Trabajo 1

3,5

Estudiantes

Paula Díaz Omen
Natalia Giraldo
Franklin Rodríguez López

Equipo 2

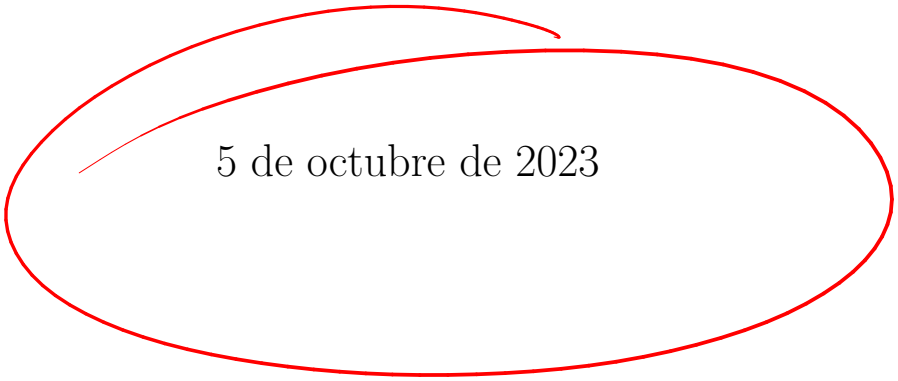
Docente

Julieth Veronica Guarin

Asignatura

Estadística II

Sede Medellín



5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

17,5 pt

Teniendo en cuenta la base de datos brindada, en la cual hay 5 variables regresoras dadas por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 54$$

Donde

- 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

3 pt

Al ajustar el modelo, se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	0.8503
β_1	0.1650
β_2	0.0122
β_3	0.0438
β_4	0.0076
β_5	0.0005

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 0.8503 + 0.165 X_{1i} + 0.0122 X_{2i} + 0.0438 X_{3i} + 0.0076 X_{4i} + 5 \times 10^{-4} X_{5i}$$

1.2. Significancia de la regresión

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

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

Cuyo estadístico de prueba es:

$$F_0 = \frac{MST}{MSE} \overset{H_0}{\sim} f_{5,48} \tag{1}$$

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	28.2249	5	5.644983	6.32367	0.000138821
Error	42.8484	48	0.892675		

De la tabla Anova, se observa un valor p aproximadamente igual a 0.000138821, siendo este mucho menor que el nivel de signifancia 5 %(0.05) por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $0 \leqslant j \leqslant 5$, esto implica, que no hay suficiente evidencia estadística significativa para aceptar la hipótesis nula, por ende, algún $\beta_j \neq 0$ con $0 \leqslant j \leqslant 5$. Esto significa que al menos una de las covariables en el modelo tiene un efecto significativo

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	0.8503	1.7124	0.4966	0.6218
β_1	0.1650	0.0733	2.2509	0.0290
β_2	0.0122	0.0326	0.3748	0.7094
β_3	0.0438	0.0168	2.6106	0.0120
β_4	0.0076	0.0079	0.9676	0.3381
β_5	0.0005	0.0007	0.7050	0.4842

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_1 y β_3 son estadisticamente significativos, pues sus P-valores son menores a α. Esto sugiere que X_1 y X_3 son variables importante para predecir el riesgo de infección.

1.4. Interpretación de los parámetros

$\hat{\beta}_1$**:** Este coeficiente está asociado a X_1:*Duración de la estadía*, significa que; manteniendo el resto de las covariables constantes, un aumento en la *Duración de la estadia*:X_1, de los pacientes en un día se asocia en, promedio, a un aumento 0.1650 de riesgo de infección en los hospitales del estudio.

$\hat{\beta}_3$**:** El coeficiente está asociado a X_3:*Número de camas*, significa que; manteniendo el resto de las covariables constantes, que un aumento en una unidad del *Número de camas*:X_3, se relaciona en, promedio, a un aumento 0.0438 del riesgo de infección en los hospitales del estudio.

