Trabajo 1

4,45

Estudiantes

Maryory Aguedelo Goez
Vanessa Martínez De Ornelas
María Isabel Carvajal Londoño
David Quiroz García

Equipo #17

Docente

Verónica Guarín Escudero

Asignatura

Estadística II

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medellín
30 de marzo de 2023

Índice

Índice de cuadros

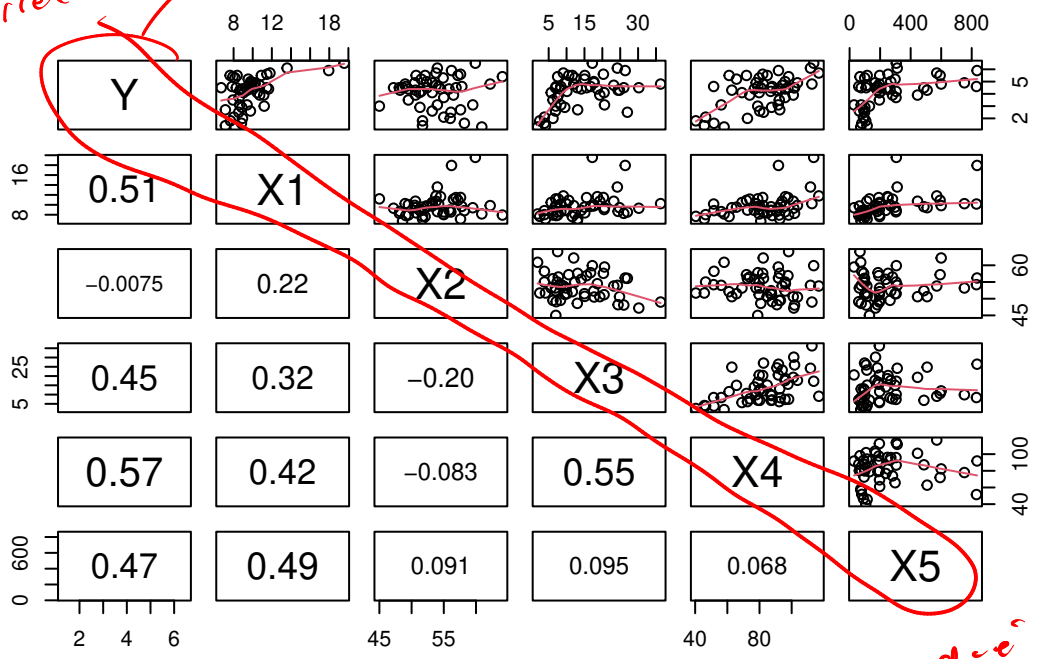
Nota: En el presente trabajo, para cada una de las pruebas de hipótesis que se desarrollan, se usó un nivel de significancia estadística de $\alpha = 0.05$.

1. Pregunta 1

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.

Solución

Inicialmente, puede ser útil realizar chequeos gráficos de la naturaleza de las asociaciones entre las variables predictoras con la variable respuesta. Para esto, haremos uso de una matriz de gráficas de dispersión. La cual nos permite visualizar rápida y simultáneamente estas relaciones. Veamos:



Buscamos con la matriz entender el comportamiento y la relación que puede existir entre las variables regresoras y la variable respuesta. Se espera que no existan relaciones lineales fuertes entra las variables regresoras para no incurrir en problemas de multicolinealidad.

Teniendo claro lo anterior, damos paso al planteamiento de un modelo de RLM:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i, \quad i = 1, 2, \ldots, 50$$

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 y X_5 : Número de enfermeras.

Que tiene como supuestos lo siguiente:

$$\varepsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \quad \forall_{i=1,2,\ldots,50}$$

Agregamos que también es posible especificar el modelo en términos matriciales como sigue a continuación:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{con} \quad \boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

1.1. Estimando los parámetros del modelo de regresión lineal múltiple

Para ello, a través del resumen del modelo, obtenemos la estimación de cada parámetro.

