4,15

Eficacia en el Control de Infecciones Hospitalarias

Diego Mauricio Ballesteros Osorio

Valentin Jose Padilla Marimon

León Felipe Restrepo Perez

Santiago Molina Muñoz

Estadística II

Equipo 22

Escuela de Estadística, Universidad Nacional de Colombia

Medellín

2023

Problema planteado:

En un estudio a gran escala realizado en **EE.UU** sobre la eficacia en el control de infecciones hospitalarias se recogió información en 113 hospitales. Y se obtuvo una muestra aleatoria con **60** registros. Según la base de datos considere lo siguiente:

Tabla 1. Descripción de variables regresoras y de respuesta.

Variable	Descripción
Y: Riesgo de infección	Probabilidad promedio estimada de adquirir infección en el hospital (Porcentaje)
X_1: Duración de la estadía	Duración promedio de la estadía de todos los pacientes en el hospital (Días)
X_2: Rutina de cultivos	Razón del número de cultivos realizados en pacientes sin síntomas de infección hospitalaria, por cada 100
X_3: Número de camas	Número promedio de camas en el hospital durante el periodo de estudio
X_4: Censo promedio diario	Número promedio de pacientes en el hospital por dia durante el periodo del estudio
X_5: Número de enfermeras	Número promedio de enfermeras, equivalente a tiempo completo, durante el periodo de estudio

Problemas a resolver:

Pregunta 1.

18 pt

1.1 Estimación del modelo de regresión lineal múltiple

El modelo que se va a emplear para aproximar la relación entre la variable **Y** y las variables $\mathbf{X_k}$; **k=1,2,3,4,5.** (Ver tabla 1). Está dado por: $1 \leq i \leq 60$.

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

Al hacer una estimación de los parámetros: $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, $\boldsymbol{\beta_3}$, $\boldsymbol{\beta_4}$, $\boldsymbol{\beta_5}$. Se obtienen los parámetros estimados: $\mathbf{B_j}$; **j = 0, 1, 2, 3, 4, 5.** Expuestos en la siguiente tabla:

Tabla 2. Coeficientes de regresión estimados de $\mathbf{B_j}$

	Coeficientes de regresión estimados
$\mathbf{B_0}$	-1.0186398
$\mathbf{B_1}$	0.2313154
$\mathbf{B_2}$	0.0226572
$\mathbf{B_3}$	0.0643261
$\mathbf{B_4}$	0.0053465
$\mathbf{B_5}$	0.0018688

3pt

Con los datos dados por la **tabla 2**. La regresión estimada o modelo de regresión ajustado será de la forma:

$$\hat{Y}_i = -1.0186398 + 0.2313154\, X_{1i} + 0.0226572\, X_{2i} + 0.0643261\, X_{3i} + 0.0053465\, X_{4i} + 0.0018688\, X_{5i}$$

1.2 Significancia de la regresión

5pt

Para plantear la significancia de la regresión se formula la siguiente prueba de hipótesis:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0\ ,\ \ H_1: \text{Algún } \beta_k \neq 0\ ,\ k = 1, 2, 3, 4, 5.$$

Para continuar con el análisis de la significancia de la regresiòn. Es necesario calcular la tabla de análisis de varianza que estará compuesta por distintos elementos que son: Fuente de variación, suma de cuadrados, grados de libertad, cuadrados medios y finalmente un estadístico de prueba con el que se probaràn las hipótesis planteadas anteriormente. La forma de la tabla queda de la siguiente manera:

Tabla 3. Análisis de varianza

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	Valor estadístico de prueba F
Regresiòn	67.1743	5	13.434854	$\mathbf{F_0} = 14.2182$
Error	51.0251	54	0.944909	
Total	118.1994	59		