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple aproximadamente de $R^2 = 0.3971238$, significa que las covariables del modelo explican aproximadamente el 40 % de la variabilidad total de la variable de respuesta Y :*Riesgo de infección*, mientras que aproximadamente un 60 % restante de la variabilidad se debe a otros factores no incluidos o a la aleatoriedad según el modelo propuesto.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

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

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_3 = \beta_4 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 3, 4 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	42.848	X1 X2 X3 X4 X5
Modelo reducido	43.347	X1 X3 X4

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 54$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$F_0 = \frac{(SSE_R - SSE_F)/r}{SSE_F/(n-p)} \tag{2}$$

En este caso especifico $$

$$\begin{aligned} F_0 &= \frac{(43.347 - 42.848)/2}{42.848/48} \\ &= 0.2795 \end{aligned} \tag{3}$$

Ahora, comparando el F_0 con $f_{0.95,2,48} = 3.1907$, se puede ver que $F_0 < f_{0.95,2,48}$ y por tanto no se rechaza la hipotesis nula. Esto sugiere que las variables excluidas en el modelo reducido no son significativas y que el modelo reducido es suficiente para explicar la variabilidad en los datos. En otras palabras no hay evidencia estadística para concluir que el modelo completo es significativamente mejor que el modelo reducido. $$

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si ¿Existen pruebas estadísticas suficientes para concluir que los coeficientes de las variables predictoras X1 y X2 son iguales entre sí, al igual que los coeficientes de las variables predictoras X4 y X5 en un modelo de regresión lineal que explique el riesgo de infección en los hospitales? por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2;\ \beta_4 = \beta_5 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1(X_1 + X_2) + \beta_3 X_3 + \beta_4(X_4 + X_5) + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 54$$

Donde $X^*_{2i} = X_{2i} + X_{4i}$ y $X^*_{3i} = 3X_{1i} + X_{3i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{K\,n-p} \tag{4}$$

$$F_0 = \frac{((SSE(MR) - 42.8484)/2}{0.892675} \overset{H_0}{\sim} f_{6\,48} \tag{5}$$

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Evaluaremos el supuesto de normalidad mediante un prueba de hipótesis que se realiza por medio de Shapiro-wilk, acompañada de un gráfico cuantil-cuantil:

$$\begin{cases} \mathrm{H}_0 : \varepsilon_i \sim \text{Normal} \\ \mathrm{H}_1 : \varepsilon_i \nsim \text{Normal} \end{cases}$$

