3,45

Trabajo 1

Estudiantes

Rivera Bedoya Marhia Camila
Sanchez Vivas Harold Jhoan
Valderrama Posada Mateo
Tora Arroyave Oscar Julian

Docente

Francisco Javier Rodriguez Cortes

Asignatura

Estadística II

Sede Medellín
25 de Marzo de 2023

Índice

Índice de figuras

Índice de tablas

1. Pregunta 1

15 pt

Estime un modelo de regresión lineal múltiple que explique el riesgo de infección en términos de las variables restantes (actuando como predictoras) Analice la significancia de la regresión y de los parámetros individuales. Interprete los parámetros estimados. Calcule e interprete el coeficiente de determinación múltiple R^2

Teniendo en cuenta la base de datos asignada a nuestro equipo, la cual es **Equipo45.txt**, las variables para el modelo son

Y RI Riesgo de infección en porcentaje: Probabilidad promedio estimada de adquirir infección en el hospital.

X1 DE Duración de la estadía en días: Duración promedio de la estadía de todos los pacientes en el hospital.

X2 RC Rutina de cultivos: Razón del número de cultivos realizados en pacientes sin síntomas de infección hospitalaria, por cada 100 pacientes.

X3 NCP Número de camas: Promedio de camas en el hospital durante el periodo del estudio.

X4 CPD Censo promedio diario: Número promedio de pacientes en el hospital por día durante el periodo del estudio.

X5 ENF Número de enfermeras: Promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio.

El modelo que se propone es:

$$RI_i = \beta_0 + \beta_1 DE_i + \beta_2 RC_i + \beta_3 NCP_i + \beta_4 CPD_i + \beta_5 ENF_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

1.1. Modelo de regresión

Al ajustar el modelo de regresion para el riesgo de infeccion de una enfermedad en un hospital, se obtienen los siguientes coeficientes:

Tabla 1: Tabla de valores de los coeficientes estimados

	Valor del parámetro
$\hat{\beta}_0$	-1.68786
$\hat{\beta}_1$	0.19317
$\hat{\beta}_2$	0.03430
$\hat{\beta}_3$	0.04258
$\hat{\beta}_4$	0.01994
$\hat{\beta}_5$	0.00069

2 pt

no va en ec. ajustada

Por lo que el modelo con los respectivos valores de los parametos es:

$$\widehat{RI}_i = -1.68786 + 0.19317 DES_i + 0.0343 RC + 0.04258 NC_i + 0.01994 CPD_i + 6.9 \times 10^{-4} NENF_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Donde las variables se mueven deacuerdo $1 \leq i \leq 60$

1.2. Significancia de la regresión

2,5 pt

Se Plantea el siguiente Juego de Hipotesis

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_a : \text{Algún } \beta_j \neq 0 \; para \; j = 1, 2, 3, 4, 5 \end{cases}$$

¿Estadístico de prueba?

Se utilizará la siguiente tabla ANOVA para evaluar la significancia de la regresión:

Tabla 2: Tabla anova significancia de la regresión

	Sumas de cuadrados	g.l	Cuadrado medio	F_0	Valor-P
Modelo de regresión	67.3216	5	13.46431	11.8112	9.69569e-08
Error	61.5578	54	1.13996		

Los resultados obtenidos de la Tabla Anova indican que la hipótesis nula debe ser rechazada de lo cual podemos concluir que el modelo al menos alguna de las variables es significativa

1.3. Significancia de los parámetros

$$\begin{cases} \mathrm{H}_0 : \beta_j = 0 \\ \mathrm{H}_a : \text{Algún } \beta_j \neq 0 \ para \ j = 1, 2, 3, 4, 5 \end{cases}$$

La tabla a continuación muestra los criterios utilizados para evaluar la significancia de los parámetros de forma individual:

Tabla 3: Resumen de los coeficientes

	Estimación β_j	$se(\hat{\beta}_j)$	T_{0j}	Valor-P
β_0	-1.6879	1.8239	-0.9254	0.3589
β_1	0.1932	0.0846	2.2832	0.0264
β_2	0.0343	0.0354	0.9698	0.3365
β_3	0.0426	0.0147	2.8922	0.0055
β_4	0.0199	0.0082	2.4384	0.0181
β_5	0.0007	0.0008	0.8141	0.4192

Los resultados de las pruebas: valor del estadístico de prueba y el valor p para la prueba se obtiene en las dos últimas columnas de la tabla de los parámetros estimados.

