4,3

Trabajo 1

Estudiantes

Brian Alexander Guerrero
Kevin Daniel Guio Covilla
Sebastián Soto Arcila

Equipo #44

Docente

Francisco Javier Rodrigez 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

17,5 pt

Teniendo en cuenta la base de datos del equipo 44, en la cual hay 5 variables regresoras y una variable dependiente, denominadas por:

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.

Al plantear el modelo de regresión lineal múltiple queda como:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i;\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2);\ 1 \leq i \leq 55$$

1.1. Modelo de regresión

1,5 pt

Los valores de los coeficientes, luego de ajustar el modelo se muestran en la siguiente tabla:

Cuadro 1: Valores aproximados de los coeficientes del modelo

	Valor ajustado del parámetro
β_0	-1.1564
β_1	0.0884
β_2	0.0391
β_3	0.0308
β_4	0.0197
β_5	0.0016

Con estos valores podemos escribir el modelo de regresión ajustado, queda de la siguiente manera:

¿no va en ec. ajustada

$$\hat{Y}_i = -1.1564 + 0.0884X_{1i} 0.0391X_{2i} + 0.0308X_{3i} + 0.0197X_{4i} + 0.0016X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant \boxed{50}$$

55

1.2. Significancia de la regresión

4 pt

Para analizar la significancia de la regresión nos basamos en la siguiente hipótesis:

$$\begin{cases} H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1: \text{Algun } \beta_j \neq 0; \text{ j=1, 2, 3, 4, 5} \end{cases}$$

$F_0 = \frac{MSR}{MSE} \neq \frac{MST}{MSE}$

Y rechazamos o aceptamos con el estadístico de prueba:

$$F_0 = \frac{\cancel{MST}}{MSE} \overset{H_0}{\sim} f_{5,49} \tag{1}$$

Para definir la significancia del modelo podemos también hacer uso de el valor P y compararlo con una significancia $\alpha = 0.05$ que obtenemos de la tabla Anova:

Cuadro 2: ANOVA del modelo

	Sumas Cuadrática	Grados de libertad.	Media Cuadrática	F_0	Valor-P
Regresión	39.2257	5	7.84514	7.80019	1.83603e-05
Error	49.2823	49	1.00576		

Como podemos observar, el valor P es muy pequeño, cercano a 0 lo cual nos indica que debemos rechazar la hipótesis nula aceptando así la hipótesis alternativa, evidenciando que existe una relación de regresión, es decir, podemos decir que la regresión es significativa, sin embargo, aún no podemos garantizar que el modelo sea útil para hacer predicciones.

Como la regresión es significativa sabemos que al menos un parámetro es significativo, cosa que comprobaremos a continuación.

1.3. Significancia de los parámetros

Cap1

Para analizar la significancia de los parametros del modelo utilizamos la siguiente hipótesis:

$$\begin{cases} H_0: \beta_j = 0; \text{ j=}\boxed{\text{1, 2, 3, 4, 5}} \\ H_1: \beta_j \neq 0; \text{ j=}\boxed{\text{1, 2, 3, 4, 5}} \end{cases}$$

$\to$ y β_0 ?

Y rechazamos o aceptamos con el estadístico de prueba que tenemos a continuación:

$$T_{j,0} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \overset{bajo H_0}{\sim} t_{49}; \text{ j=1, 2, 3, 4, 5} \tag{2}$$

La siguiente tabla organiza la información de cada parámetro y es necesaria para determinar cuáles de ellos son significativos:

Cuadro 3: Resumen de los coeficientes

	Valor Ajustado del parámetro	$se(\hat{\beta}_j)$	T_{0j}	Valor-P
β_0	-1.1564	2.0138	-0.5742	0.5684
β_1	0.0884	0.1240	0.7125	0.4795
β_2	0.0391	0.0354	1.1030	0.2754
β_3	0.0308	0.0156	1.9712	0.0544
β_4	0.0197	0.0088	2.2408	0.0296
β_5	0.0016	0.0008	2.0694	0.0438

El Valor-P de cada parámetro presente en la tabla permite concluir con un nivel de significancia $\alpha = 0.05$ que, como se acepta la hipótesis nula para los parámetros $\beta_0, \beta_1, \beta_2$ y β_3 entonces, los únicos parámetros que resultan ser significativos son β_4 y β_5 ya que rechazan la hipótesis nula, pues sus Valores-P son menores a α.

