Trabajo 1

4,8

Estudiantes

Omar David Mercado Turizo
Laura Juliana Insignares Montes
Sayaana Valeria Diaz Rivero
Juan Camilo Bastidas Alvarez

Equipo 18

Docente

Julieth Veronica Guarin Escudero

Asignatura

Estadística II

Sede Medellín
5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

20 pt

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2),\ i = 1, .., 69$$

Donde las variables son las siguientes:

- 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

1.1. Modelo de regresión

Al ajustar el modelo, se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
$\hat{\beta}_0$	-0.5788
$\hat{\beta}_1$	0.1225
$\hat{\beta}_2$	0.0224
$\hat{\beta}_3$	0.0668
$\hat{\beta}_4$	0.0139
$\hat{\beta}_5$	0.0015

3 pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.5788 + 0.1225X_{1i} + 0.0224X_{2i} + 0.0668X_{3i} + 0.0139X_{4i} + 0.0015X_{5i},\quad i = 1, 2, ..., 69$$

1.2. Significancia de la regresión

Para analizar la significancia de la regresión, se plantea el siguiente juego de hipótesis:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2, 3, 4, 5.} \end{cases}$$

Donde el estadístico de prueba es el siguiente:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,63} \tag{1}$$

5pt

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	77.3476	5	15.469516	17.6076	7.07584e-11
Error	55.3498	63	0.878568		

De la tabla Anova, se observa un valor P muy pequeño, por lo que se rechaza la hipótesis nula en la que $\hat{\beta}_j = 0$ con $j = 1, .., 5$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$, por lo tanto, la regresión es ~~globalmente~~ significativa.

1.3. Significancia de los parámetros

Planteamos las siguientes hipotesis:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_a : \beta_j \neq 0 \text{ con } 0 \leqslant j \leqslant 5 \end{cases}$$

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

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-0.5788	1.4839	-0.3901	0.6978
β_1	0.1225	0.0721	1.6998	0.0941
β_2	0.0224	0.0275	0.8124	0.4196
β_3	0.0668	0.0135	4.9399	0.0000
β_4	0.0139	0.0069	2.0152	0.0482
β_5	0.0015	0.0007	2.2106	0.0307

6pt

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros $\hat{\beta}_3$, $\hat{\beta}_4$ y $\hat{\beta}_5$ son significativos, pues sus P-valores son menores a $\alpha = 0.05$.

1.4. Interpretación de los parámetros

3pt

como el 0 no está incluido en el intervalo no tiene interpretación.

$\hat{\beta}_3$:

Indica que por cada unidad que aumenta el número de camas, aumenta en promedio 0.0668 el riesgo de infección en el hospital dado que las otras variables predictoras se mantienen constantes.

$\hat{\beta}_4$:

Indica que por cada unidad que aumenta el número de pacientes, aumenta en promedio 0.0139 el riesgo de infección en el hospital dado que las otras variables predictorias se mantienen constantes.

$\hat{\beta}_5$:

Indica que por cada unidad aumenta el número de enfermeras, aumenta en promedio el riesgo de infección en el hospital en 0.0015 dado que las otras variables predictorias se mantienen constantes.

1.5. Coeficiente de determinación múltiple R^2

3pt

Extrayendo valores de la tabla ANOVA, tenemos que:

$$R^2 = \frac{SSR}{SST} = \frac{77.3476}{(77.3476 + 55.3498)} = 0.5829$$

- Es decir, el 58.29 % de la variabilidad total de la probabilidad promedio de adquirir infección en el hospital es explicado por el modelo RLM propuesto.
- Simultaneamente, el 41.71 % de la variabilidad total de la probabilidad promedio de adquirir infección en el hospital es explicado por el error del modelo.

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

Se calcula como se muestra a continuación:

$$R^2 adj = 1 - \frac{(n-1)MSE}{SST} = 1 - \frac{(69-1)0.878568}{(77.3476 + 55.3498)} = 0.549783$$

El valor de $R^2 adj = 0.549783 < R^2 = 0.5829$, lo que indica que en el modelo pueden haber variables que no aporten significativamente a la estimación.