Con un nivel de significancia$\alpha = 0.05$ se concluye que los parámetros individuales β_1,β_3,β_4 son significativos cada uno en presencia de los demás parametros Por el contrario los parametros β_0,β_2,β_5 individualmente no son significativos en presencia de los demas parametros

1.4. Interpretación de los parámetros

A continuacion se hará la interpretacion de los parametros que son significativos, ya que los otros parametros no tiene interpretacion y no aportan al modelo:

- $\hat{\beta}_1$= 0.19317:Si todas las demás variables predictoras se mantienen iguales, un aumento de un día en la Duración de la estancia en el hospital daría como resultado un aumento esperado en el promedio del Riesgo de infección en un porcentaje determinado por el valor de 0.19317 %.
- $\hat{\beta}_3$= 0.04258:Si el número promedio de camas en el hospital durante el periodo de estudio aumenta en una unidad, manteniendo constantes las demás variables predictoras, se espera que el promedio del Riesgo de infección se incremente en un 0.04258 %

- $\hat{\beta}_4$= 0.01994: si el número censo del promedio Diario del paciente en el hospital durante el periodo de estudio se incrementa en una unidad, cuando las demas variables se mantienen constantes, se espera que el promedio del Riesgo de infección aumenta en un 0.01994 %

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un R^2 de 0.5224 lo cual significa que aproximadamente el 52.24 % de la variabilidad total en el porcentaje de Riesgo de infeccion es explicado por el modelo RLM

1.6. Comentarios

En el modelo de regresión, se puede notar que las variables que contribuyen significativamente son la Duración de la estadía en el hospital, el Censo promedio diario de pacientes en el hospital y el número de camas. Esto se refleja en la importancia de los parámetros.

2. Pregunta 2

Use la tabla de todas las regresiones posibles, para probar la significancia simultánea del subconjunto de tres variables con los valores p más grandes del punto anterior. Según el resultado de la prueba es posible descartar del modelo las variables del subconjunto? Explique su respuesta.

2.1. Planteamiento prueba de hipotesis y modelo reducido

Los parametros cuyos valores P fueron los más altos corresponden a β_2 con VP=0.3365, β_5 con VP= 0.4192, β_1 con VP= 0.02634. Por tanto, se plantea la siguiente prueba de hipótesis:

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

El modelo completo es el definido en la sección 1.1, y el modelo reducido es:

$$\text{MR: } Rinf_i = \beta_0 + \beta_3 NCP_i + \beta_4 PD4_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Se presenta la siguiente tabla con el resumen de todas las regresiones para plantear el estadístico de prueba:

Tabla 4: Resumen de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	61.558	X1 X2 X3 X4 X5
Modelo reducido	73.462	X3 X4

2.2. Estadístico de prueba y conclusiones

Se construye el estadístico de prueba como:

$$F_0 = \frac{(SSR(\beta_0, \beta_3, \beta_4 | \beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,54}$$

$$F_0 = \frac{(SSE(\beta_0, \beta_3, \beta_4) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5))/2}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)} \overset{H_0}{\sim} f_{2,54}$$
$$= \frac{(73.462 - 61.558)/2}{61.558/54} = 5.2212$$

Cuando comparamos F_0 con $f_{0.05,2,54} = 3.168246$ a un nivel de significancia de $\alpha = 0.05$, y usamos un valor p de 0.0084527, vemos que el valor p es pequeño, lo que sugiere que debemos rechazar la hipótesis nula H_0. Por lo tanto, llegamos a la conclusión de que no se puede descartar este subconjunto de datos del modelo.