1.4. Interpretación de los parámetros 3pt

$\hat{\beta}_4$: Por cada aumento en el número promedio de pacientes en el hospital, la probabilidad promedio de adquirir infección en el hospital aumenta en 0.0197 unidades cuando las demás variables se mantienen constantes.

$\hat{\beta}_5$: Por cada aumento en el número promedio de enfermeras, la probabilidad promedio de adquirir infección en el hospital aumenta en 0.0016 unidades cuando las demás variables se mantien constantes.

1.5. Coeficiente de determinación múltiple R^2 3pt

El coeficiente de determinación múltiple del modelo $R^2 = 0.4432$, indica que la varibilidad total observada de la respuesta explicada por la regresion es aproximadamente 44.32 %. Sin embargo como medida de bondad de ajuste se prefiere usar el R^2 ajustado, ya que penaliza al modelo por la cantidad de variables incluidas, al contrario del R^2 que no decrece cuando existen variables que no aportan significativamente. El valor del R^2 ajustado es $= 0.3864$, que similarmente en el caso de R^2, significa que aproximadamente el 33.64 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto.

¿cómo se calculan?

2. Pregunta 2 4pt

2.1. Planteamiento prueba de hipótesis y modelo reducido

Las 3 covariables con Valor-P más grande son $X_1, X_2, \text{y} X_3$, por lo tanto a través de la tabla de todas las regresiones posibles se hace la siguiente prueba de hipótesis:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1 : \text{Algun } \beta_j \neq 0; \text{ j=1, 2, 3} \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	49.282	X1 X2 X3 X4 X5
Modelo reducido	55.430	X4 X5

El modelo reducido para la prueba de significancia entonces quedaría de la forma:

$$Y_i = \beta_0 + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i; \;\; \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2); \;\; 1 \leq i \leq 55$$

2pt

2.2. Estadístico de prueba y conclusión

El estadístico de prueba se construye de la siguiente manera:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_4, \beta_5) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4,\; \beta_5))/3}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4,\; \beta_5)} \overset{H_0}{\sim} f_{3,49} \\ &= \frac{(41.663 - 40.739)/3}{1.00576} \\ &= 0.3062 \end{aligned} \quad (3)$$

$$\frac{(55{,}430 - 49{,}282)/3}{49{,}282/49}$$

Ahora, comparando el F_0 con $f_{0.95,3,49} = 2.7939$, se puede ver que $F_0 < f_{0.95,3,49}$ Entonces, el subconjunto de covariables con Valor-P más grande no es significativo, por lo tanto estas variables se pueden descartar del modelo ya que el subconjunto de estas es igual a 0. ~ estadísticamente

2pt

3. Pregunta 3

9pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Las preguntas planteadas son; ¿Existe una relación significativa entre el número promedio de enfermeras equivalentes a tiempo completo y la duración promedio de la estadía de todos los pacientes en el hospital durante el periodo del estudio? y ¿Existe una relación entre el número promedio de camas ocupadas y el número promedio de pacientes por día durante el periodo de estudio? Por tanto la prueba de hipótesis a plantear seria:

es el efecto, no la variable en sí.

$$\begin{cases} \mathrm{H}_0 : \beta_1 = 54\beta_5;\ \beta_3 = 1.6\beta_4;\ \beta_1 = \beta_4 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

Escribiéndolo matricialmente, tenemos:

$$\begin{cases} \mathrm{H}_0 : \mathbf{L}\underline{\beta} = \underline{\mathbf{0}} \\ \mathrm{H}_1 : \mathbf{L}\underline{\beta} \neq \underline{\mathbf{0}} \end{cases}$$