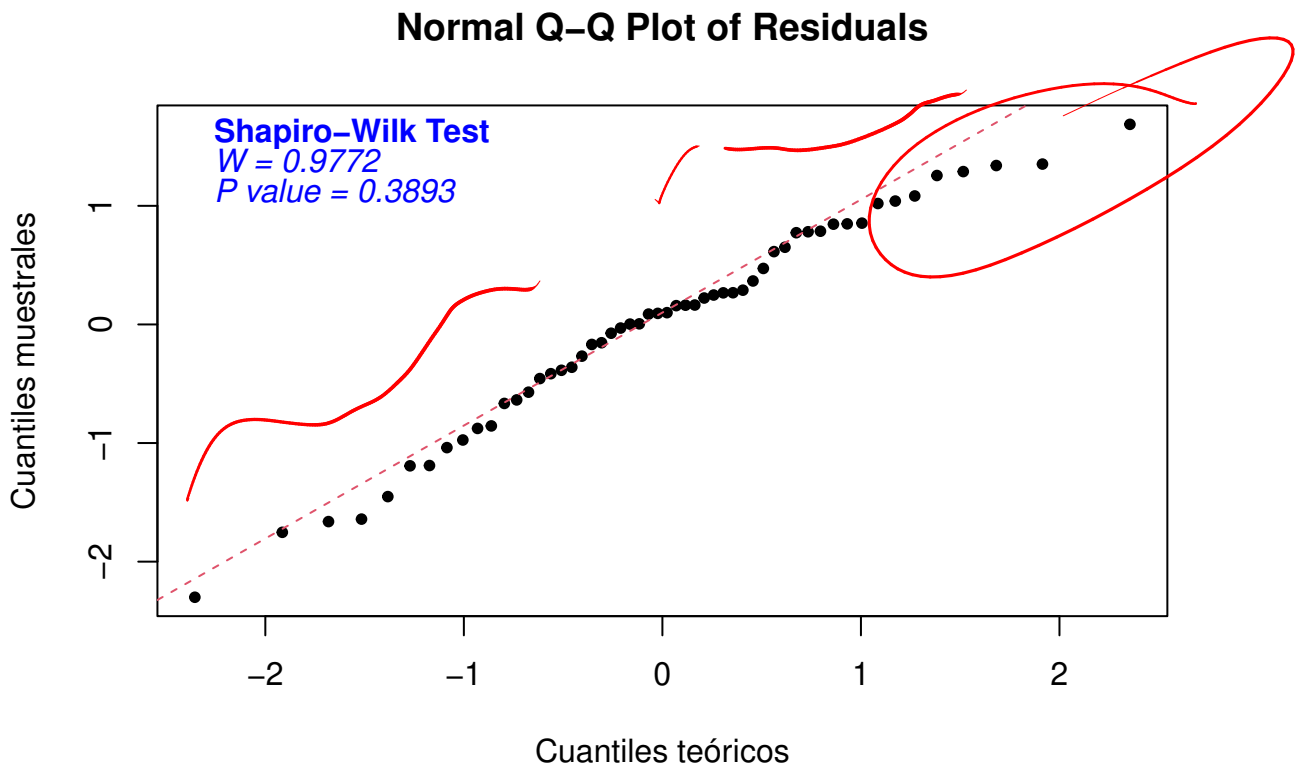


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.3893 y teniendo en cuenta que el nivel de significancia $\alpha = 0.05$, el P-valor es mucho mayor y por lo tanto, no se rechazaría la hipótesis nula, es decir que los datos distribuyen normal, sin embargo la gráfica de comparación de cuantiles permite ver patrones irregulares que posiblemente se deben a datos atípicos, pero son leves, así que, a partir del gráfico tampoco se rechaza la hipótesis nula.

Ahora se validará si la varianza cumple con el supuesto de ser constante.

2pt

ellos no cad san eso

Si observaron esos patrones tan fuertes, ¿por qué no rechazaron?