3. Pregunta 3

Plantee una pregunta donde su solución implique el uso exclusivo de una prueba de hipótesis lineal general de la forma H0 : L$\beta = 0$ (solo se puede usar este procedimiento y no SSextra). Especifique claramente la matriz L, el modelo reducido y la expresión para el estadístico de prueba (no hay que calcularlo).

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \text{H}_0 : \beta_4 = \beta_5, \beta_2 = \beta_3 \\ \text{H}_a : \text{Alguna de las desigualdades no se cumple} \end{cases}$$

Reescribendo matricialmente:

$$\begin{cases} \text{H}_0 : \mathbf{L}\underline{\beta} = 0 \\ \text{H}_a : \mathbf{L}\underline{\beta} \neq 0 \end{cases}$$

Donde **L** está dada por:

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

Donde el modelo reducido está dado por:

$$RI = \beta_0 + \beta_1(DES_i) + \beta_2(RC_i + NCP_i) + \beta_4(CPD_i + ENF_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

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

Obteniendo esto podemos definir la region de rechazo de la hipotesis nula como $F_0 > F_{0.05,2,54} = 3.168246$ y con valor p:P($F_{2,54} > |F_0|$)

4. Pregunta 4

Realice una validación de los supuestos en los errores y examine si hay valores atípicos, de balanceo e influenciales. Qué puede decir acerca de la validez de éste modelo?. Argumente su respuesta.

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Para la validación de este supuesto, se plantea la siguiente prueba de hipótesis (Shapiro-Wilk)

$$\begin{cases} \mathrm{H}_0: & \varepsilon_i \sim N(\mu, \sigma^2) \\ \mathrm{H}_a: & \varepsilon_i \nsim N(\mu, \sigma^2) \end{cases}$$

acompañado de un grafico cuantil-cuantil:

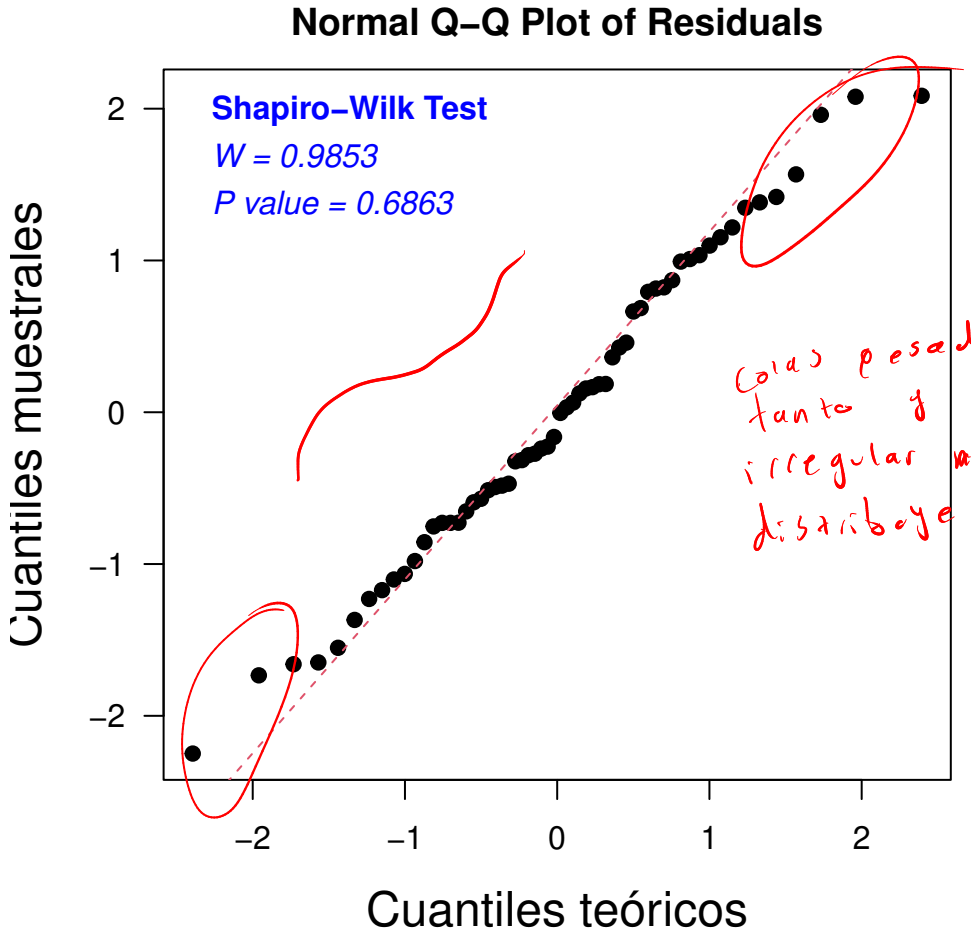


Figura 1: Gráfico cuantil-cuantil y normalidad de los residuales

Dado que el valor de p es alto, se puede concluir que no hay suficiente evidencia para rechazar la hipótesis nula H_0. Por lo tanto, se puede inferir que el modelo es consistente con la suposición de que los residuos siguen una distribución normal.

4.1.2. Media 0 y Varianza constante

En esta prueba se quiere probar

$$H_0: \mathbf{V}[\varepsilon_i] = \sigma^2 \qquad \textbf{vs} \qquad \mathbf{V}[\varepsilon_i] \neq \sigma^2$$