Cuadro 1: Tabla de coeficientes del modelo

	Estimación	Error Estándar	Valor T	Pr(>|t|)
β_0	0.27908	1.94959	0.14315	0.88683
β_1	0.06062	0.07662	0.79113	0.43311
β_2	-0.00016	0.03540	-0.00446	0.99646
β_3	0.02483	0.02035	1.22007	0.22894
β_4	0.02846	0.00919	3.09607	0.00341
β_5	0.00234	0.00076	3.07190	0.00364

Por lo tanto, se llega a la ecuación de regresión ajustada:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2} + \cdots + \widehat{\beta}_5 X_{i5}, \quad i = 1, 2, \ldots, 50$$

$$\widehat{Y}_i = 0.27908 + 0.06062 X_{i1} - 0.00016 X_{i2} + 0.02483 X_{i3} + 0.02846 X_{i4} + 0.00234 X_{i5}, \quad i = 1, 2, \ldots, 50$$

1.2. Analizando la significancia de los parámetros

Para llevar a cabo este proceso se establece el siguiente juego de hipótesis:

$$\begin{array}{l} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{array} \quad \text{para} \quad j = 0, 1, \ldots, 5.$$

Además, utilizando el estadístico de prueba t-student

$$T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{44} \text{ bajo } H_0$$

obtenemos los resultados de las pruebas: **valor del estadístico de prueba** y el **valor-P**, los cuales se encuentran en las dos últimas columnas de la tabla de parámetros estimados (Cuadro 1).

A un nivel de significancia $\alpha = 0.05$ se concluye que los parámetros individuales β_4, β_5 son significativos cada uno en presencia de los demás parámetros.

Por otro lado, se encuentra que $\beta_0, \beta_1, \beta_2, \beta_3$ son individualmente no significativos en presencia de los demás parámetros.

1.3. Interpretando los parámetros estimados

El primer paso es identificar aquellos parámetros susceptibles de interpretación, es decir, solo se podrán interpretar parámetros que resultaron significativos individualmente, como ya vimos, en este caso son: β_4, β_5.

- $\widehat{\beta}_4 = 0.02846$ indica que ante un aumento en una unidad del número promedio de pacientes que están en el hospital por día durante el periodo de estudio, la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje) aumenta en 0.02846 %, cuando las demás predictoras se mantienen fijas.
- $\widehat{\beta}_5 = 0.00234$ indica que ante un aumento en una unidad del número promedio de enfermeras presentes a tiempo completo en el hospital durante el periodo del estudio, la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje) aumenta en 0.00234 %, cuando las demás predictoras se mantienen fijas.

1.4. Analizando la significancia de la regresión

Conocemos que la prueba de significancia de la regresión establece lo siguiente:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_5 = 0, \quad \text{vs.}$$
$$H_1 : \text{algún } \beta_j \neq 0, j = 1, \ldots, 5.$$

Para poder concluir, se usará la tabla de análisis de varianza (ANOVA) que se muestra a continuación

Cuadro 2: Tabla ANOVA

	Suma de cuadrados	g.l.	Cuadrados medios	F Value	Valor P
Modelo	42.6514	5	8.530289	9.94355	2.05896e-06
Error	37.7464	44	0.857872		

De la tabla ANOVA se obtienen los valores del estadístico de prueba $F_0 = \frac{\text{MSR}}{\text{MSE}} = 9.94355$ y su correspondiente valor-P: VP $= 2.05896e - 06$.

Como $VP < 0.05 = \alpha$ se rechaza H_0 concluyendo que el modelo de RLM propuesto es significativo. Esto quiere decir, que la probabilidad promedio estimada de adquirir infección

en el hospital depende significativamente de al menos una de las variables predictoras del modelo.

1.5. Calculando el coeficiente de determinación múltiple R2

Sabemos que $R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$, De la tabla ANOVA extraemos los valores necesarios para su cálculo:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{42.6514}{37.7464 + 42.6514} = 0.53051$$

- Es decir, el 53.05 % de la variabilidad total de la probabilidad promedio de adquirir infección en el hospital es explicado por el modelo de RLM propuesto.
- De esta forma, el 46.95 % de la variabilidad total de la probabilidad promedio de adquirir infección en el hospital es explicado por el error.

No obstante, este R^2 no es la medida de preferencia para sacar conclusiones del modelo y su ajuste, por el contrario, se prefiere calcular el R^2 ajustado como una medida que sí penaliza al modelo por el número de variables incluidas.

Se calcula como se muestra a continuación:

$$R^2_{\text{adj}} = 1 - \frac{(n-1)\,\text{MSE}}{\text{SST}} = 1 - \frac{(50-1)\,0.857872}{37.7464 + 42.6514} = 0.47715$$

El valor de $R^2_{\text{adj}} = 0.47715$ es menor que $R^2 = 0.53051$, lo que indica que en el modelo pueden haber variables que no aporten significativamente. En otras palabras, se podrían quitar variables del modelo que no aporten.

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.