2. Pregunta 2

5pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con el P-valor más bajos en el modelo fueron X_3, X_4, X_5, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_3 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 3, 4, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	55.350	X1 X2 X3 X4 X5
Modelo reducido	93.904	X1 X2

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); i = 1, ..., 69$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

3pt

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_2) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,63} \\ &= \frac{93.904 - 55.350/3}{0.878568} \\ &= 14.62759 \end{aligned} \tag{2}$$

2pt

Ahora, comparando el F_0 con $f_{0.95,3,63} = 2.7505$, se puede ver que $F_0 > f_{0.95,3,63}$ y por lo tanto se rechaza la hipotesis nula y decimos que hay evidencia suficiente para decir que algún β_j es distinto de 0 para j=3,4,5. Por lo tanto alguno de ellos es significativo.

Con base a los resultados no es posible descartar del modelo anterior las variables del subconjunto que involucran a β_3, β_4, β_5, o sea el número de camas en el hospital, censo promedio diario y el número de enfermeras ya que al menos una de ellas es significativa para el modelo.

3. Pregunta 3

4pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si $\beta_3 = 6\beta_1$; $\beta_2 = 2\beta_5$ por consiguiente se plantea la siguiente prueba de hipótesis:

$\neq$

$$\begin{cases} \mathrm{H}_0 : \beta_1 = 6\beta_3; \ \beta_2 = 2\beta_5 \\ \mathrm{H}_1 : \text{Alguna de las ecuaciones no se cumple} \end{cases}$$

Lo que es equivalente a lo siguiente:

$$\begin{cases} \mathrm{H}_0 : \beta_1 - 6\beta_3 = 0; \ \beta_2 - 2\beta_5 = 0 \\ \mathrm{H}_1 : \text{Al menos una de las ecuaciones no se cumple} \end{cases}$$

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:

$$L = \begin{bmatrix} 0 & 1 & 0 & -6 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -2 \end{bmatrix}$$

1pt

El modelo completo esta dado por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i, \ \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); i = 1, .., 69$$

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_4 X_{4i} + \beta_3 X^*_{3i} + \beta_5 X^*_{5i} + \varepsilon_i; \ \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); i = 1, .., 69$$

1pt

Donde:

$X^*_{3i} = 6X_{1i} + X_{3i}$ y $X^*_{5i} = 2X_{2i} + X_{5i}$

3.2. Estadístico de prueba

2pt

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(\beta_0, \ \cdots, \ \beta_5)))/2}{MSE(\beta_0, \ \cdots, \ \beta_5))} \overset{H_0}{\sim} f_{2,63} \tag{3}$$

Al reemplazar con los valores conocidos, se encuentra lo siguiente:

$$F_0 = \frac{(SSE(MR)) - 77.3476)/2}{0.878568} \overset{H_0}{\sim} f_{2,63} \tag{4}$$

