4,45

Trabajo 1

Estudiantes

Andrés Stevens Arrieta Muñoz
Isaac Mesa Maya
Federico Toro Alvarez
María José Uribe Henao

Equipo #31

Docente

Francisco Javier Rodriguez Cortes

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín

30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

19,5 pt

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

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 60$$

Básandonos en una muestra aleatoria de 60 hospitales, se plantea el modelo de regresión, donde:

Variable respuesta:

$$Y_i = Riesgo\ de\ infección\ (porcentaje)$$

Variables Regresoras:

$$X_1 = Duración\ de\ la\ estadía\ (días)$$

$$X_2 = Rutina\ de\ cultivos\ (por\ cada\ 100\ pacientes)$$

$$X_3 = Numero\ de\ camas$$

$$X_4 = Censo\ promedio\ diario\ (número\ de\ pacientes\ por\ día)$$

$$X_5 = Número\ de\ enfermeras$$

1.1. Modelo de regresión

Al ajustar el modelo, se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del estimado del parámetro
β_0	-0.7960
β_1	0.1699
β_2	0.0135
β_3	0.0490
β_4	0.0203
β_5	0.0011

3 pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.796 + 0.1699 X_{1i} 0.0135 X_{2i} + 0.049 X_{3i} + 0.0203 X_{4i} + 0.0011 X_{5i}$$

1.2. Significancia de la regresión

5 pt

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:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,54} \tag{1}$$

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	73.2337	5	14.64674	15.6581	1.66117e-09
Error	50.5121	54	0.93541		

De la tabla Anova, se observa un valor P igual a $1.66117e-09$ el cual es aproximadamente igual a 0, por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$, así que se acepta la hipótesis alternativa en la que algún $\beta_j \neq 0$, por lo tanto la regresión del modelo es significativa.

1.3. Significancia de los parámetros

6 pt

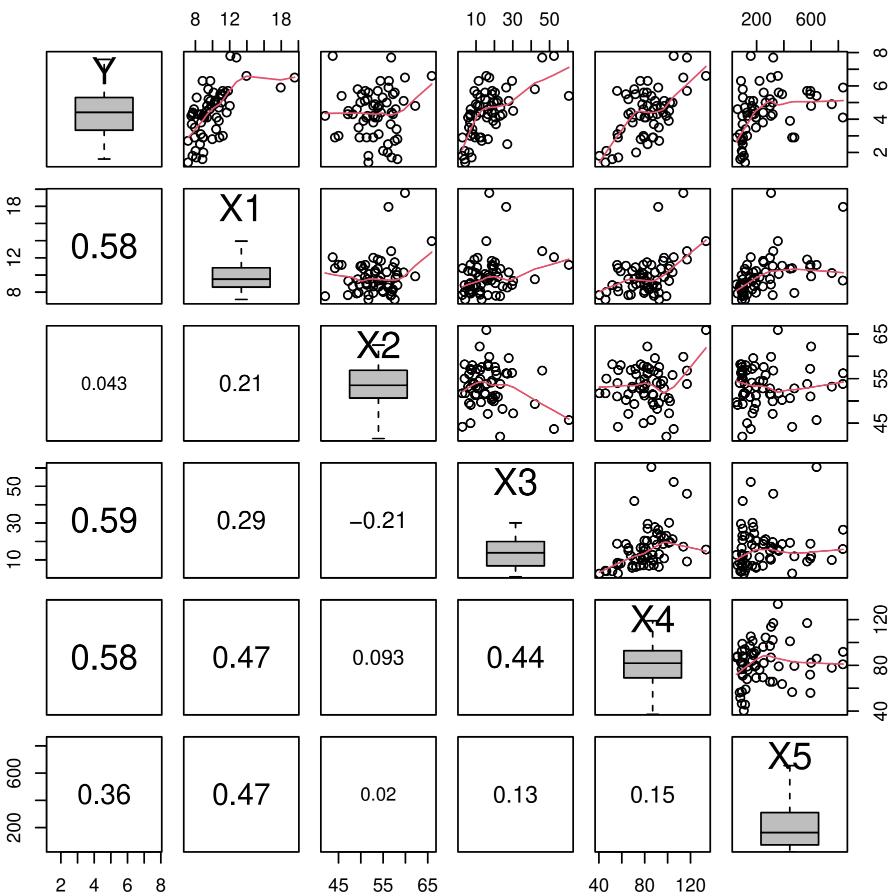
En el siguiente cuadro se presenta información de los parámetros, la cual permitirá determinar cuáles de ellos son significativos.