Solución

Obtenemos la tabla de todas las regresiones posibles

Cuadro 3: Resumen tabla de todas las regresiones posibles

K	R^2	Adj R^2	SEE	Cp	Variables en el modelo
1	0.319	0.305	54.724	17.791	X4
1	0.263	0.247	59.270	23.090	X1
...	...	...	...	...	...
2	0.504	0.483	39.858	2.461	X4 X5

K	R^2	Adj R^2	SEE	Cp	Variables en el modelo
...	...	...	...	...	...
5	0.531	0.477	37.746	6.000	X1 X2 X3 X4 X5

Ahora, para probar la significancia simultánea de dicho subconjunto planteamos las siguientes hipótesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0, \quad \text{vs.}$$
$$H_1 : \text{algún } \beta_j \neq 0, j = 1, 2, 3.$$

Para llevar a cabo esta prueba de hipótesis, se usará el estadístico de prueba:

$$F_0 = \frac{\text{MSextra}}{\text{MSE (MF)}} = \frac{\text{MSR}(\beta_1, \beta_2, \beta_3 | \beta_0, \beta_4, \beta_5)}{\text{MSE (MF)}} = \frac{[\text{SSR}(\beta_1, \beta_2, \beta_3 | \beta_0, \beta_4, \beta_5)]/3}{\text{MSE (MF)}}$$
$$= \frac{[\text{SSE}(\beta_0, \beta_4, \beta_5) - \text{SSE}(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)]/3}{\text{MSE (MF)}}$$
$$= \frac{[39.858 - 37.746]/3}{0.857872} = 0.82064$$

¿Dónde está el modelo reducido?

Para el criterio de decisión se requiere obtener el valor crítico de una distribución $f_{3,50-6} = f_{3,44}$ a un nivel de significancia $\alpha = 0.05$, esto es, $f_{0.05,3,44} = 2.8164$.

Como $F_0 = 0.82064 < f_{0.05,3,44} = 2.8164$, entonces no se rechaza H_0 y se concluye que la probabilidad promedio de adquirir infección en el hospital no está influenciada por la duración de la estadía, ni por la rutina de cultivos, ni por el número de camas presentes en el hospital en el periodo de estudio. Con base a estos resultados, podemos afirmar que es posible descartar del modelo anterior las variables de este subconjunto que involucran a $\beta_1, \beta_2, \beta_3$.

3. Pregunta 3

Plantee una pregunta donde su solución implique el uso *exclusivo* de una prueba de hipótesis lineal general de la forma $H_0 : L\beta = 0$ (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).

Solución

Al analizar el contexto del problema, se plantea lo siguiente:

¿Al aumentar en una unidad el número promedio de enfermeras presentes a tiempo completo en el hospital durante el periodo del estudio (X_5), la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje) incrementará en la misma medida que si aumentara en una unidad la duración promedio de la estadía de todos los pacientes en el hospital (en días)(X_1)?

Por otro lado, es interesante verificar si el efecto en la probabilidad promedio estimada de adquirir infección en el hospital que causa aumentar en una unidad el número promedio de camas (X_3), es igual al efecto que tendría aumentar en una unidad el promedio de pacientes en el hospital (X_4) en el mismo periodo.

Notemos que a partir de los anterior, se puede proponer una hipótesis nula de la en la forma $H_0 : \boldsymbol{L\beta} = \mathbf{0}$, de manera que se tiene una prueba de hipótesis lineal general de la forma:

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

En este sentido, veamos H_0 como un sistema de dos ecuaciones:

$$H_0 : \begin{cases} \beta_1 - \beta_5 = 0 \\ \beta_3 - \beta_4 = 0 \end{cases}$$

Que en forma matricial se puede expresar como:

$$H_0 : \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Notemos que la matriz $\boldsymbol{L}$ tiene $r = 2$ filas linealmente independientes (ninguna de las dos fila puede escribirse como un múltiplo escalar de la otra).

Por lo tanto, el modelo bajo H_0 es:

$$\text{RM: } Y = \beta_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_3 X_{i4} + \beta_1 X_{i5} + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right),$$

Operando obtenemos:

$$\begin{aligned} \text{RM: } Y_i &= \beta_0 + \beta_1 \left(X_{i1} + X_{i5}\right) + \beta_2 X_{2i} + \beta_3 \left(X_{i3} + X_{i4}\right) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right) \\ &= \beta_0 + \beta_1 X_{1,5} + \beta_2 X_2 + \beta_3 X_{3,4} + \varepsilon_i \end{aligned}$$

Donde $X_{1,5} = X_{i1} + X_{i5}$, y $X_{3,4} = X_{i3} + X_{i4}$

Finalmente, la expresión para el estadístico de prueba tiene la forma:

$$F_0 = \frac{\text{MSH}}{\text{MSE}} = \frac{\text{SSH}/2}{\text{MSE}} = \frac{[\text{SSE(RM)}^* - \text{SSE(FM)}]/2}{\text{MSE}} = \frac{[\text{SSE(RM)}^* - 37.746]/2}{0.85786}$$

4. Pregunta 4

18,5 pt

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.