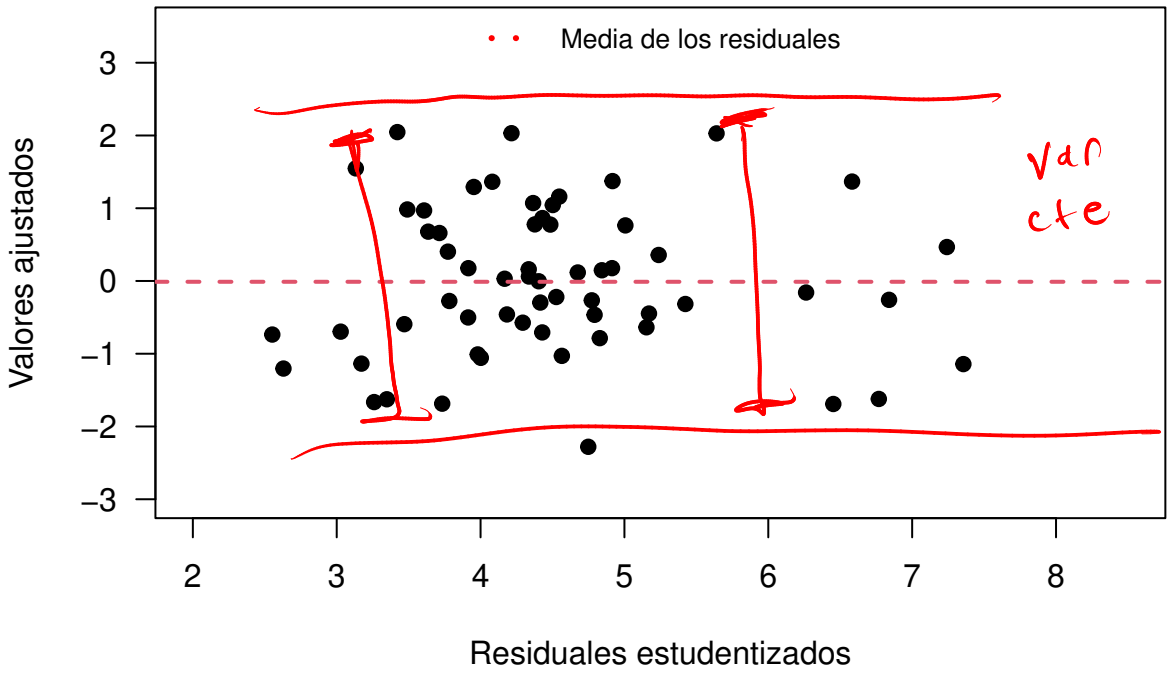


Figura 2: Gráfico residuales estudentizados vs valores ajustados

Se puede observar que la línea punteada roja, que representa la media de los errores, se encuentra cerca o en cero, lo que sugiere que los errores tienen una media cercana a cero. Además, al examinar los residuos, no se puede detectar ningún patrón y se ven uniformemente distribuido, lo que indica que la varianza de los errores es constante a través de todo el rango de los valores observados.

4.2. Observaciones extremas

4.2.1. Datos atípicos

1,5 pt

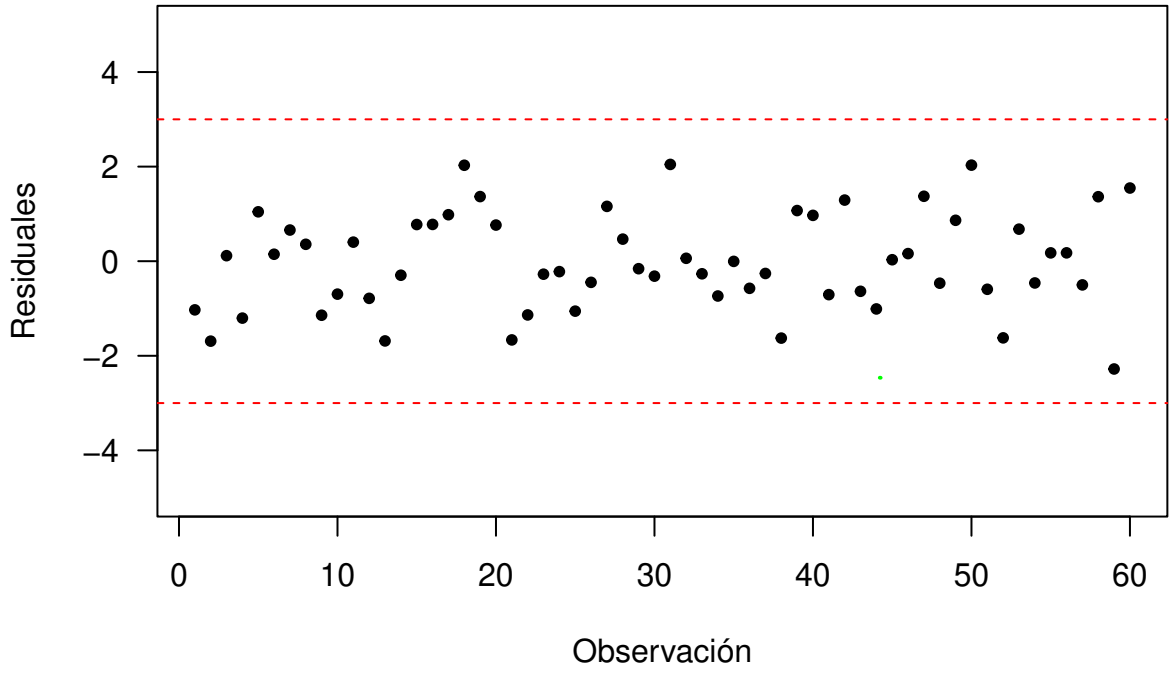


Figura 3: Identificación de datos atípicos

Notese que segun este criterio no existen puntos atipicos que deban ser investigados

Según cuál criterio? qué dice?