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-0.7960	1.5709	-0.5067	0.6144
β_1	0.1699	0.0744	2.2833	0.0264
β_2	0.0135	0.0298	0.4540	0.6517
β_3	0.0490	0.0127	3.8418	0.0003
β_4	0.0203	0.0083	2.4370	0.0181
β_5	0.0011	0.0007	1.4948	0.1408

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_1, β_3 y β_4 son significativos, pues sus P-valores son menores a α.

Mientras que los parametros β_0, β_2 y β_5 no son significativos puesto que sus P-valores son mayores a α.



De la gráfica de correlaciones entre variables predictoras y respuesta Y podemos observar que si existe una mayor relación entre las predictoras X_1 X_3 X_4 con la variable de la respuesta Y. Este análisis también tiene concordancia con el análisis por medio de la prueba de hipótesis anterior donde concluimos que β_1, β_3 y β_4 son significativos para el modelo. → Realmente no tanto, pero muy bien por el análisis.

1.4. Interpretación de los parámetros

2,5 pt

→ la probabilidad promedio

$\hat{\beta}_1$: Por cada día que aumenta la duración de la estadía, el riesgo de infección aumenta en un 0.169866653, es decir un 16.99 % cuando los valores de las demas predictoras son constantes fijas.

$\hat{\beta}_3$: Por cada nueva cama que se ocupa en el hospital, el riesgo de infección aumenta en un 0.048970542, es decir un 4.897 % cuando los valores de las demas predictoras son constantes fijas.

$\hat{\beta}_4$: Por cada aumento en el número promedio de pacientes diarios, el riesgo de infección aumenta en un 0.020263877, es decir un 2.03 % cuando los valores de las demas predictoras son constantes fijas.

1.5. Coeficiente de determinación múltiple R^2

3 pt

El modelo de regresión tiene un coeficiente de determinación múltiple $R^2 = 0.592$, lo que significa que explica aproximadamente el 59.2 % de la variabilidad total observada en la respuesta. Si analizamos el R^2 ajustado de todo el modelo, el cual es 55.4 %, se puede observar que hay una disminución en su valor lo que nos indica que existe la posibilidad de que el modelo presenta variables que no están aportando al resultado.

¿Cómo se calcula?

2. Pregunta 2

4,5 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con el P-valor más alto en el modelo fueron X_1, X_2, X_5, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 2, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	50.512	X1 X2 X3 X4 X5
Modelo reducido	64.258	X3 X4

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

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 60$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

3pt

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_4) - SSE(\beta_0,\ \cdots,\ \beta_5))/3}{MSE(\beta_0,\ \cdots,\ \beta_5)} \overset{H_0}{\sim} f_{3,54} \\ &= \frac{64.258 - 50.512/3}{0.93451} \\ &= 4.89387 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,54} = 2.7758$, se puede ver que $F_0 > f_{0.95,1,45}$ 1,5pt

No es posible descartar las variables del subconjunto, ya que se rechaza la hipótesis nula.

Primero se dice que el subconjunto es significativo y luego que no se descarta 4 pt

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si el número de camas es igual al doble del número de enfermeras, y además se pregunta si la duración de la estadía es igual a la rutina de cultivos. Por consiguiente se plantea la siguiente prueba de hipótesis: → mal planteadas las preguntas

$$\begin{cases} \mathrm{H}_0 : \beta_3 = 2\beta_5;\ \beta_1 = \beta_2 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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

2pt

$$L = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & -2 \\ 0 & 1 & -1 & 0 & 0 & 0 \end{bmatrix}$$

El modelo reducido está dado por:

0pt

$$Y_i = \beta_o + \beta_1 X^*_{1,2i} + \beta_3 X^*_{3,5i} + \beta_4 X_{4i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 60$$

Donde $X^*_{1i} = X_{1i} + X_{2i}$ y $\boxed{X^*_{3i} = X_{3i} + 2X_{5i}}$ ✗ $X_{3i} = 2X_{3i} + X_{5i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