Y la matriz **L** dada por:

2pt

$$L = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & -54 \\ 0 & 0 & 0 & 1 & -1.6 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \end{bmatrix}$$

El modelo reducido es:

$X^*_{3i} = 1{,}6\, X_{3i} + X_{4i}$

$$Y_i = \beta_o + \beta_1 X^*_{1i} + \beta_2 X_{2i} + \beta_3 X^*_{3i} + \beta_4 X^*_{4i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 55$$

0,5 pt

Donde $\boxed{X^*_{1i} = X_{1i} + 54X_{5i}}$ y $\boxed{X^*_{3i} = X_{3i} + 1.6X_{4i}}$ y $X^*_{4i} = X_{1i} + X_{4i}$

$x^*_{1i} = 54 x_{1i} + x_{3i}$

3.2. Estadístico de prueba

1,5 pt

El estadístico de prueba F_0 para comprobar la prueba de hipótesis es:

$$F_0 = \frac{(SSE(MR) - 40.739)/3}{1.00576} \overset{H_0}{\sim} f_{3,49} \tag{4}$$

Fórmula general? RR?

4. Pregunta 4

17,5 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

4pt

Para la validación del supuesto de los residuales, se plantea la siguiente prueba de hipótesis ~~Shapiro-wilk~~ junto con un gráfico de cuantil-cuantil