4.2.2. Puntos de balanceo

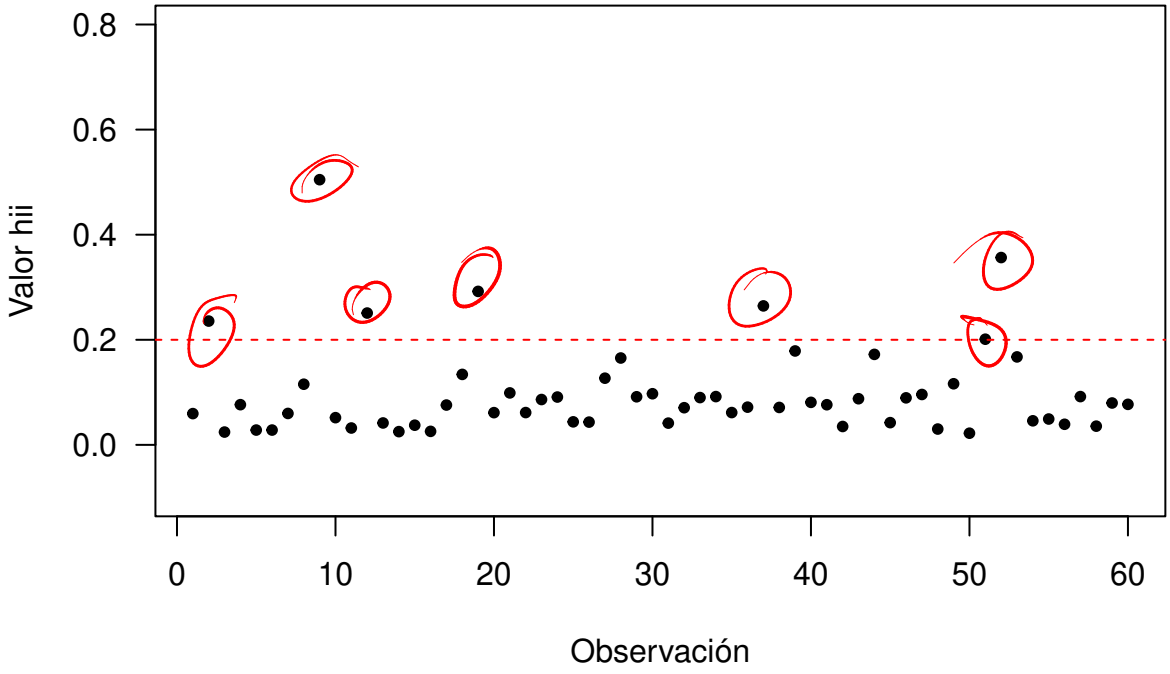


Figura 4: Identificación de puntos de balanceo

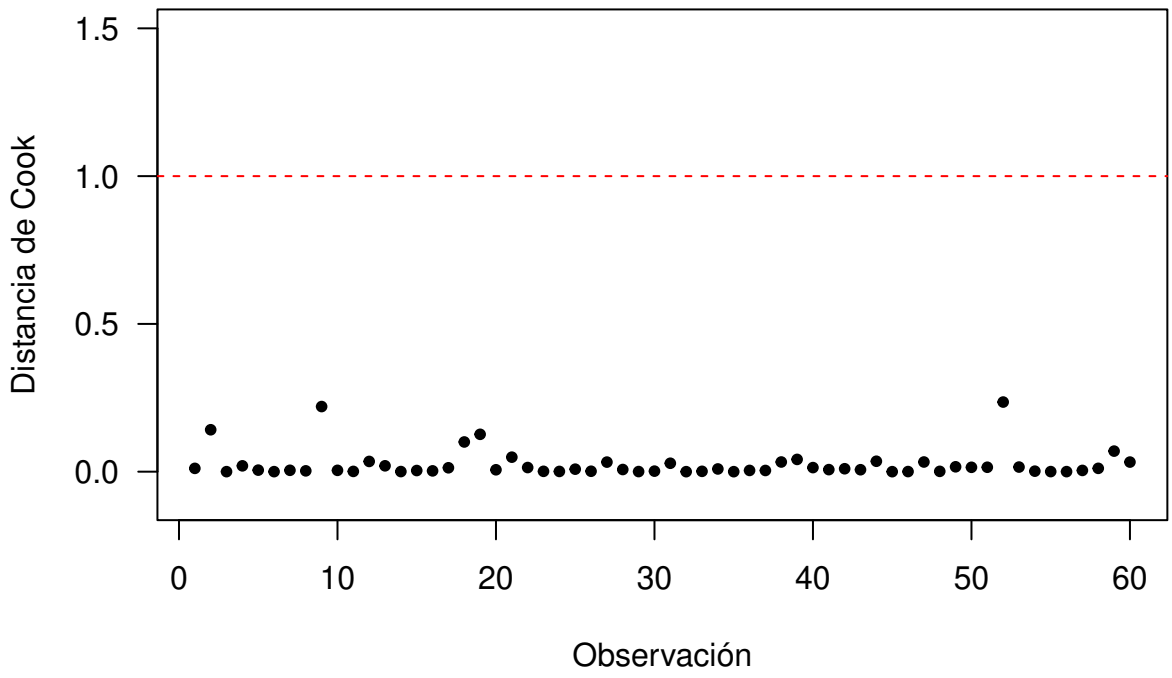
Tabla 5: Tabla de puntos de Balanceo

	Errores Estudentizados	D.Cook	Valor hii	DFFITS
2	-1.6907	0.1419	0.2356	-0.9385
9	-1.1425	0.2204	0.5047	-1.1533
12	-0.7858	0.0347	0.2509	-0.4547
19	1.3664	0.1264	0.2921	0.8777
37	-0.2592	0.0041	0.2645	-0.1555
51	-0.5936	0.0150	0.2011	-0.2978
52	-1.6214	0.2356	0.3565	-1.2067

Es importante destacar que hay 7 datos datos que deben ser estudiados en términos de su impacto en el ajuste del modelo y sus propiedades. Estos datos corresponden a los puntos 2, 9, 12, 19, 37, 51 y 52, ya que son mayores al criterio $\frac{2p}{n}$. Estos puntos de balanceo pueden tener una gran influencia en el modelo

4.2.3. Puntos influenciales

Bajo el criterio de Cook, se hace la siguiente gráfica:



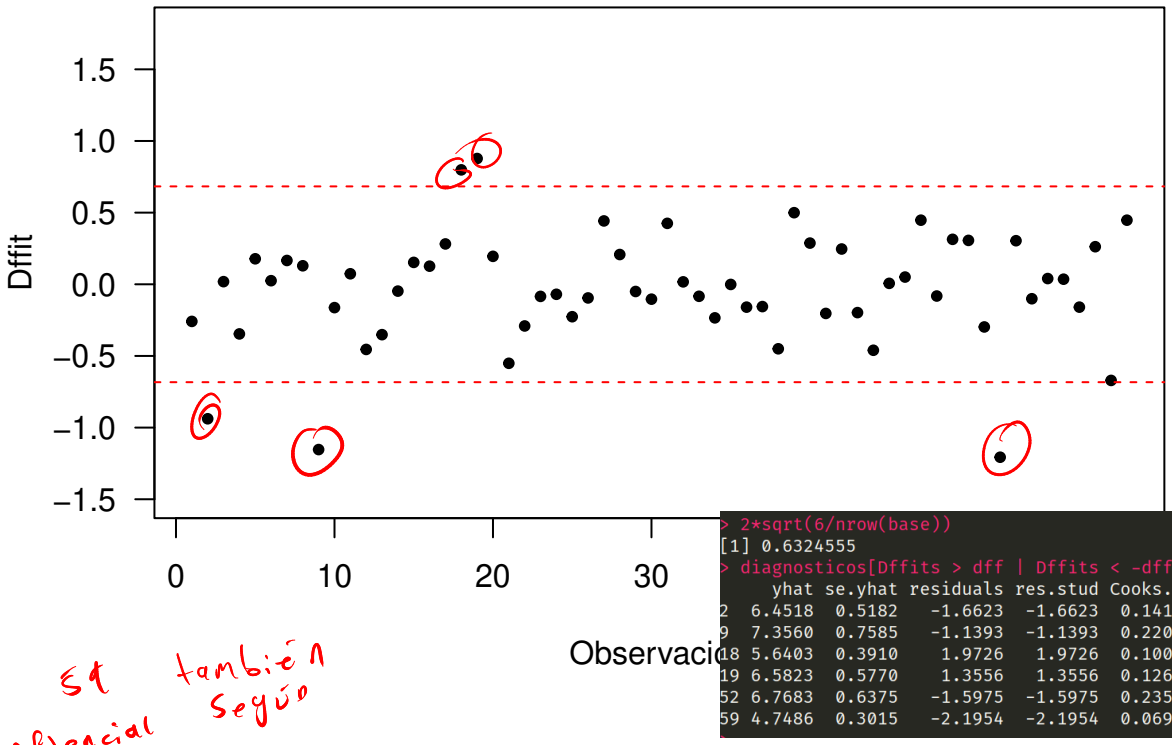
redundantes

1,5 pt

Figura 5: Criterio distancias de Cook para puntos influenciales

Bajo el criterio de cook, se obtuvo la anterior gráfica. A partir de la gráfica podemos concluir que no existen puntos influenciales bajo este criterio

¿Qué dice el criterio?




```
> 2*sqrt(6/nrow(base))
[1] 0.6324555
> diagnosticos[Dffits > dff | Dffits < -dff, ]
    yhat se.yhat residuals res.stud Cooks.D hii.value  Dffits
2  6.4518  0.5182   -1.6623  -1.6623  0.1419    0.2356 -0.9385
9  7.3560  0.7585   -1.1393  -1.1393  0.2204    0.5047 -1.1533
18 5.6403  0.3910    1.9726   1.9726  0.1005    0.1341  0.7985
19 6.5823  0.5770    1.3556   1.3556  0.1264    0.2921  0.8777
52 6.7683  0.6375   -1.5975  -1.5975  0.2356    0.3565 -1.2067
59 4.7486  0.3015   -2.1954  -2.1954  0.0696    0.0797 -0.6708
```

Figura 6: Criterio Dffits para puntos influenciales

Tabla 6: Tabla del criterio DFFITS para encontrar puntos influenciales

	Errores Estudentizados	D.Cook	Valor hii	DFFITS
2	-1.6907	0.1419	0.2356	-0.9385
9	-1.1425	0.2204	0.5047	-1.1533
18	2.0287	0.1005	0.1341	0.7985
19	1.3664	0.1264	0.2921	0.8777
52	-1.6214	0.2356	0.3565	-1.2067

Usando el criterio de Dffits, se ha generado el gráfico anterior, el cual sugiere que hay varios valores influyentes en el modelo. Específicamente, las observaciones 2, 9, 18, 19 y 62 que pueden tener un impacto significativo en el modelo y deben ser investigadas con más detalle.

4.3. Conclusiones

El modelo cumple con los supuestos básicos de regresión lineal de que la media de los residuos es cercana a cero y la varianza es constante. Sin embargo, se observó una gran cantidad de datos de balanceo y puntos influenciales, lo que sugiere la necesidad de investigar si estos datos afectan significativamente el modelo y sus supuestos, incluida la normalidad de los residuos. En resumen, aunque el modelo cumple con los supuestos subyacentes, no podemos decir si el modelo es adecuado para hacer predicciones, y se requieren análisis

adicionales para evaluar el impacto de los datos de equilibrio e influencia para determinar si el modelo es un modelo preciso

Modelo válido o no?