Usando los datos dados por la tabla de análisis de varianza. **(Ver tabla 3).** El estadístico de prueba està dado por:

$$\mathbf{F_0} = \frac{CUADRADO\ MEDIO\ DE\ LA\ REGRESIÒN}{CUADRADO\ MEDIO\ DEL\ ERROR} = \frac{\left(\frac{SUMA\ DE\ CUADRADOS\ DE\ LA\ REGRESIÒN\ (SCR)}{GRADOS\ DE\ LIBERTAD\ ASOCIADO\ A\ SCR}\right)}{\left(\frac{SUMA\ DE\ CUADRADOS\ DEL\ ERROR\ (SCE)}{GRADOS\ DE\ LIBERTAD\ ASOCIADO\ A\ SCE}\right)}\ ;\ \mathbf{F_0 \sim f_{5,54}}$$

Es decir,

$$\mathbf{F_0} = \frac{13.434854}{0.944909} = \frac{\frac{67.1743}{5}}{\frac{51.0251}{54}} = \mathbf{14.2182}$$

¿sobra el valor absoluto

Se rechaza $\mathbf{H_0}$ con un nivel de significancia $\boldsymbol{\alpha} = \mathbf{0.05}$ si: $\mathbf{|\ F_0\ | > f_{\alpha,5,54}}$ ò $\mathbf{VP < \alpha}$ donde $\mathbf{VP = P(f_{5,54} > |\ F_0\ |)}$. Entonces se tiene que: $\mathbf{(f_{\alpha,5,54} = 2.38607) \Rightarrow\ |\ F_0\ | = |\ 14.2182\ | > 2.38607}$ **y** $\mathbf{VP = 7.10924e\text{-}09 < 0.05}$.

Con los resultados dados tanto por la regiòn de rechazo como por el **VP**, concluimos el rechazo de la hipótesis nula $\mathbf{H_0}$ y decimos que podrìa haber al menos un $\boldsymbol{\beta_k} \neq 0$ en la regresiòn que la hace significativa.

1.3 Significancia individual de parámetros 6pt

Para probar la significancia individual de los parámetros y confirmar a su vez el rechazo de la hipótesis nula del subpunto **1.2** para los parámetros $\boldsymbol{\beta_k}$. Debemos plantear nuevas pruebas de hipótesis pero esta vez para cada parámetro de la regresiòn estimada o ajustada $\boldsymbol{\beta}_j$, es decir, incluiremos el intercepto en este caso. Las pruebas de hipótesis tienen la forma:

$$\mathbf{H_0: \beta_j = 0\ ,\ \ H_1: \beta_j \neq 0\ ;\ j = 0, 1, 2, 3, 4, 5.}$$

El estadístico de prueba adquiere la forma:

ojo con la notación, β, no B

$T_{0,j}$

$$\mathbf{T_j = B_j\ /\ ee(B_j)\ ;\ T_j \sim t_{54}}$$

Para calcular cada estadístico de prueba $\mathbf{T_j}$ según corresponda usamos las tablas **2** y **4**.

Tabla 4. Errores estándar (ee) y valores P de $\mathbf{B_j}$

	Errores estándar	Valores P
$\mathbf{B_0}$	1.4895450	0.497
$\mathbf{B_1}$	0.0769191	0.004
$\mathbf{B_2}$	0.0270241	0.405
$\mathbf{B_3}$	0.0151851	8.91e-05
$\mathbf{B_4}$	0.0074696	0.477
$\mathbf{B_5}$	0.0007184	0.012

Asì cada estadístico de prueba es:

$$\mathbf{T_0 = -0.684\ ,\ T_1 = 3.007\ ,\ T_2 = 0.838,\ T_3 = 4.236\ ,\ T_4 = 0.716\ ,\ T_5 = 2.602\ ;\ T_j \sim t_{54}}$$