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

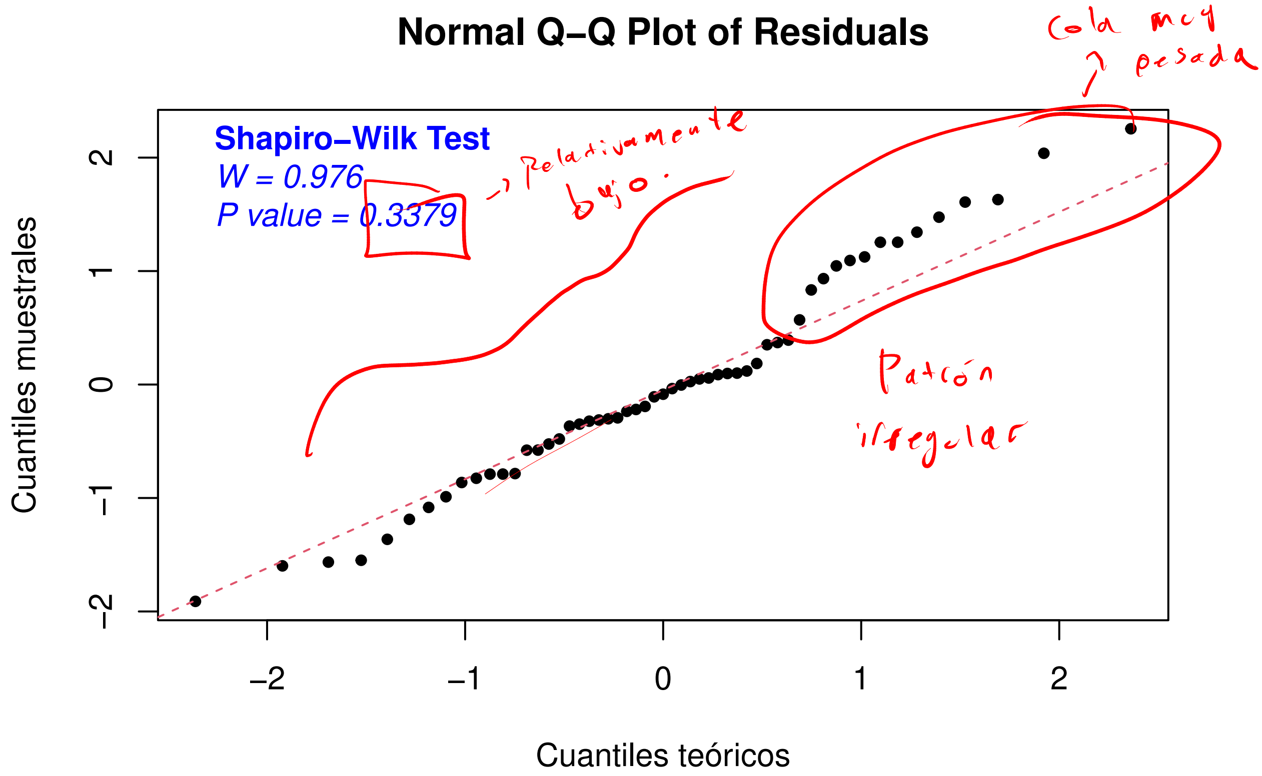

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

y cola pesada

En primer lugar se puede hacer una interpretación general de la gráfica de comparación de cuantiles: al examinar el intervalo de valores entre [-0.5,0.5] se puede observar que los valores parecen seguir una distribución normal, sin embargo en los extremos de la gráfica se evidencian patrones irregulares, lo que sugiere que la distribución no es completamente normal, ahora, si nos en fijamos en el Valor-P que es aproximadamente igual a 0.3379 y teniendo en cuenta un nivel de significancia de $\alpha = 0.05$, como el Valor-P es mucho mayor al alpha, no se rechazaría la hipótesis nula, es decir que los datos distribuyen normal con media $\boxed{\mu = 0}$ y varianza constante σ^2 no obstante al tener más poder el análisis gráfico, se termina por rechazar la hipótesis nula y concluimos que los $\varepsilon_i \not\sim$ Normal.

no están probando eso

con normalidad prueban var cte?