4.1.2. Varianza constante

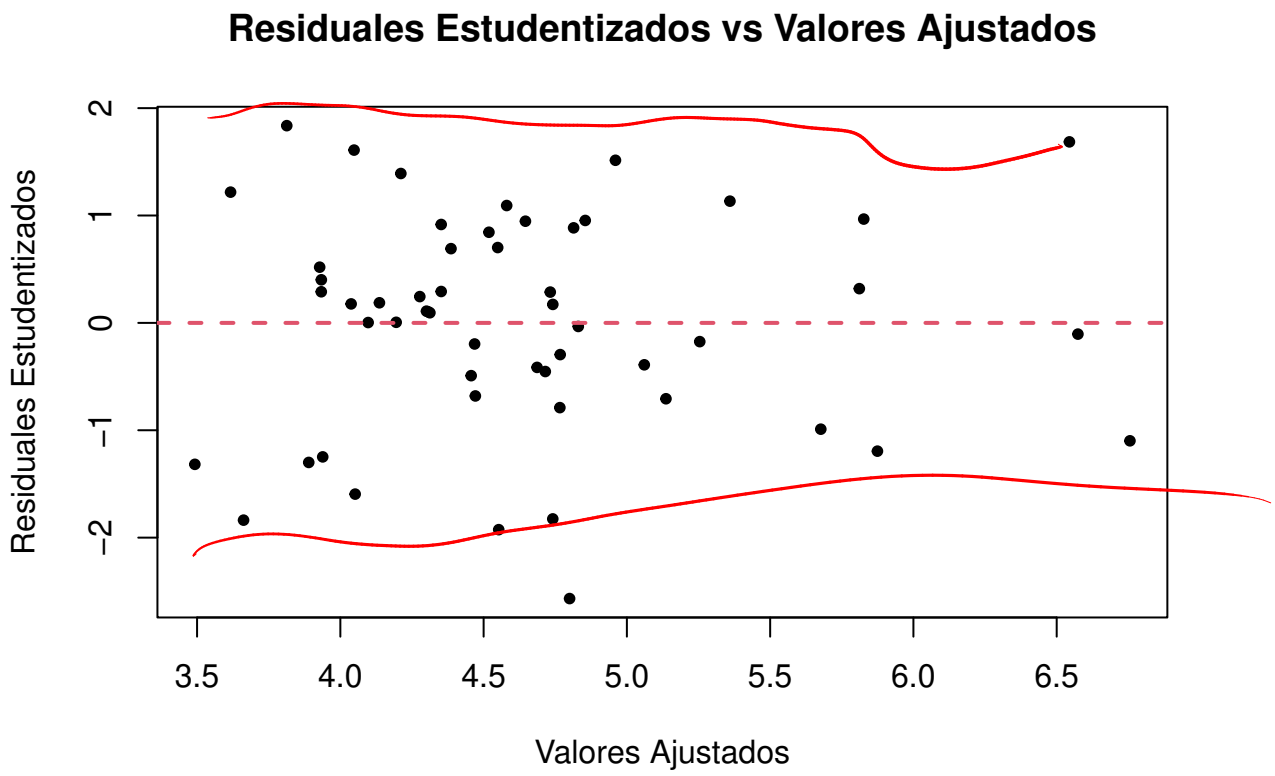


Figura 2: Gráfico residuales estudentizados vs valores ajustados

2pt

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que no hay patrones en los que la varianza aumente, decrezca ni un comportamiento que permita descartar una varianza constante, al no haber evidencia suficiente en contra de este supuesto se acepta como cierto. Además es posible observar media 0.