Se rechaza $\mathbf{H_0}$ con un nivel de significancia $\boldsymbol{\alpha} = \mathbf{0.05}$ cuando: La regiòn de rechazo de $\mathbf{H_0}$ està dada por $\mathbf{RR = \{T_{0,j} / |\ T_{j,0}| > t_{\alpha/2,54}\}}$ ò $\mathbf{VP < \alpha}$ cuando $\mathbf{VP = P(t_{54} > |\ T_j\ |) < \alpha.\ (t_{\alpha/2,54} = 2.004879).}$

Podemos observar que tanto por el criterio de **valor P (Ver tabla 4)** como de **región de rechazo (RR)**, solo los parámetros $\mathbf{B_1, B_3, B_5}$ son significativos, es decir, solo en estos casos rechazamos la hipótesis $\mathbf{H_0}$ que nos dice que son iguales a 0, es decir, no significativos. Mientras que $\mathbf{B_0, B_2}$ **y** $\mathbf{B_4}$ no son significativos.

Sólo se interpretaban las significativas

1.4 Interpretación de parámetros estimados 1pt

B

- **~~B~~$_0$:** Este parámetro no tiene interpretación debido a que ninguna de las variables regresoras tomadas en cuenta son 0 simultáneamente en la muestra, es decir, el "punto" (0, 0, 0, 0, 0) no es considerado en la muestra.
- **~~B~~$_1$:** Este parámetro nos dice que la probabilidad porcentual promedio estimada de adquirir infección en el hospital aumenta a razón de 0.2313154 unidades por cada unidad de aumento de duración de días promedio de la estadía de todos los pacientes en el hospital. Cuando las demás se quedan constantes
- **~~B~~$_2$:** Este parámetro nos dice que la probabilidad porcentual promedio estimada de adquirir infección en el hospital aumenta a razón de 0.0226572 unidades por cada unidad de aumento en la razón del número de cultivos realizados en pacientes sin síntomas de infección hospitalaria, por cada 100.
- **~~B~~$_3$:** Este parámetro nos dice que la probabilidad porcentual promedio estimada de adquirir infección en el hospital aumenta a razón de 0.0643261 unidades por cada unidad de aumento en el número promedio de camas en el hospital durante el periodo de estudio.
- **~~B~~$_4$:** Este parámetro nos dice que la probabilidad porcentual promedio estimada de adquirir infección en el hospital aumenta a razón de 0.0053465 unidades por cada unidad de aumento en el número promedio de pacientes en el hospital por día durante el periodo del estudio.
- **~~B~~$_5$:** Este parámetro nos dice que la probabilidad porcentual promedio estimada de adquirir infección en el hospital aumenta a razón de 0.0018688 unidades por cada unidad de aumento en el número promedio de enfermeras, equivalente a tiempo completo, durante el periodo de estudio.

1.5 Coeficiente de determinación múltiple e interpretaciòn R^2

3pt

El coeficiente de determinaciòn múltiple $\mathbf{R^2}$, està definido de la forma:

$$\mathbf{R^2} = 1 - \frac{\textit{SUMA DE CUADRADO DEL ERROR (SCE)}}{\textit{SUMA TOTAL DE CUADRADOS CORREGIDOS (STCC)}} \text{ ò } \frac{\textit{SUMA DE CUADRADOS DE LA REGRESIÒN (SCR)}}{\textit{SUMA TOTAL DE CUADRADOS CORREGIDOS (STCC)}}$$

✓

Asì:

$$\mathbf{R^2} = 1 - \frac{51.0251}{118.1994} \text{ ò } \frac{67.1743}{118.1994} = \mathbf{0.56831}$$

$$\mathbf{R^2 = 0.56831(100) = 56.831\%}$$

El coeficiente de determinación múltiple (en porcentaje) nos dice que el **56.831%** de la variabilidad total en la probabilidad promedio estimada de adquirir infección en el hospital es explicada por el modelo planteado. Hay que recordar que este coeficiente estima su valor teniendo en cuenta todas las variables, independientemente de si sus parámetros que las

✓

acompañan son o no significativos. Es por eso, que existe otra medida sin interpretación que reduce este error de estimación y es el $\mathbf{R^2}_{\mathbf{ajustado}}$.