4.1.2. Varianza constante

2pt

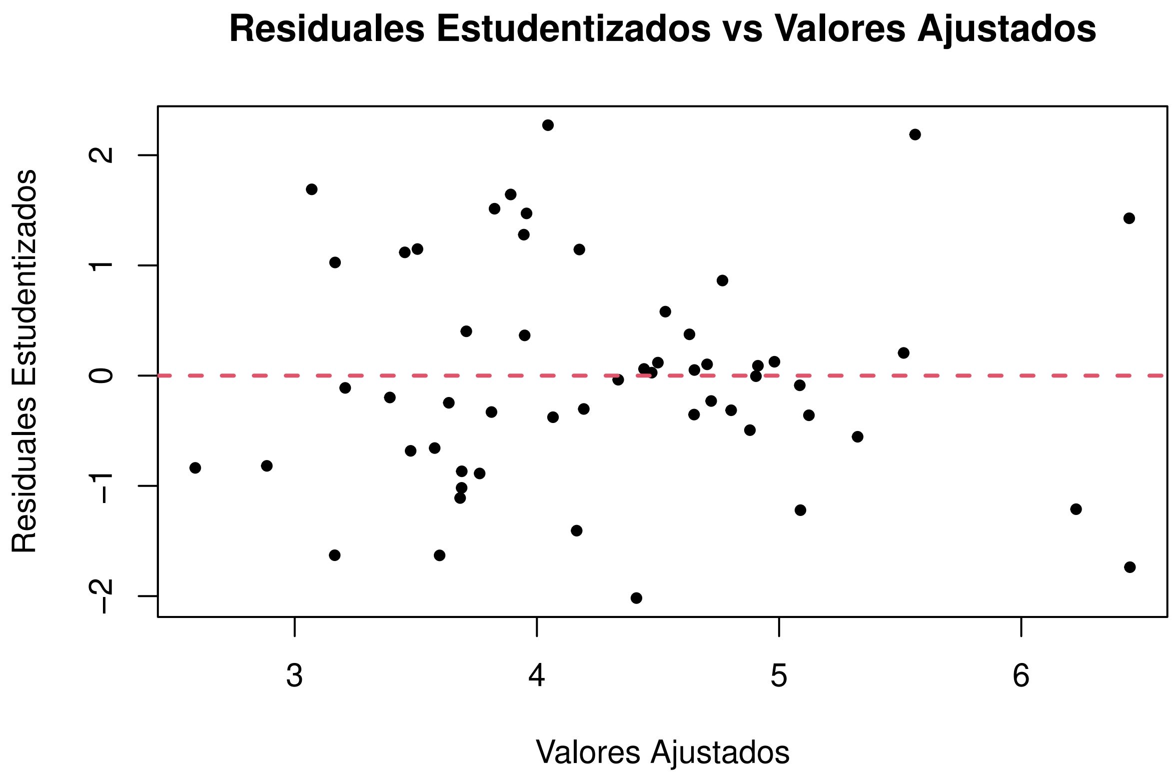

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Al analizar el gráfico de Valores Ajustados vs Residuales Estudentizados se puede observar que entre los intervalos desde [0,4.3) mantienen una varianza alta pero constante, sin embargo en los intervalos siguientes [4.3, 5.5) se puede ver que la varianza deja de ser constante debido a que la variabilidad en los datos se distribuyen de manera aproximadamente uniforme al rededor del valor 0 y en términos generales se puede concluir que el comportamiento de los datos no se asemeja a una varianza constante.

Falso, es debido a que decrece.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

3pt

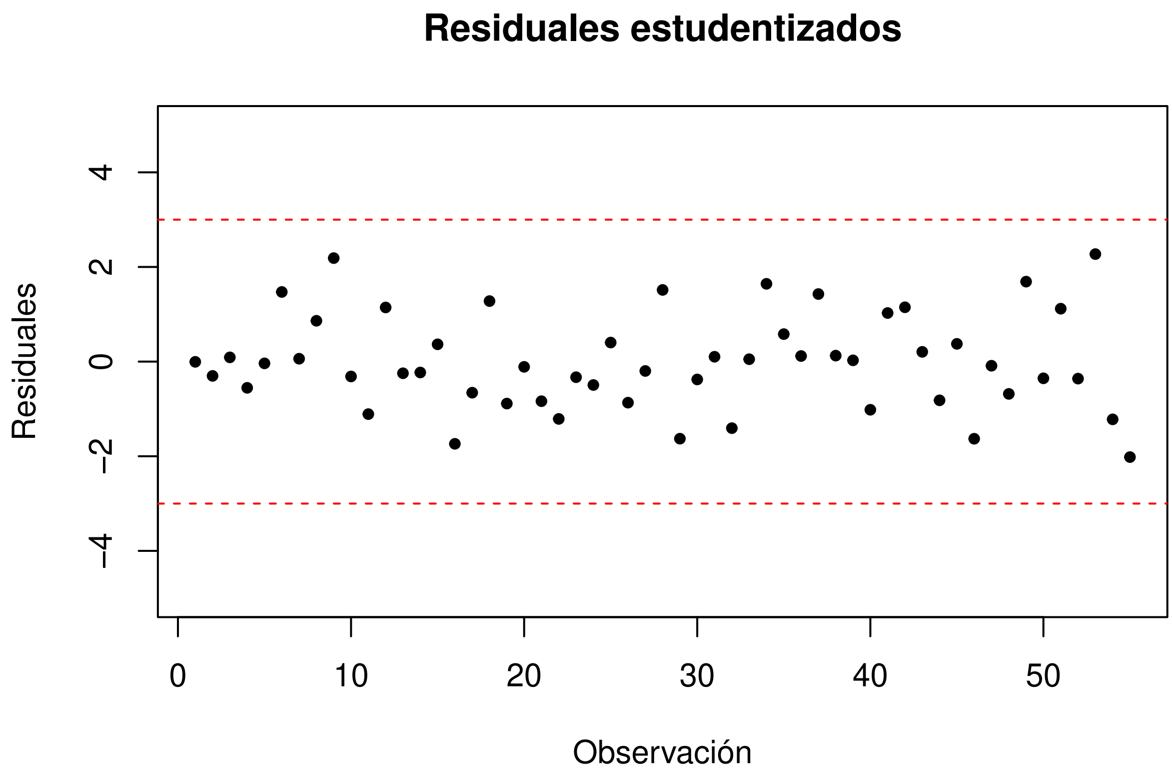

Figura 3: Identificación de datos atípicos

Como se puede evidenciar en la gráfica anterior, ningúna observación de residuales estudentizados sobrepasa el criterio de $|r_{estud}| > 3$ por tanto se puede decir que no existen observaciones malas, ya sea error de registro o error de medición por consiguiente concluimos que no hay datos atípicos o outliner en el modelo.