2pt

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,54} F_0 = \frac{(SSE(MR) - 50.5121)/2}{0.9354} \overset{H_0}{\sim} f_{2,54} \tag{3}$$

4. Pregunta 4

16,5 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

4 pt

Para la validación de este supuesto, se plantea la siguiente prueba de hipótesis ~~Shapiro-Wilk~~, acompañada de un gráfico cuantil-cuantil:

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

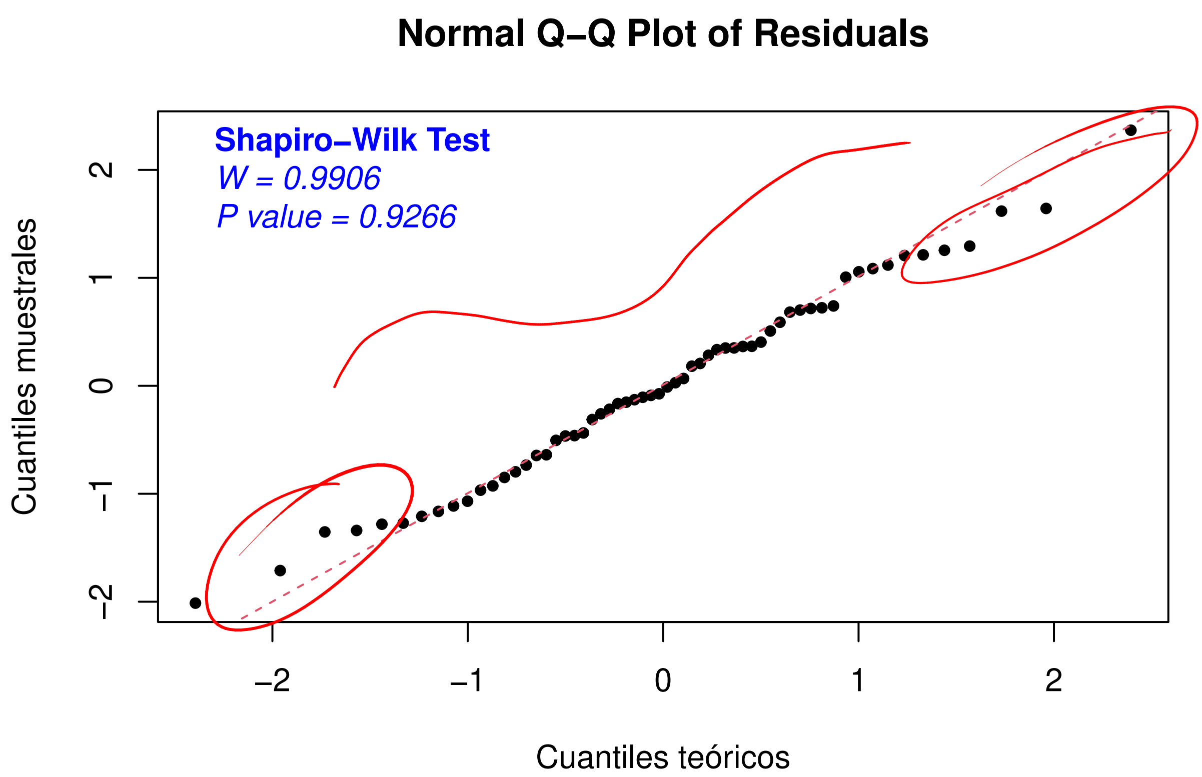


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.9266 y con un nivel de significancia $\alpha = 0.05$, el P-valor es muchísimo mayor y por lo tanto no se rechazaría la hipótesis nula, es decir que los datos distribuyen normal, con media μ y varianza σ^2. Sin embargo, observando la gráfica de comparación de cuantiles se puede ver patrones irregulares especialmente en las colas, y como es más importante el análisis gráfico, se termina por rechazar el cumplimiento del supuesto de normalidad.