Pregunta 2.

5 pt

2.1 Planteamiento de pruebas de hipótesis y modelo reducido

Las 3 co-variables con el **Valor P** más alto fueron $\mathbf{X_2; X_4; X_5}$, con el uso de la tabla de todas las regresiones posibles pretendemos hacer la siguiente prueba de hipótesis:

¿Por qué ese formato?

$$\boxed{\begin{cases} H_0: \beta_2 = \beta_4 = \beta_5 = 0 \\ H_1: Algun\ \beta_j\ distinto\ de\ 0\ para\ j = 2,4,5 \end{cases}}$$

Tabla 5. Resumen tabla de todas las regresiones

	SSE	Covariables
Modelo completo	51.025	X_1, X_2, X_3, X_4, X_5
Modelo reducido	58.098	X_1, X_3

Teniendo en cuenta los datos de la **Tabla 5**, podemos armar un modelo reducido para la prueba de significancia del subconjuto:

$$\mathbf{Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \epsilon_i ; \epsilon_i \sim N(0,\sigma^2).\ iid}$$

i=1, 2, ..., 60

2.2 Cálculo y análisis del estadístico de prueba

$$\mathbf{F_0} = \frac{\frac{SSE(\beta 0, \beta 1, \beta 3) - SSE(\beta 0, \beta 1, \beta 2, \beta 3, \beta 4, \beta 5)}{3}}{MSE(\beta 0, \beta 1, \beta 2, \beta 3, \beta 4, \beta 5)} ; \mathbf{F_0 \sim f_{3,54}}$$

Por lo tanto

3 pt

$$\mathbf{F_0} = \frac{\frac{58.098 - 51.0251}{3}}{0.944909} = \mathbf{2,4952}$$

F

Ya teniendo este resultado, obtenemos un valor crítico de la distribución $F_{3,54}$ con un nivel de significancia **$\alpha = 0.05$** , que es igual a $F_{0.05,3,54}$ **= 2.7758.**

2pt

Como F_0 = **2,4952** < $F_{0.05,3,54}$ = **2.7758**, entonces se acepta **H_0** por cual el subconjunto no es significativo, quiere decir que es posible descartar las variables del subconjunto, se concluye que no hay suficiente evidencia para decir que al menos una de las variables **X_2, X_4** y **X_5** tiene un efecto significativo en la variable de respuesta **Y_i**. En este caso, se ~~debe~~ utilizar el modelo reducido que solo incluye las variables **X_1, X_3** y la constante para hacer predicciones.

→ No se debe, se puede

Pregunta 3.

1 pt

¿El valor de la ~~variable~~ parámetro $\boldsymbol{\beta}_1$ es equivalente al valor de $\boldsymbol{\beta}_2$ y el valor de $\boldsymbol{\beta}_3$ es equivalente al valor de $\boldsymbol{\beta}_4$ en el modelo de RLM?

Para resolver, se tiene la siguiente prueba de hipótesis:

$$H_0: \begin{cases} \beta_1 = \beta_2 \\ \beta_3 = \beta_4 \end{cases} \quad vs \quad H_1: \begin{cases} \beta_1 \neq \beta_2 \\ \beta_3 \neq \beta_4 \end{cases}$$

ó

o en forma matricial:

$$H_0 : \boldsymbol{L}\underline{\beta} = \underline{\mathbf{0}} \quad \text{vs.} \quad H_1 : \boldsymbol{L}\underline{\beta} \neq \underline{\mathbf{0}},$$

1 pt

donde tenemos que:

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

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

por lo tanto Ho se puede plantear como:

$$H_0: \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

congruentes con el error, sin embargo L debe tener todos los β's del modelo, así como incluyendo a β0, también

El modelo reducido está dado por:

$$\mathbf{Y_i = \beta_0 + \beta 1(X_{1i} + X_{2i}) + \beta_3 (X_{3i} + X_{4i}) + \epsilon_i \,;\; \epsilon_i \sim N(0,\sigma^2).\ iid\ 1 <= i <= 60}$$

β5 0pt

¿Y β5? ¿Lo sacaron definitivamente del modelo o qué pasó ahí?