no se puede decir eso sólo con atípicos

4.2.2. Puntos de balanceo

3pt

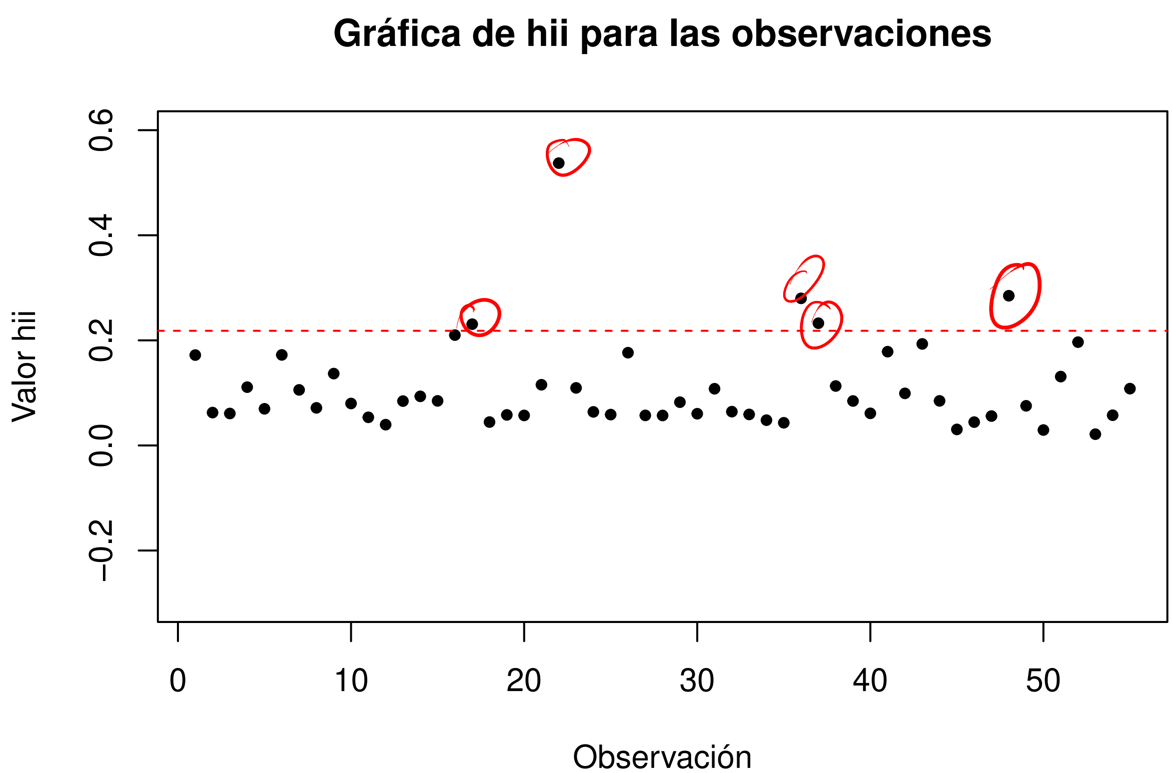

Figura 4: Identificación de puntos de balanceo

Cuadro 5: Resumen tabla para identificar puntos de balanceo

Observaciones	Valores h_{ii}
17	0.2308
22	0.5373
36	0.2799
37	0.2326
48	0.2849

Concluyendo de la gráfica de observaciones vs valores h_{ii}, se evidencia que existen 5 datos del conjunto, los cuales son los presentados en la tabla, que son puntos de balanceo según el criterio $h_{ii} > 2\frac{p}{n}$, en donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 0.2181818$. Estos puntos de balanceo pueden afectar el porcentaje de confianza R^2 del modelo y los errores estándar de los coeficientes estimados.