4.1.2. Varianza constante

3 pt

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

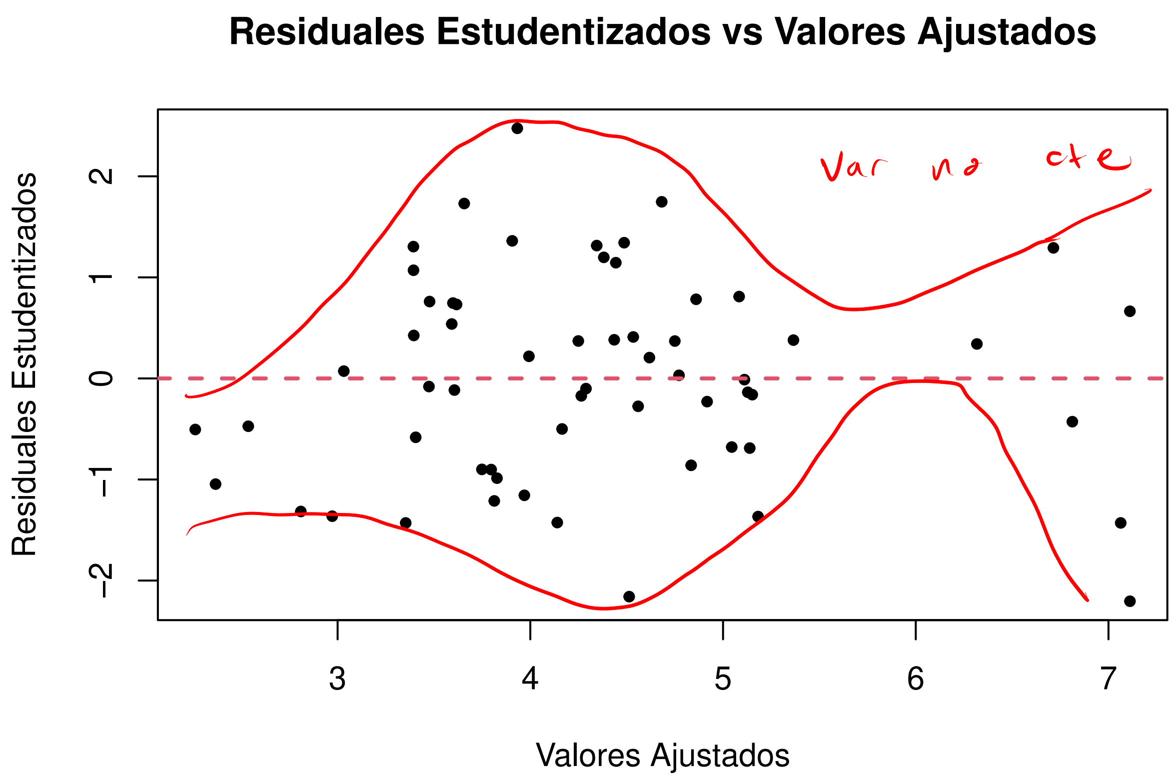


Figura 2: Gráfico Residuales estudentizados vs Valores ajustados

En el gráfico de Residuales estudentizados vs Valores ajustados, se puede observar que hay un patrón donde los valores ajustados entre 1.4 hasta 3.4 presentan unos pocos residuales estudentizados negativos y pequeños. Los valores ajustados entre 3.5 a 5.6 aumenta la dispersión de los residuales y se acumula la mayor cantidad de valores ajustados. Para terminar los valores ajustados entre 6.3 hasta 7.7 que es el máximo son menores y con una menor distribución de los datos. Con esto concluimos que el modelo presenta un patrón de crecimiento y luego decrecimiento. Como se presentan patrones en el la gráfica de Residuales Estudentizados vs Valores Ajustados determinamos que la varianza no es constante en el modelo.

Excelente!