✓

Sí lo hay, pero no es fuerte

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

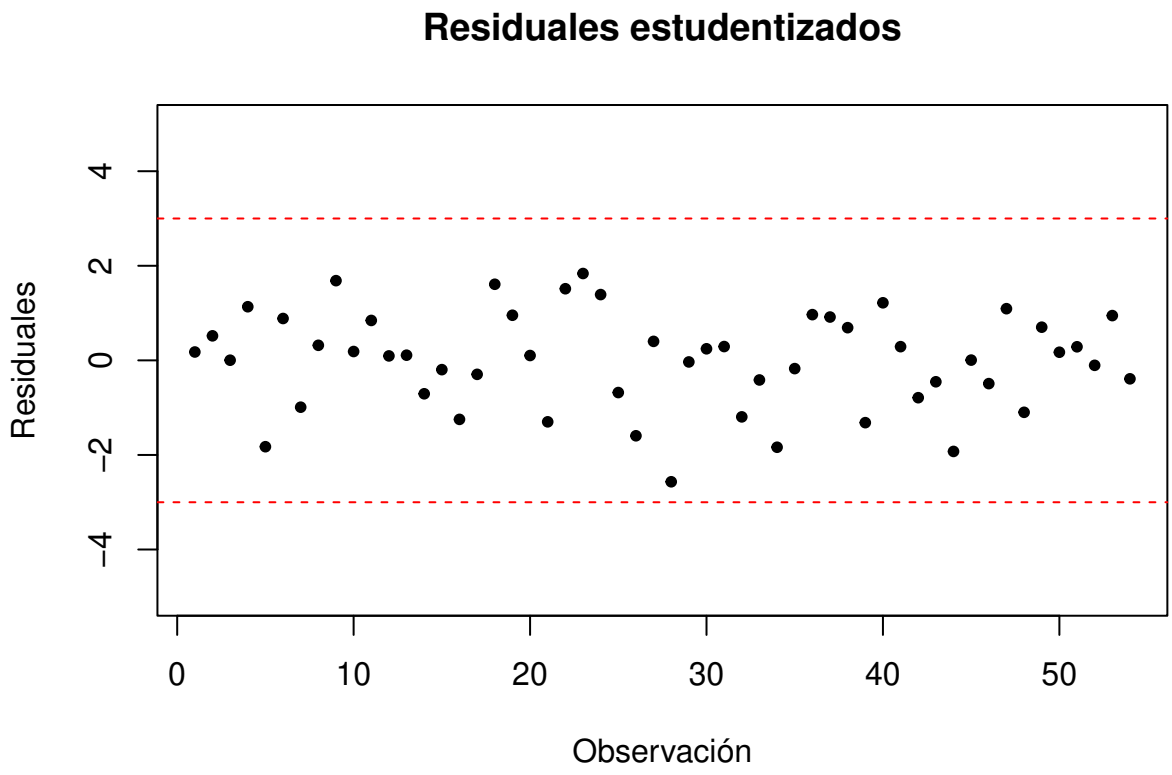


Figura 3: Identificación de datos atípicos

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

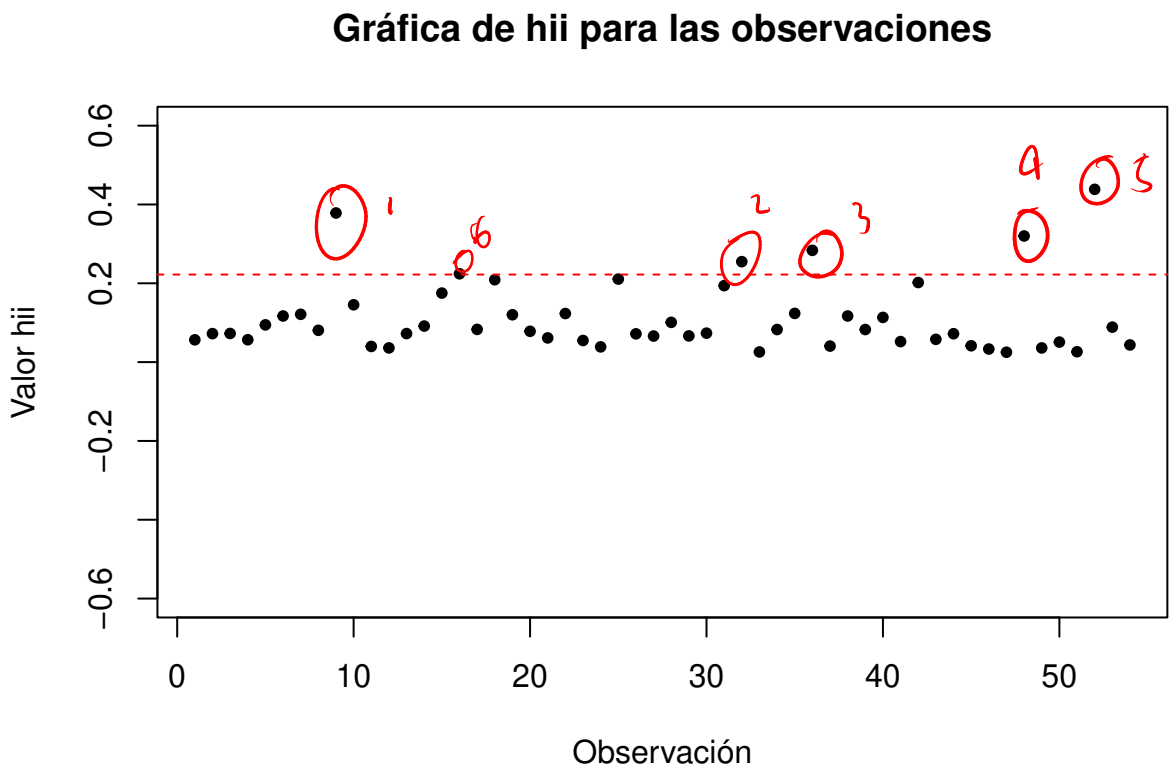


Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 9    1.6858  0.2885    0.3785  1.3423
## 16  -1.2479  0.0750    0.2241 -0.6748
## 32  -1.1947  0.0814    0.2549 -0.7019
## 36   0.9669  0.0617    0.2836  0.6079
## 48  -1.0983  0.0946    0.3201 -0.7552
## 52  -0.1047  0.0014    0.4384 -0.0915
```

A la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $2\frac{p}{n} = 0.22$ se puede apreciar que existen 6 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{p}{n}$, los cuales son los presentados en la tabla Bajo el criterio de puntos de balanceo y generan ruido o aumentan la variabilidad en el modelo.