Para simplificar se tiene que:

$$\mathbf{Y_i = \beta_0 + \beta 1 X^*_{1i} + \beta_3 X^*_{3i} + \epsilon_i \,;\; X^*_{1i} = X_{1i} + X_{2i} \;,\; X^*_{3i} = X_{1i} + X_{2i}}$$

la expresión del estadístico de prueba es:

0pt

$$F_0 = \frac{MSH}{MSE} = \frac{SSH/2}{MSE} = \frac{(SSE(RM) - SSE(FM))/2}{MSE} \sim F_{2,54}$$

$$\mathbf{F_0} = \frac{SSE(RM) - 25{,}51}{0.944909} \cancel{\sim F_{2,45}}$$

¿Por qué 45 g.l MF? Su modelo Full tiene n=60 y p=6

Pregunta 4.

17,5 pt

4.1 Normalidad de los residuales

4 pt

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

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

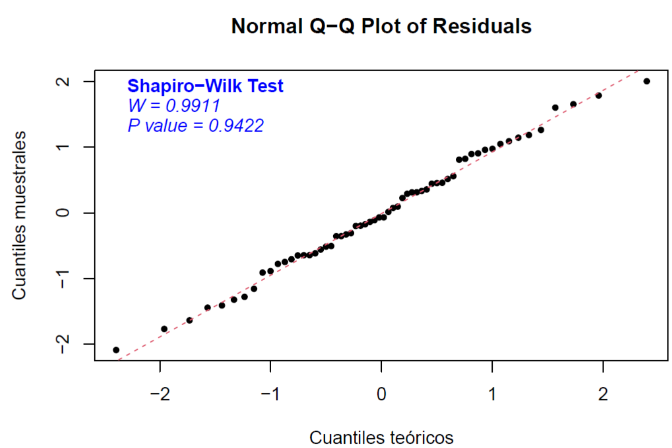


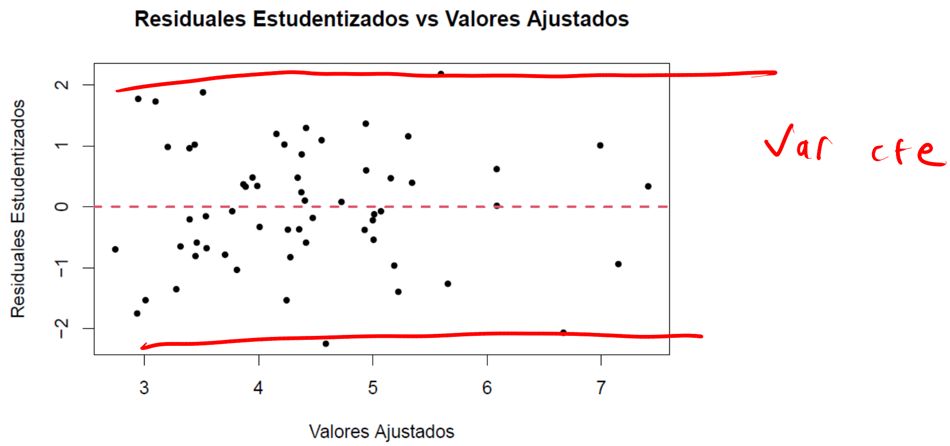
Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el **valor P** aproximadamente igual a **0.9422** y teniendo en cuenta que el nivel de significancia **α = 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 con media 0 y varianza $\boldsymbol{\sigma}^2$ constante, esto es soportado además por el gráfico de comparación de cuantiles, se puede observar en la figura anterior que las colas no son tan pesadas con respecto a la diagonal trazada. Se determina por aceptar el cumplimiento de este supuesto.