Solución

Procedemos a validar primero los supuestos de normalidad y varianza constante de los errores del modelo.

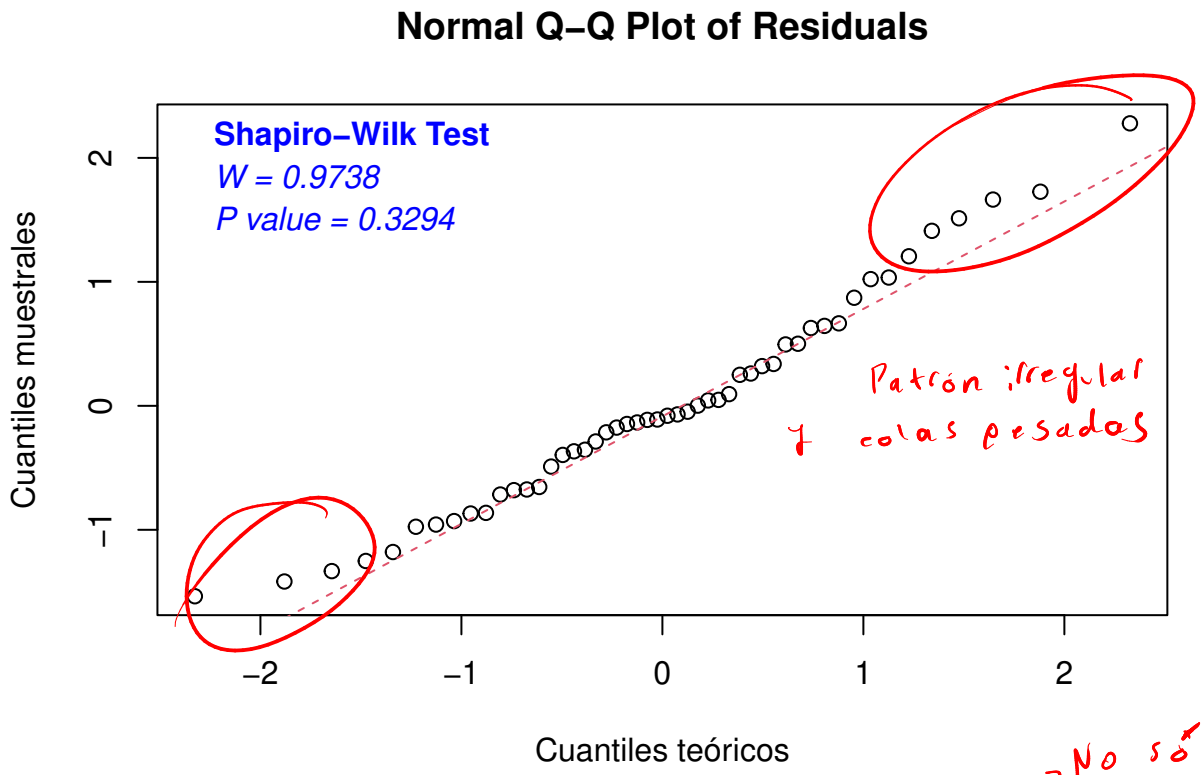
Empecemos por probar:

$$H_0: \ \varepsilon_i \sim \text{Normal. vs. } \ H_1: \ \varepsilon_i \not\sim \text{Normal}$$

4.1. Supuesto de normalidad

3,5 pt

Para aceptar o rechazar este supuesto, haremos uso de la gráfica de normalidad y prueba de Shapiro-Wilk.



No sólo eso, también el patrón!

Si bien la prueba de normalidad Shapriro-Wilk indica que los errores son normales (valor-P $= 0.3294 > 0.05$). El patrón de los residuales no sigue estrictamente la línea roja que representa el ajuste de la distribución de los residuales a una distribución normal, por lo tanto, podemos concluir que, debido a este motivo, el supuesto de normalidad no se cumple.