4. Pregunta 4

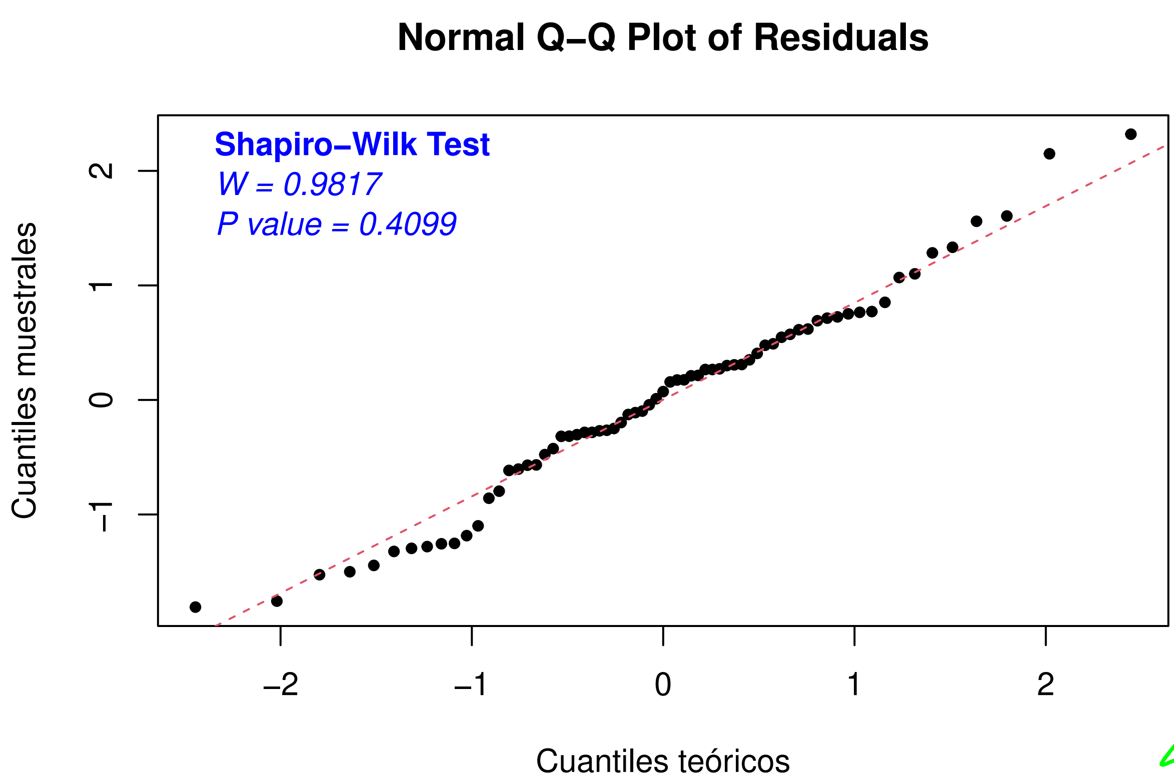
19pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Para la validación de este supuesto, se planteará el siguiente test de Shapiro-Wilk que se utiliza para determinar si un conjunto de datos puede distribuirse mediante la distribucción normal, acompañada de un gráfico cuantil-cuantil:

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



4pt

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Si bien la prueba de normalidad Shapiro-wilk indica que los errores son normales (P-valor $= 0.4099 > 0.05$), con un nivel de confianza al 95 %. El patrón de los residuales no sigue estrictamente la línea roja que representa el ajuste de la distribución normal y se evidencia patrones irregulares, por lo que podemos concluir que el supuesto de normalidad no se cumple.