Excelente

4.2.3. Puntos influenciales

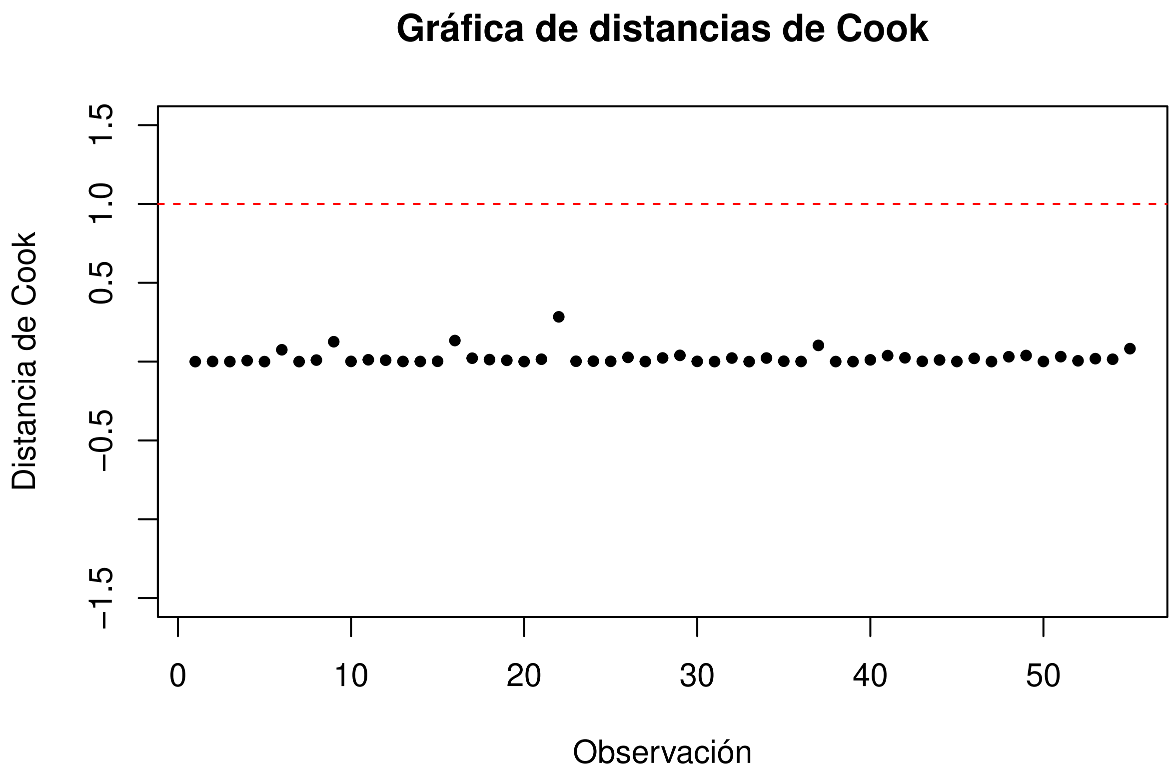

Figura 5: Criterio distancias de Cook para puntos influenciales

Como se puede analizar, la mayoría de los datos observados están por debajo de la línea roja punteada que referencia al criterio $D_i > 1$. No hay un dato que destaque del gráfico alejándose de la tendencia de todas las observaciones, que pueda afectar significativamente los coeficientes de regresión ajustados, es decir no hay observaciones atípicas que sean influyentes de manera individual.