4.2.3. Puntos influenciales

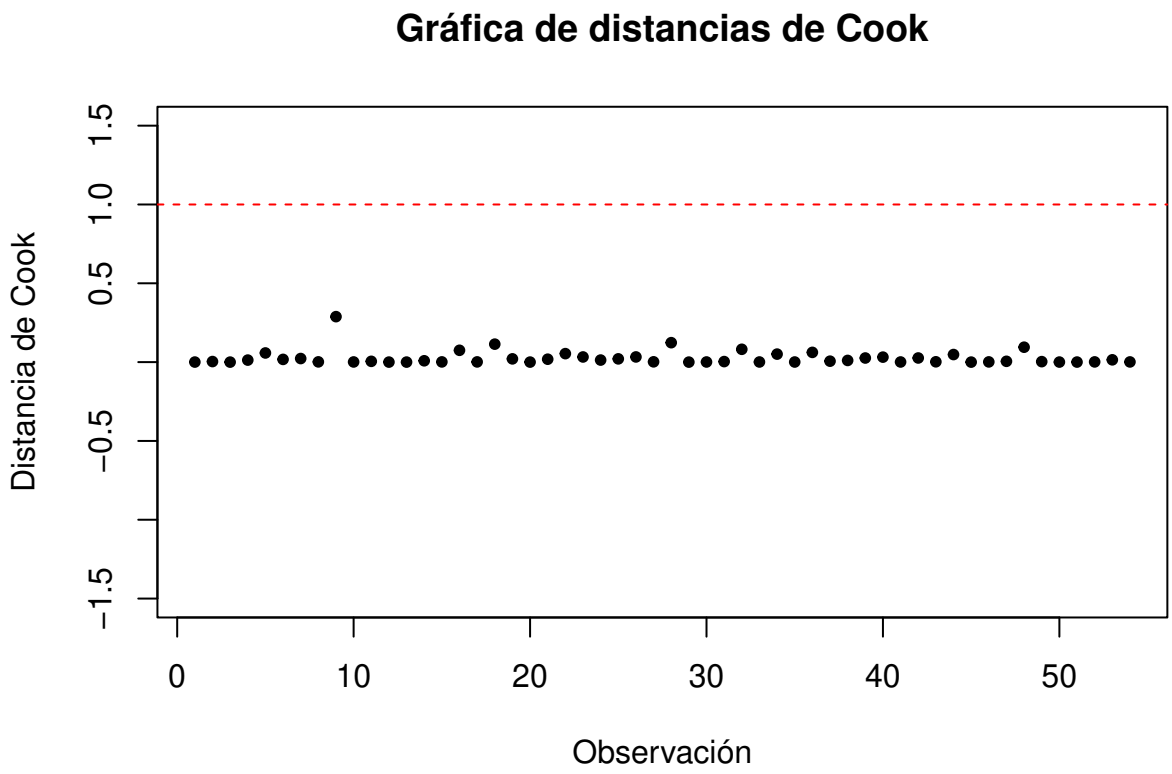


Figura 5: Criterio distancias de Cook para puntos influenciales

Como se puede visualizar en la gráfica de distancias de Cook, donde la línea punteada roja representa el valor 1, se puede inferir que no hay medidas de influencia a partir del criterio $D_i > 1$.

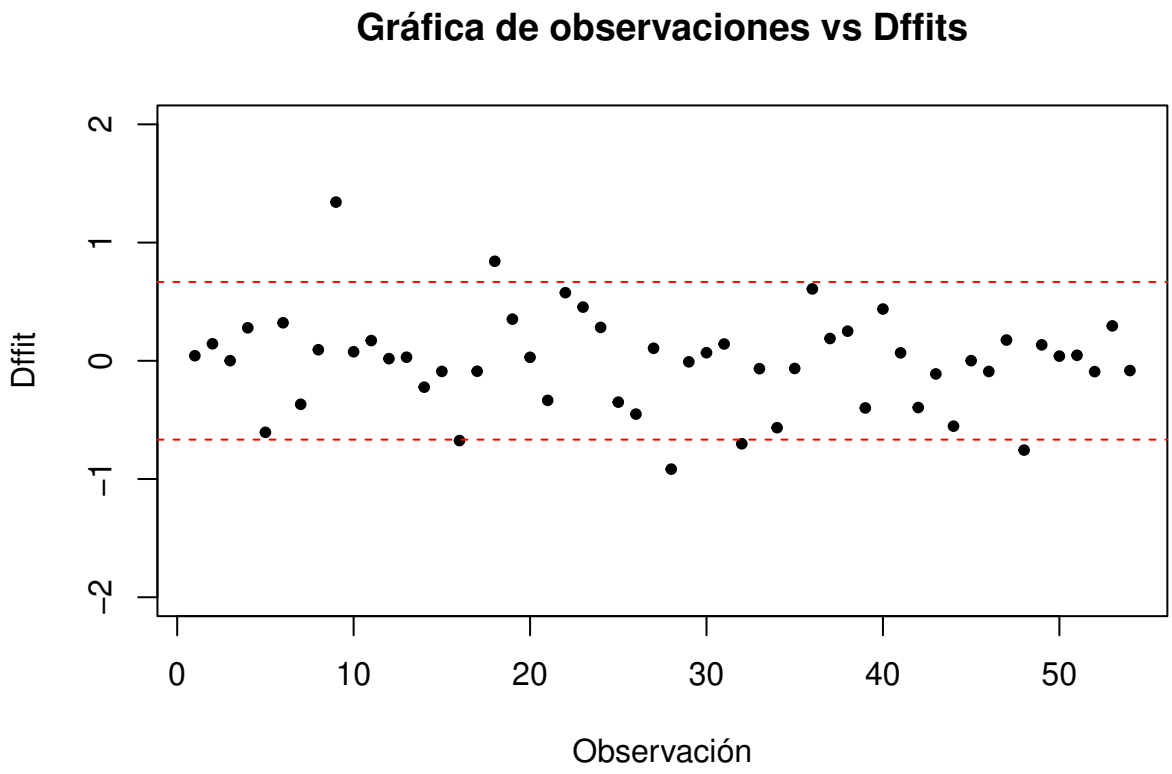


Figura 6: Criterio Dffits para puntos influenciales

3 pt

```
##    res.stud Cooks.D hii.value  Dffits
## 9    1.6858  0.2885    0.3785  1.3423
## 16  -1.2479  0.0750    0.2241 -0.6748
## 18   1.6092  0.1141    0.2091  0.8417
## 28  -2.5674  0.1233    0.1009 -0.9165
## 32  -1.1947  0.0814    0.2549 -0.7019
## 48  -1.0983  0.0946    0.3201 -0.7552
```

Como se puede ver, las observaciones 9, 16, 18, 28, 32 y 48 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{6}{54}} = 0.667$, es un punto influencial a pesar de 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.

¿Qué causan?

4.3. Conclusión

1 pt

Del modelo se puede inferir que tiene dos variables significativas, adicionalmente también encontramos que los supuestos del modelo se cumplen, con algunos patrones irregulares

en las colas los errores se distribuyen normal y aunque la varianza tiene algunas irregularidades como puntos de balanceo e influénciales, es constante. El R^2 no explica gran porcentaje de variabilidad del modelo, sin embargo el R^2 no nos lleva a descartar el modelo.

uno no descarta con R^2

es o no es válido?