4.1.2. Varianza constante

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

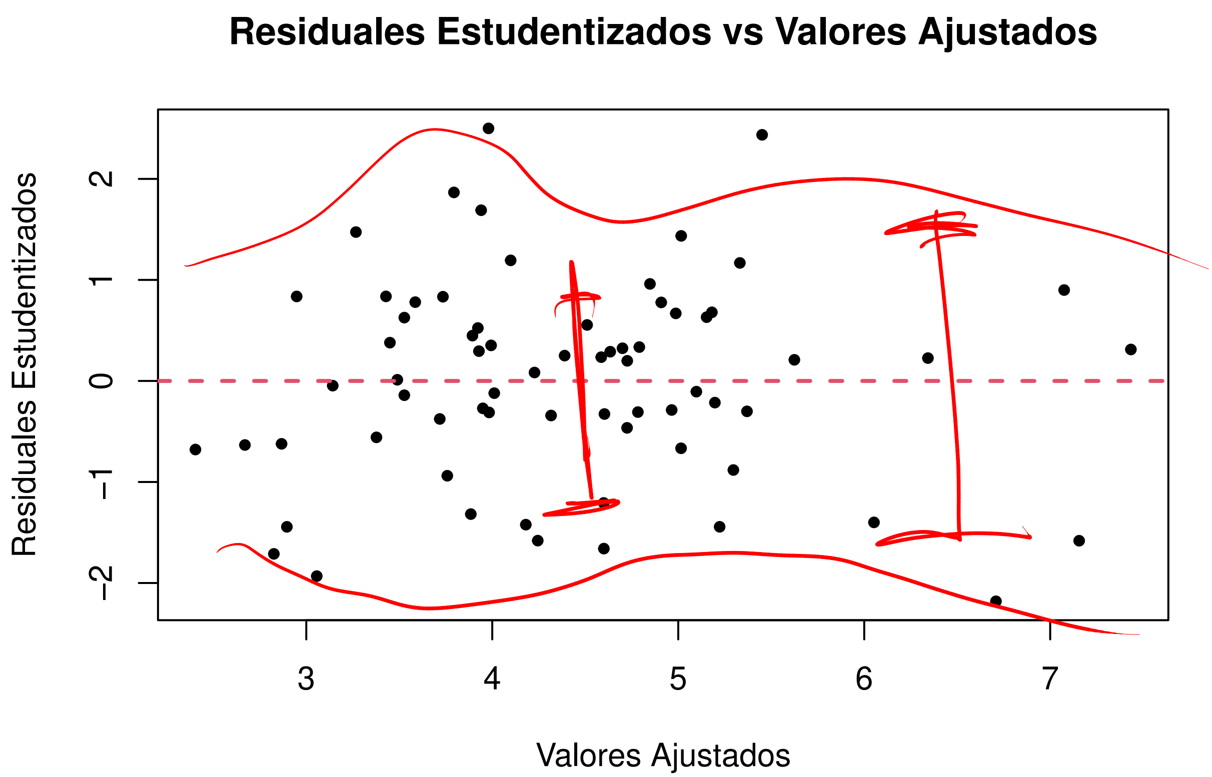


Figura 2: Gráfico residuales estudentizados vs valores ajustados 2 pt

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que no hay patrones en los que la varianza aumente ni decrezca por lo que podemos concluir que el supuesto de que los errores tienen varianza constante. Además es posible observar la media igual a 0.

Ls observan mejor.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

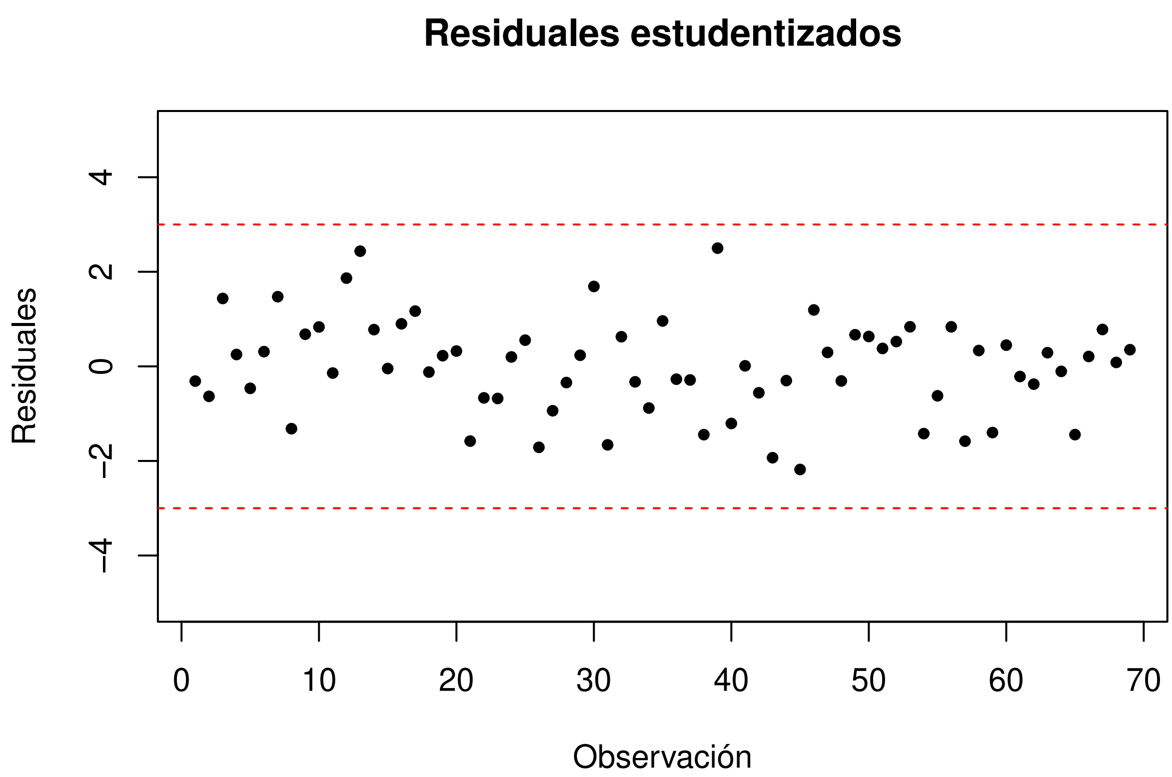


Figura 3: Identificación de datos atípicos

3pt

Graficamente no observamos ningún valor mayor a 3 ni menor a -3.