2pt

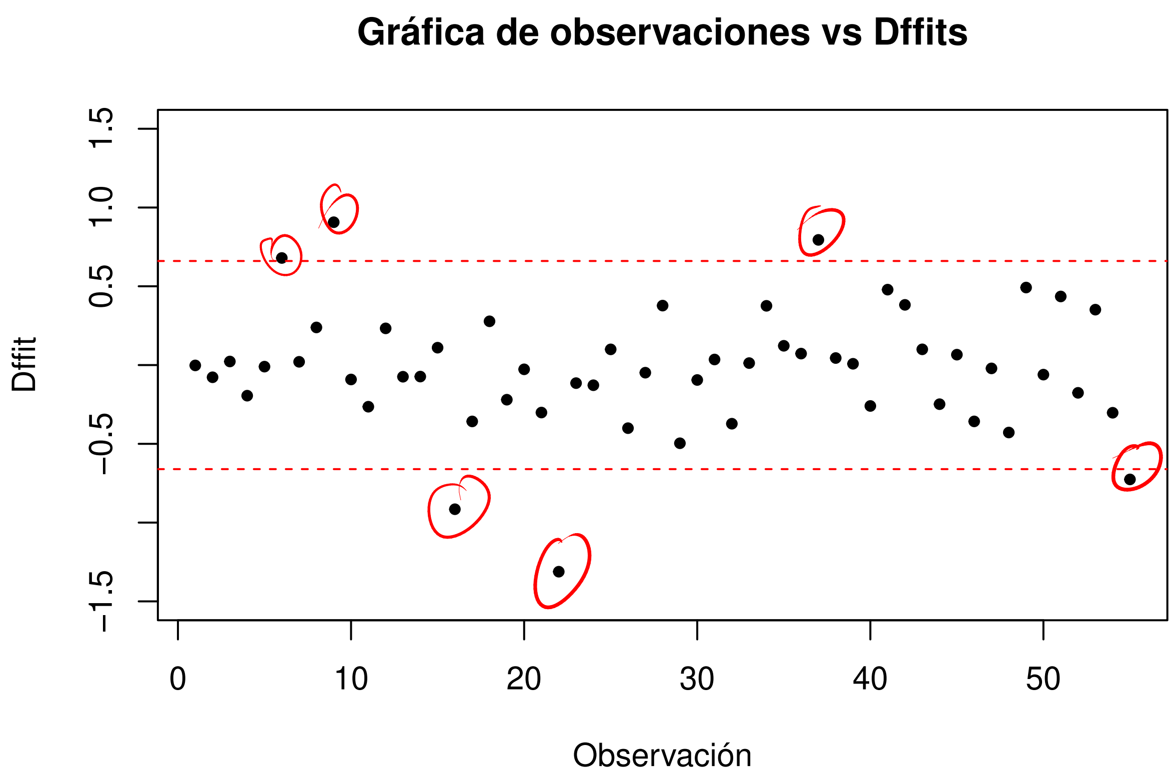

Figura 6: Criterio Dffits para puntos influenciales

Cuadro 6: Resumen tabla para identificar puntos influenciales Dffits

Observaciones	Dffits
6	0.6798
9	0.9067
16	-0.9152
22	-1.3118
37	0.7946
55	-0.7259

1,5 pt

Al observar la gráfica se puede ver hay 6 observaciones que según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 0.6605783$, es un punto influencial y se puede verificar en la tabla. Cabe destacar que también con el criterio de distancias de Cook, se verifican punto influencial pero ninguno de los datos cumple con serlo. Los puntos de influencia pueden afectar tanto por un error de ajuste grande como por un gran balanceo, por eso, los puntos detectados por estos criterios deben ser investigados.

Influenciales según este criterio qué afectan?

un modelo puede tener buen ajuste y no cumplir supuestos.

normalizar es muy distinto a distribuir normal

2 pt

4.3. Conclusión

Podemos decir que los resultados de la regresión indican una asociación significativa entre las variables. Sin embargo, se debe tener en cuenta que los residuos no están normalizados y que la varianza no es constante, lo que sugiere que el modelo puede no ajustarse bien a los datos y que los resultados pueden no ser confiables. Es posible que se requieran técnicas adicionales, como la transformación de variables o la selección de un modelo diferente, para mejorar el ajuste del modelo y garantizar la validez de los resultados y según el análisis realizado, la regresión podría mejorar si se eliminan tres de las cinco variables incluidas en el modelo. → ¿cuáles?

Modelo válido o no?