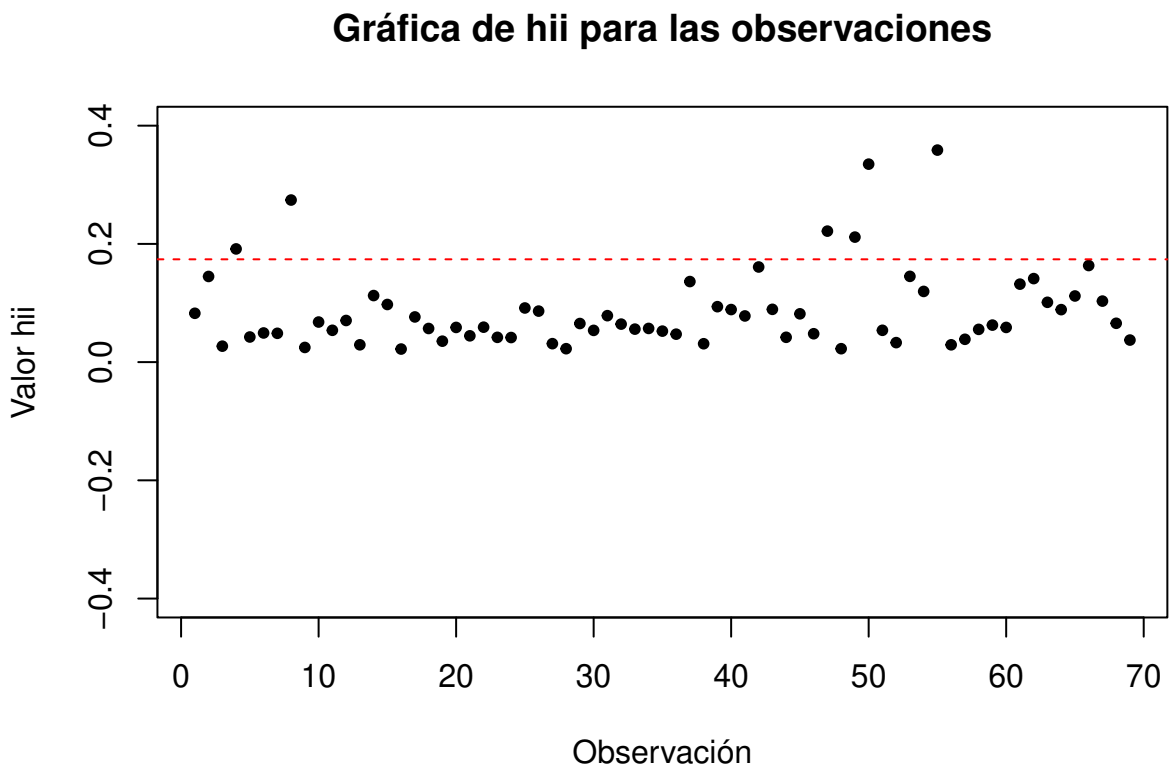

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 4   -1.1287  0.0503    0.1915 -0.5506
## 8    0.6424  0.0260    0.2742  0.3930
## 47  -1.2257  0.0713    0.2216 -0.6567
## 49  -0.7832  0.0274    0.2115 -0.4043
## 50  -1.9361  0.3146    0.3349 -1.4053
## 55  -1.4471  0.1952    0.3587 -1.0920
```

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{6}{69}$, se puede apreciar que existen 6 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{6}{69}$, los cuales son los presentados en la tabla.

4.2.3. Puntos influenciales

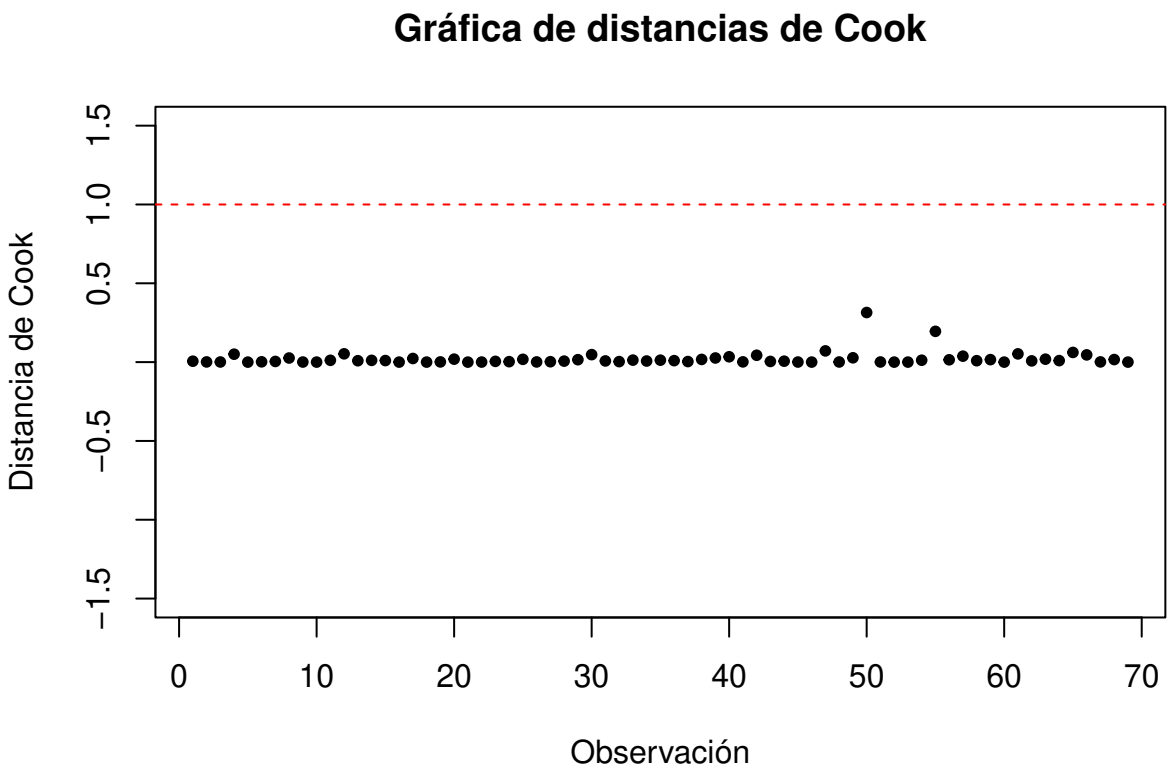

Figura 5: Criterio distancias de Cook para puntos influenciales

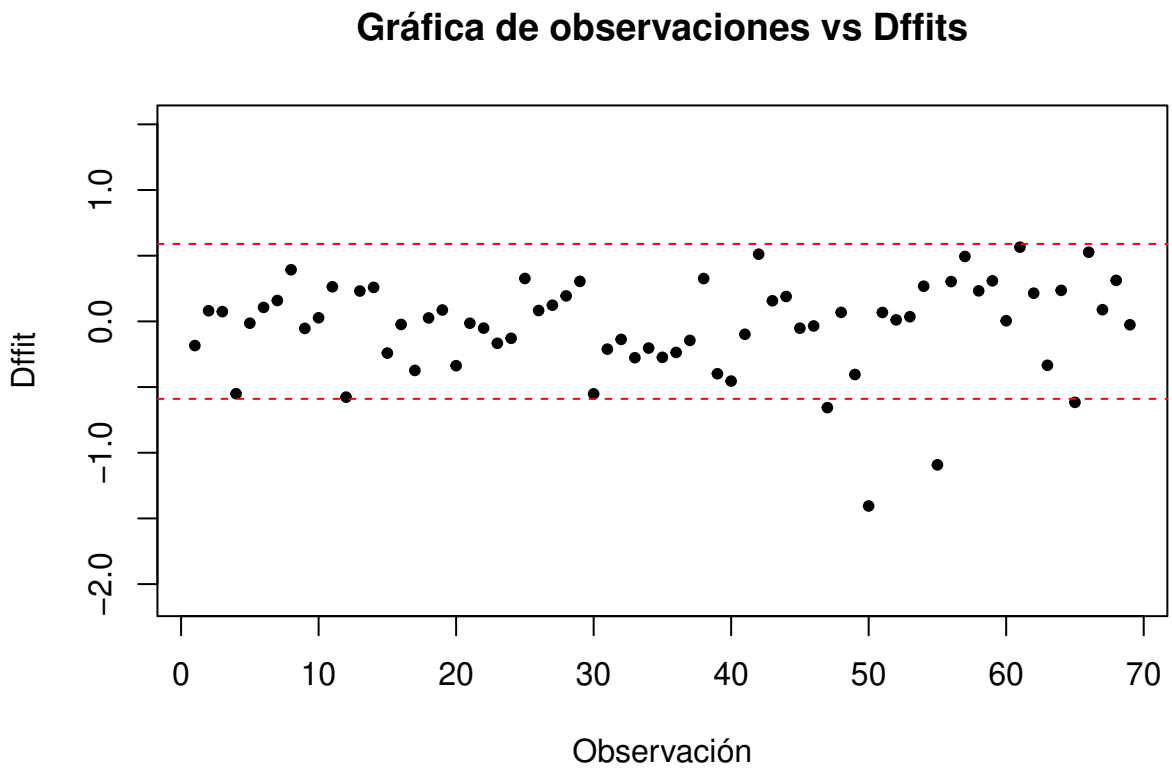

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 47  -1.2257  0.0713    0.2216 -0.6567
## 50  -1.9361  0.3146    0.3349 -1.4053
## 55  -1.4471  0.1952    0.3587 -1.0920
## 65  -1.7093  0.0614    0.1119 -0.6164
```

¿Qué cacsan? 3pt

Como se puede ver,las observaciones i = 47, 50, 55, 65 son puntos influenciales, según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{6}{69}}$, 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.

Tambien hay que notar que la observacion i=65 si bien es un punto influencial no es ni un punto de balanceo ni un valor atipico, por lo que no es punto que requiera de atención. ✗

↳ No necesariamente

4.3. Conclusión

2,5 pt

La aceptación el supuesto de normalidad en los residuales del modelo y la ausencia de valores atípicos dan indicios de la validez general del modelo. Sin embargo, β_1 y β_2 no son significativos, tambien la inconsistencia en la varianza es potencialmente problemática, ya

que evidencia la falta de homogeneidad de los residuales, criterio importante para decidir la validez del modelo. Además, la presencia de puntos de balanceo y puntos influyentes sugiere que ciertas observaciones pueden afectar significativamente los coeficientes y predicciones del modelo. En base a esto concluimos que el modelo no tiene la validez suficiente.

¿Es o no válido?