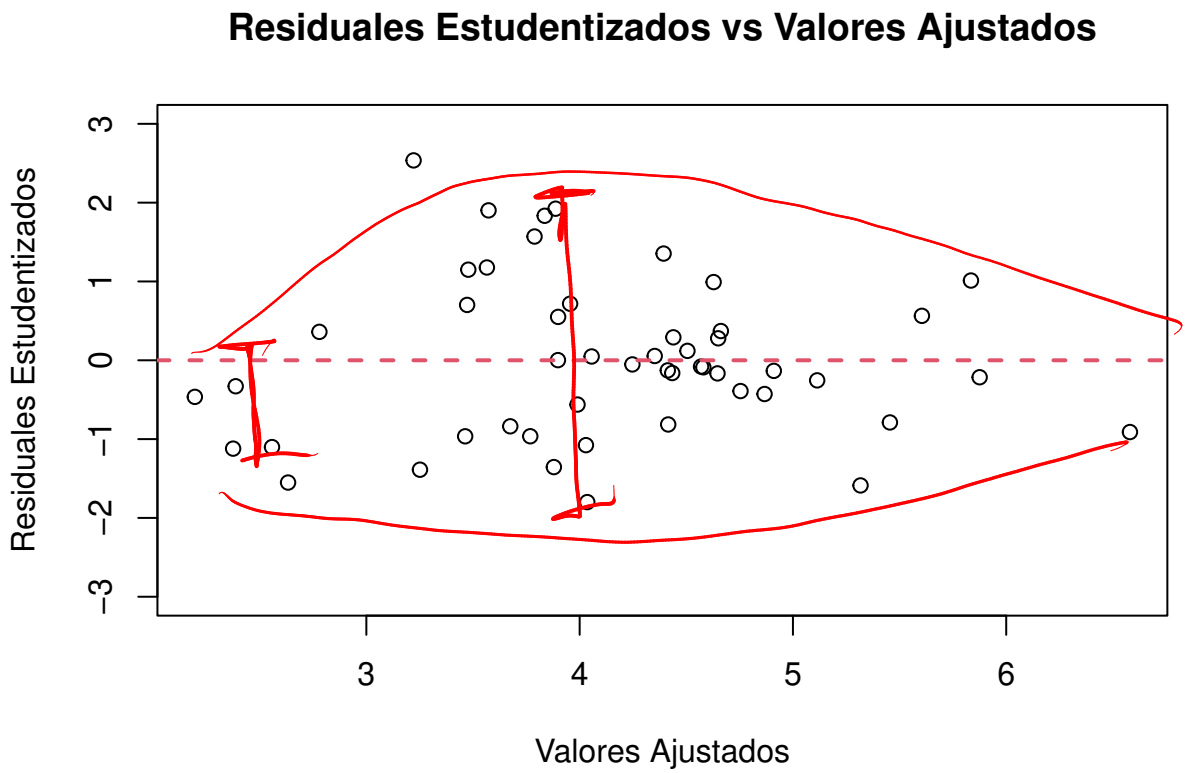
4.2. Supuesto de varianza constante

En este caso, haremos la prueba de forma gráfica, justamente a través del gráfico de residuales vs. valores ajustados

Tenemos entonces que se quiere probar:

$$H_0: \ V[\varepsilon_i] = \sigma^2 \quad \text{vs.} \quad H_1: \ V[\varepsilon_i] \neq \sigma^2$$

Para ello se usa la siguiente gráfica.



Resultado del análisis de la gráfica Valores Ajustados vs Residuales Estudentizados se tiene que la dispersión de los puntos inicia con una concentración a la izquierda, cercana al cero, que se transforma en un leve aumento de la dispersión hacia los extremos; dispersión la cual, a mitad del gráfico, toma una actitud de decrecimiento en dirección al centro de la figura analizada. Es decir, hay patrones en los que la varianza aumenta o disminuye que permiten rechazar el supuesto de varianza constante.