4.2 Varianza constante

3pt

Ahora se validará si la varianza cumple con el supuesto de ser constante. Realizaremos el análisis a través de un gráfico de residuales vs. valores ajustados en donde buscaremos patrones en la nube de puntos generada.



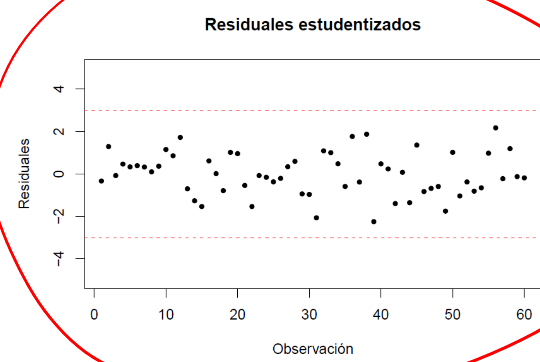
En el gráfico de Residuales Estudentizados vs Valores Ajustados se puede observar que no hay patrones en los que se note que la varianza aumenta o decrece, ni un comportamiento que permita descartar una varianza constante, al no haber evidencia suficiente en contra de este supuesto se acepta como cierto. Además es posible observar media 0.

4.3 Datos atípicos

3pt

Se propone ahora un gráfico de residuales estudentizados que nos permita observar la existencia de valores atípicos.

¿por q' tan pequeño?

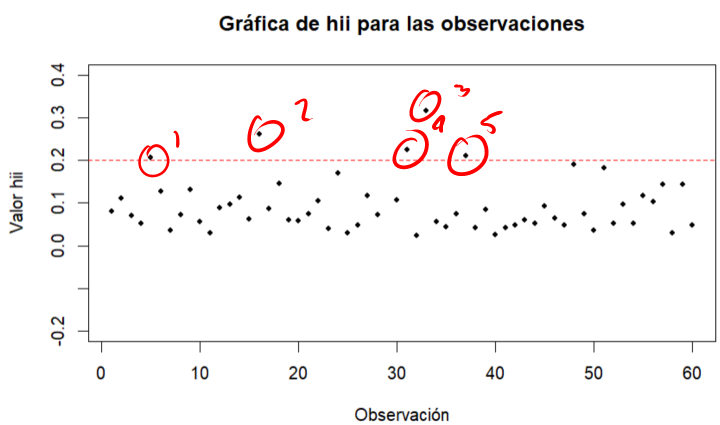


Como se puede observar en la gráfica anterior, no hay datos atípicos en el conjunto de datos pues ningún residual estudentizado sobrepasa el criterio de $|r_{estud}| > 3$.

4.4 Puntos de balanceo

2 pt

Para buscar los puntos de balanceo, que son aquellos tal que $h_{ii} > 2\frac{p}{n} = 0.2$, nos apoyamos del gráfico de las observaciones versus los valores h_{ii}



tienen 6 puntos de balanceo y en la gráfica sólo se ven 5, hii del dato 29 = 0,4959, debían ampliar límite superior a 0,5

Se identifica de la **tabla 6** y del gráfico anterior, que las observaciones **5**, 16, **29**, **31**, 33 y **37** son puntos de balanceo. Estos son puntos alejados de los valores predictores, que pueden ser parte de la causa de nuestro pequeño coeficiente de determinación.