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

3pt

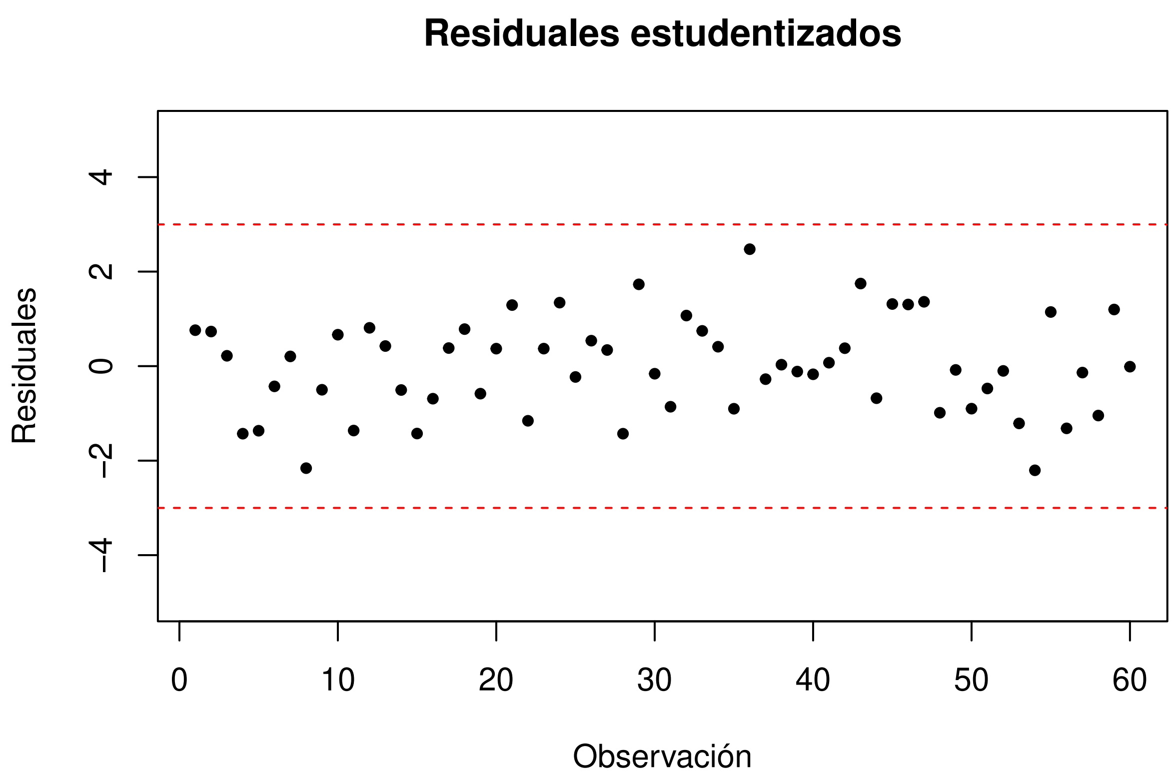


Figura 3: Identificación de datos atípicos

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

1pt

Ampliar límites

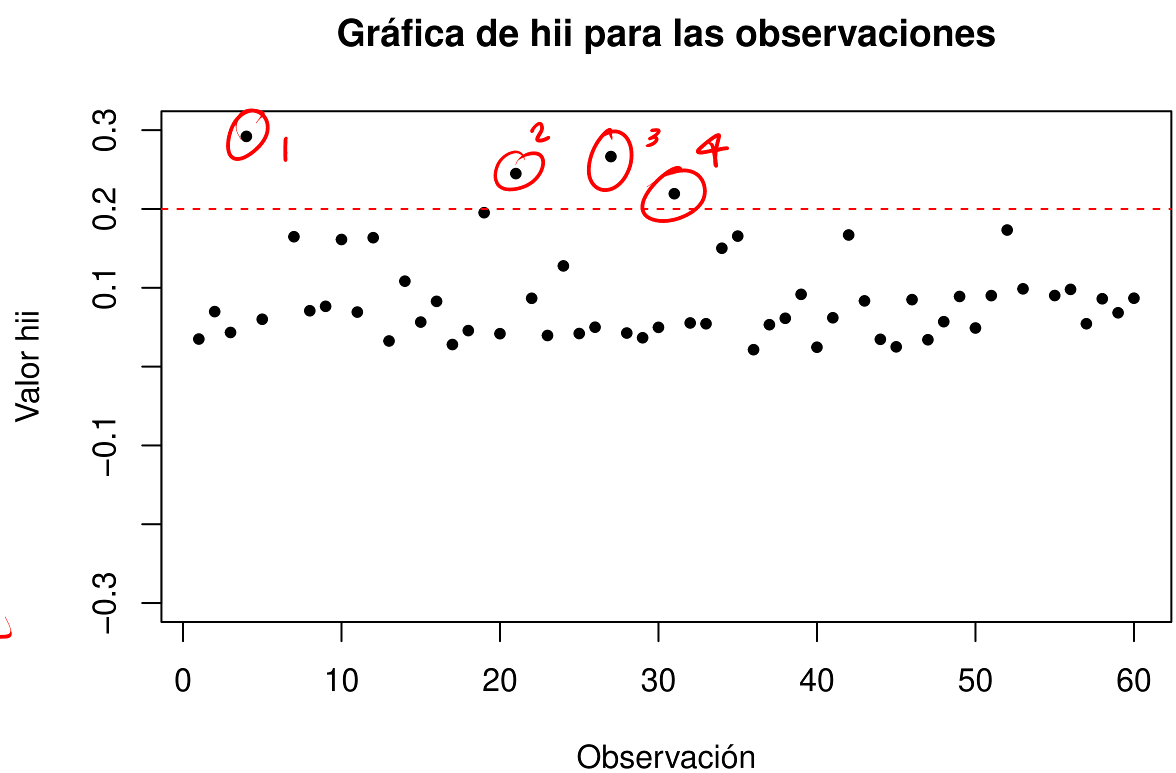


Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 4   -1.4297  0.1406    0.2921 -0.9276
## 6   -0.4281  0.0231    0.4306 -0.3695
## 21   1.2920  0.0902    0.2449  0.7406
## 27   0.3413  0.0071    0.2665  0.2040
## 31  -0.8596  0.0346    0.2194 -0.4546
## 54  -2.2045  0.4471    0.3557 -1.7010
```

-> tabla, no salida

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$ ($h_{ii} = 0.2$), 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.