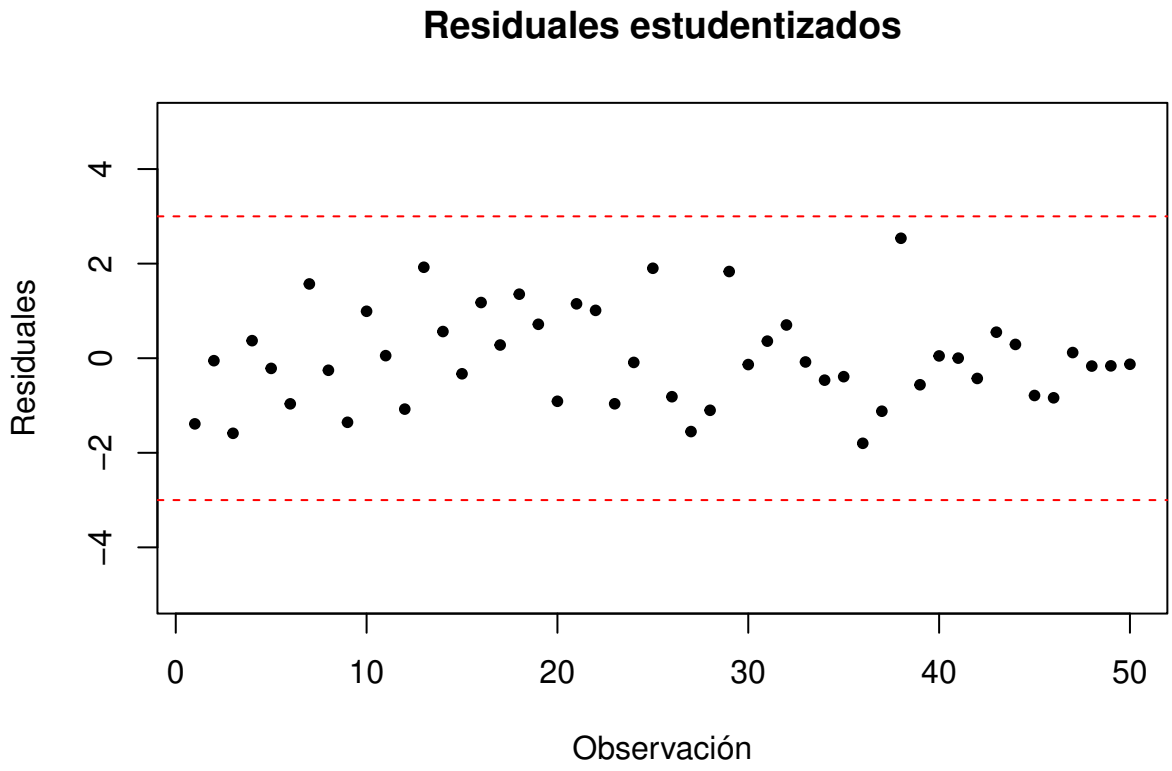
4.3. Análisis de la presencia de observaciones extremas

Para identificar si en el modelo hay observaciones extremas, se deben calcular los estadísticos que nos permiten aplicar criterios en ese sentido, los cuales incluyen: residuales estudentizados, los valores de la diagonal de la matriz $\boldsymbol{H}$ (los h_{ii}), la distancia de Cook (D_i) y los DFFITS.

4.3.1. Identificación de valores atípicos

Se considera que una observación es **atípica** cuando su residual estudentizado r_i, es tal que: $|r_i| > 3$.

Gráficamente, vemos que ningún valor es mayor a 3 o menor que -3.



4.3.2. Identificación de puntos de balanceo

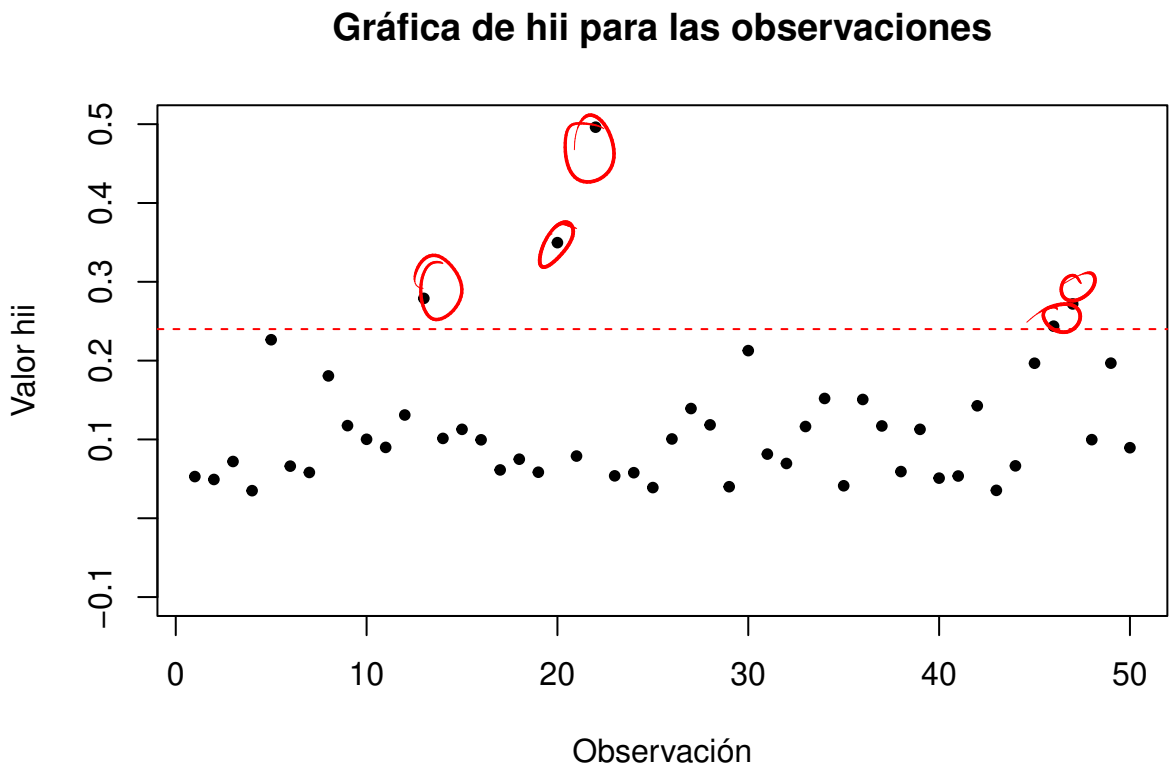
Se asume que la observación i es un **punto de balanceo** si $h_{ii} > 2p/n$. En este caso tenemos que: $h_{ii} > 2p/n = 2(6/50) = 0.24$. Por lo tanto, de la tabla de valores para el diagnóstico de valores extremos se tiene que:

Cuadro 4: Resumen tabla de puntos de balanceo

Observación	Valor hii
13	0.2790
20	0.3497
22	0.4962
46	0.2437
47	0.2719

De acuerdo a la columna **Valor hii** de valores de la diagonal de la matriz $\boldsymbol{H}$ se tiene que las observaciones 13, 20, 22, 46 y 47 son puntos de balanceo.

Para confirmar lo visto con la tabla, se muestra además la gráfica que resalta las 5 observaciones ya mencionadas:



4.3.3. Identificación de oberservaciones influenciales

Recuerde que para identificar esos valores tenemos dos criterios: Por un lado, se dice que la observación i será **influencial** si $D_i > 1$, y por el otro, una observación será **influencial** si $|\text{DFFITS}_i| > 2\sqrt{\frac{p}{n}}$. En este caso tenemos que el criterio basado en DFFITS debe superar en valor absoluto a $2\sqrt{\frac{6}{50}} = 0.6928203$.

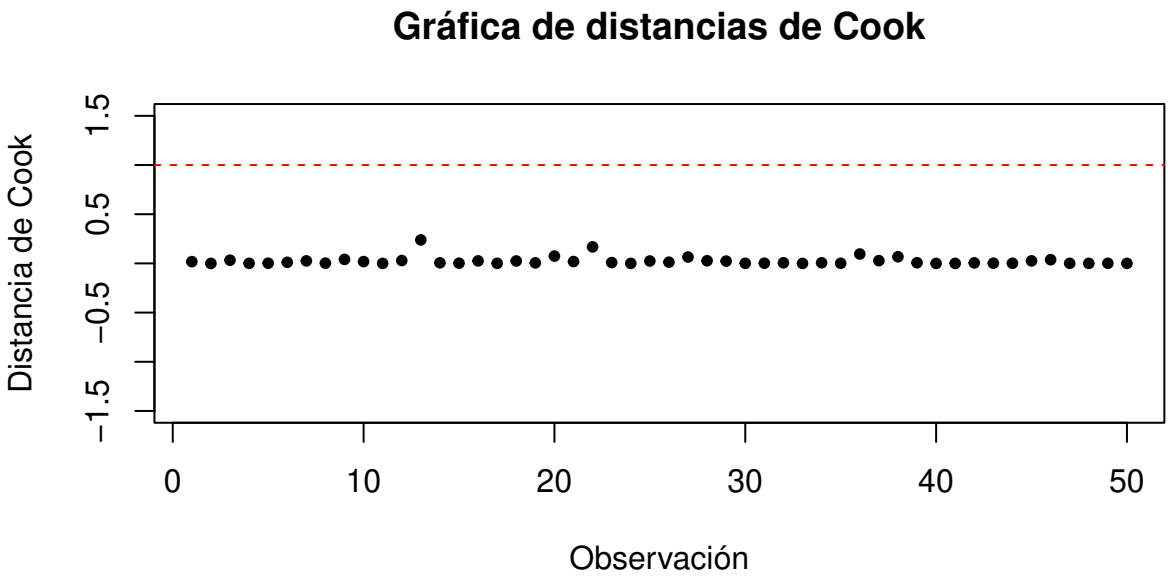
Analizando nuevamente la tabla de valores para el diagnóstico de valores extremos obtenemos