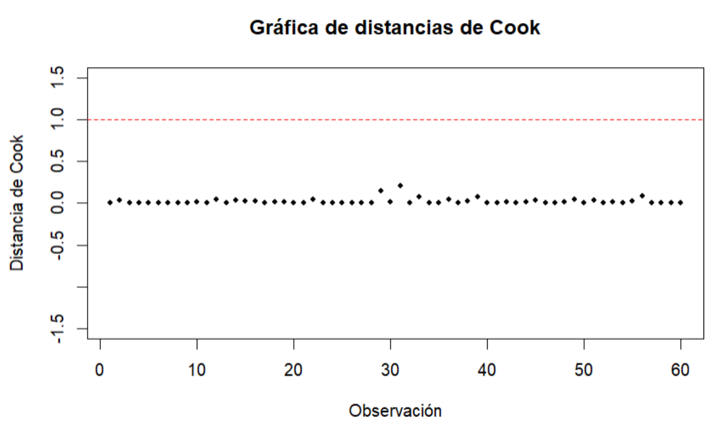
¿Qué más causan estos puntos de balanceo?

Tabla 6. listado de puntos de balanceo

	res.stud	Cooks.D	hii.value	Dffits
5	0,3346	0,0049	0,2064	0,1692
16	0,6176	0,0224	0,2609	0,3648
29	-0,9406	0,1450	0,4959	-0,9319
31	-2,0668	0,2072	0,2254	-1,1511
33	1.0066	0.0779	0.3156	0.6837
37	-0,3802	0,0064	0,2111	-0,1951

4.5 Puntos influyentes

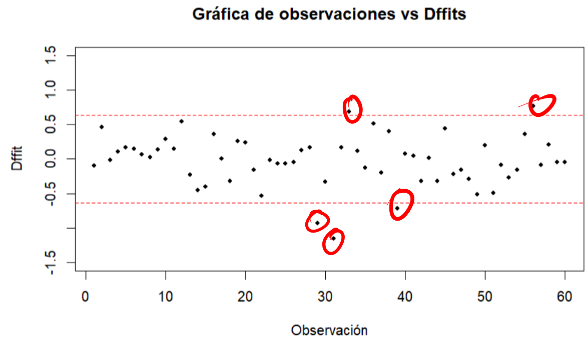
Para buscar puntos influyentes usaremos los criterios de las Distancias de Cook $\left(D_i > 1\right)$ y **DFFITS** $\left|DFFITS_i\right| > 2\sqrt{\frac{p}{n}}$.



Usando distancias de cook, en el gráfico anterior observamos que no hay $D_i > 1$, entonces por medio de este criterio no se encontraron puntos influyentes. ✓ 2pr

Para el criterio de **DFFITS,** $\left|DFFITS_i\right| > 2\sqrt{\frac{p}{n}} = 0.6324$, el gráfico nos muestra la existencia de algunos puntos ~~influenciables~~:

influenciales



1,5 pt

expliquen bien que lo que causan es mover un vector

Son 5 puntos que corresponden a las observaciones **29, 31, 33, 39** y **56** que son puntos los cuales arrastran el modelo hacia su dirección. A continuación se muestra en la tabla 7 los valores influenciales hallados:

Tabla 7. listado de puntos de influenciables (Criterio DFFITS)

	res.stud	Cooks.D	hii.value	Dffits
29	-0,9406	0,145	0,4959	-0,9319
31	-2,0668	0,2072	0,2254	-1,1511
33	1,0066	0,0779	0,3156	0,6837
39	-2,247	0,0786	0,0854	-0,7145
56	2,1769	0,0909	0,1032	0,7658

4.6 Conclusiones del modelo 2 pt

¿dado q e...?

Del análisis de los supuestos podríamos concluir que el modelo, dado que tiene un $\mathbf{R}^2$ (explica aproximadamente el 56% de la variabilidad del modelo), los errores cumplen ~~de una manera positiva~~ los supuestos, pero hay que tener en cuenta que los valores influenciales dado que arrastran el modelo hacia su dirección, podrían "desviar" el modelo de una mejor aproximación, como también podría ser algo “natural” del experimento estos puntos. Esto, se podría observar comparando modelos con y sin los puntos influenciales.

¿Es válido o no?