eso no es una tabla, es un print de un data.frame

mentira, ahí se ven 4, no son congruentes

¿Qué causan?

4.2.3. Puntos influenciales

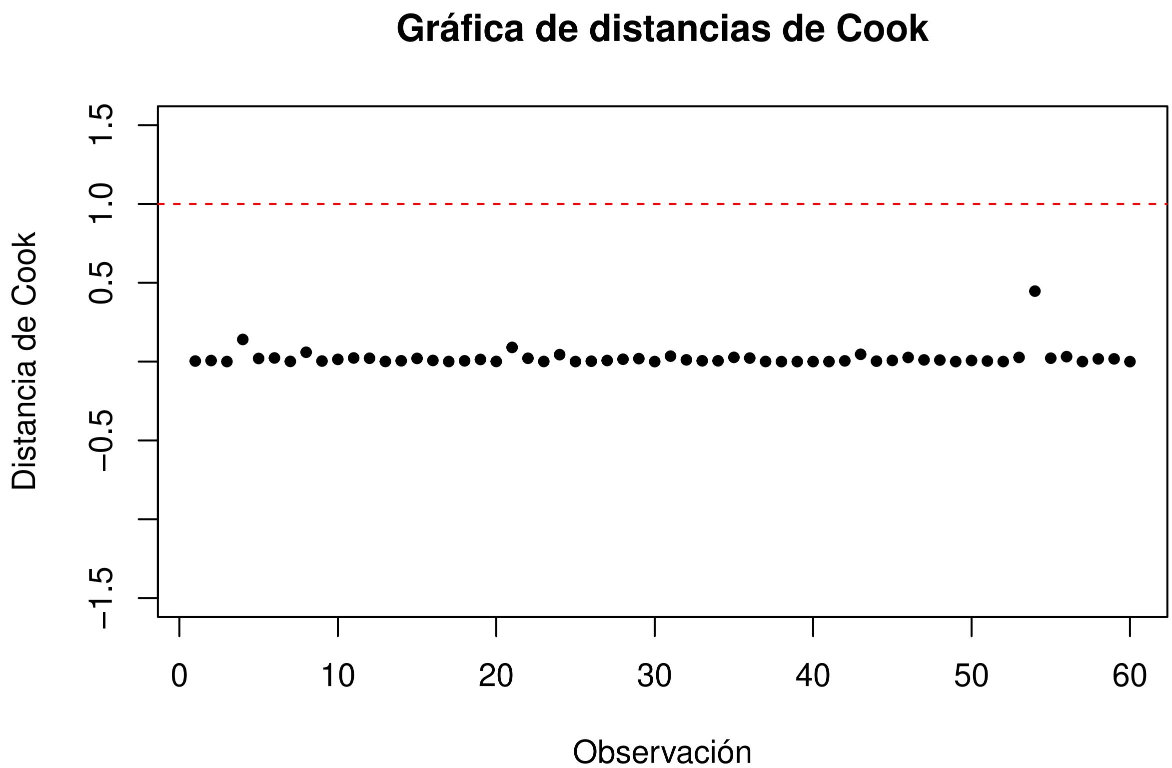


Figura 5: Criterio distancias de Cook para puntos influenciales

Ampliar límites

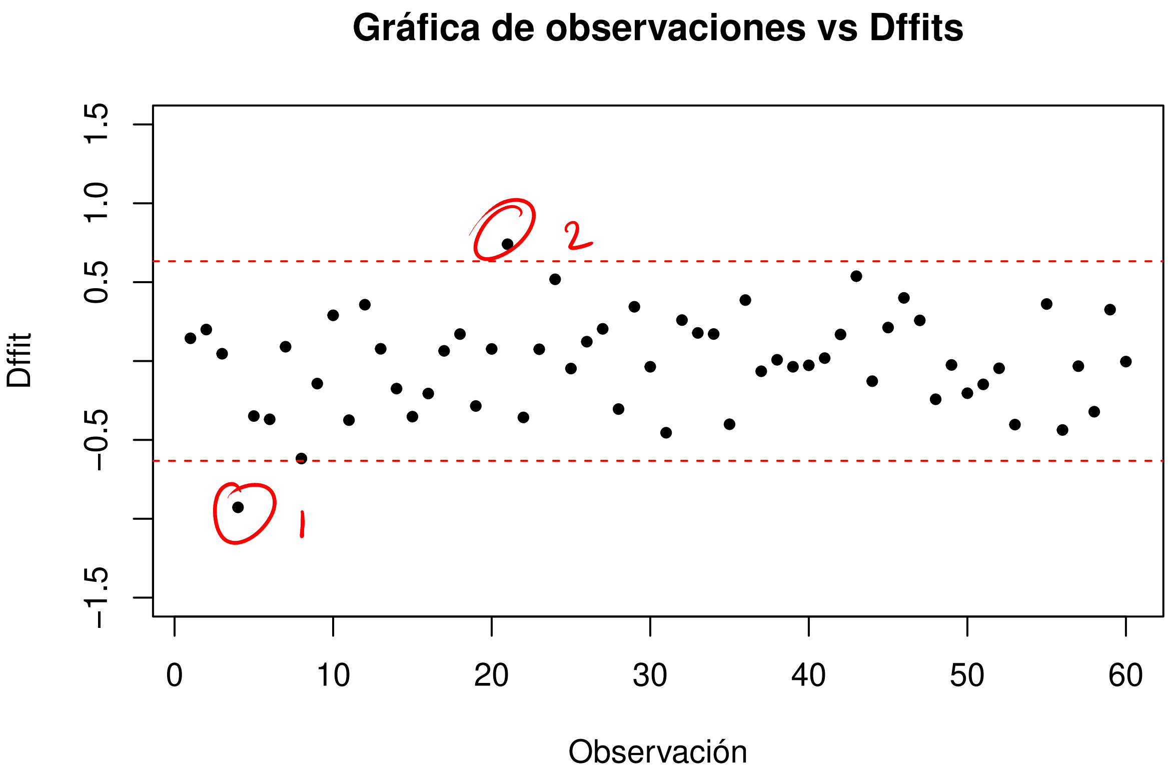


Figura 6: Criterio Dffits para puntos influenciales

```
##     res.stud Cooks.D hii.value  Dffits
## 4    -1.4297  0.1406    0.2921 -0.9276
## 21    1.2920  0.0902    0.2449  0.7406
## 54   -2.2045  0.4471    0.3557 -1.7010
```

Reportan 3 y sólo veo 2, límite inferior debía llegar al menos hasta -1,71.

~~Como se puede ver,~~ las observaciones 4, 21 y 54 son puntos influyentes según el criterio de Dffits. Este criterio establece que cualquier punto cuyo $|D_{ffits}| > 2\sqrt{\frac{p}{n}} = 0.63$, es un punto influyente. Es importante destacar que con el criterio de distancias de Cook, el cual establece que cualquier punto cuya $D_i > 1$ es un punto influyente, ninguno de los datos cumple con este criterio para ser determinados influyentes. 2,5 pt

¿Qué causan?

4.3. Conclusión

3 pt

Debido a que no se cumplen los criterios de normalidad, ni de varianza constante se determina que el modelo no es válido para realizar un análisis sobre el riesgo de contagio, basándonos en las variables predictoras descritas con anterioridad.