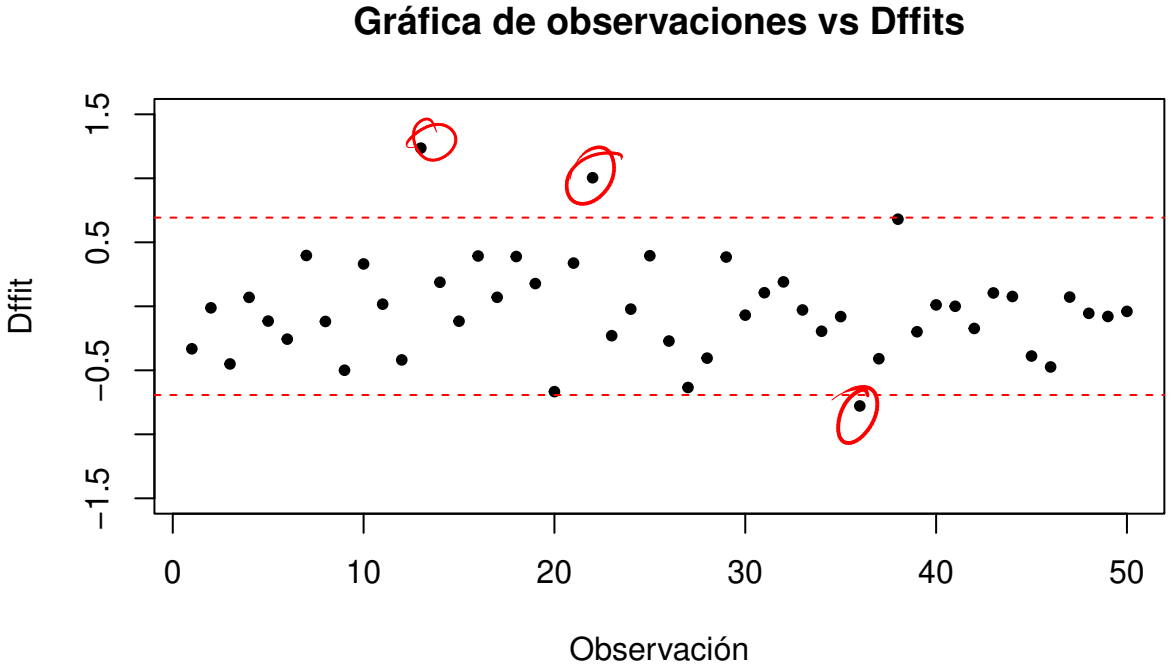
Cuadro 5: Resumen tabla de valores extremos

Observación	Cook (D_i)	Dffits
13	0.2385	1.2358
22	0.1681	1.0045
36	0.0957	-0.7783

- De acuerdo a la columna **Cooks** (D_i) de distancias de Cook tenemos que ninguna observación es influencial.
- De acuerdo a la columna **Dffits** de valores DFFITS tenemos que las observaciones 13, 22 y 36 son influenciales.

Lo anterior se puede confirmar por medio de los siguientes gráficos





¿Qué causan estos puntos según el criterio?

4.4. Conclusiones

Como vimos anteriormente, el supuestos de normalidad y varianza constante de los errores del modelo no se cumple, por lo tanto, el modelo no es valido para hacer estimaciones y predicciones. Es importante resaltar que estos procesos de prueba se hicieron aún considerando los posibles valores extremos que pudiese tener el modelo.

Para identificar justamente estos valores que pueden alterar el modelo, se parte del análisis de observaciones extremas que acabamos de desarrollar, donde se obtuvo:

- Niguna de las observaciones es atípica.
- Las observaciones 13, 20, 22, 46 y 47 son puntos de balanceo: Por lo tanto, estas son observaciones en el espacio de las variables predictoras que están alejadas del resto de la muestra y posiblemente afectan al R^2 y los Se de los coeficientes estimados.
- Las observaciones 13, 22 y 36 son influenciales: En consecuencia, son observaciones que tienen un impacto notable sobre los coeficientes de regresión ajustados. Es así, una observación que hala al modelo en su dirección.

Se aprecia que, en efecto, se detectó la presencia de observaciones extremas que tendrán que ser estudiadas antes de usar el modelo y así evaluar de nuevo su validez como predictor o estimador de valores de la respuesta.