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

Se puede apreciar que hay 7 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{p}{n} = 0.174$, los cuales son los presentados en la tabla.

Cuadro 5: Puntos de balanceo

Observación	Valor hii
4	0.1954
16	0.2611
19	0.4498
45	0.2173
57	0.2824
62	0.1875
66	0.1983

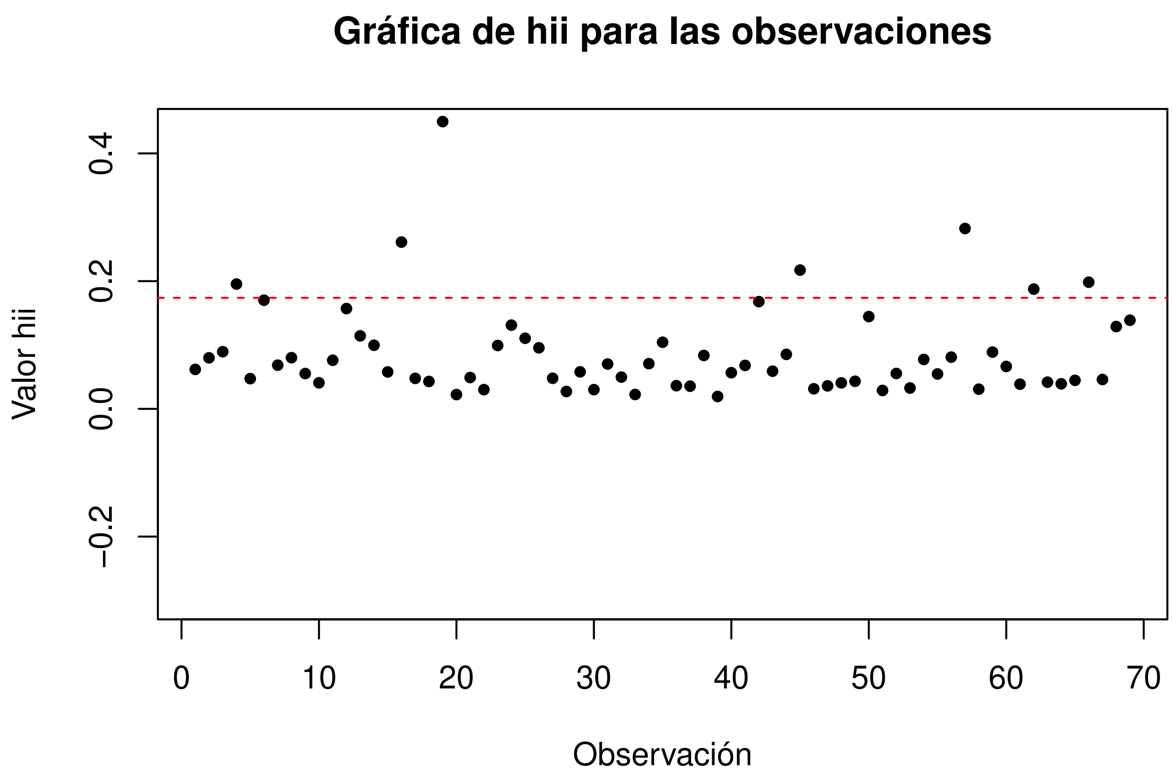


Figura 4: Identificación de puntos de balanceo

3pt

Claramente en la gráfica de observaciones vs valores h_{ii}, confirmamos lo visto con la tabla, la grafica resalta las 7 observaciones con su valor hii más alto que la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 2(6/69) = 0.174$,

4.2.3. Puntos influenciales

Recuerde que para identificar esos valores tenemos dos criterios: Por un lado, se dice que la observación i será influencial si su $D_i > 1$, y por el otro, una observación será influencial si

$|DFFITS_i| > 2\sqrt{\frac{p}{n}}$. En este caso tenemos que el criterio basado en $DFFITS$ debe superar en valor absoluto a $2\sqrt{\frac{6}{69}} = 0.5897678$.

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

Cuadro 6: Puntos influenciales

Observación	Cook (D_i)	Dffits
12	0.1081	0.8219
13	0.1276	0.9121
45	0.2200	-1.1854
57	0.1641	-1.0044

- De acuerdo al de **Cooks** (D_i) de distancias de Cook tenemos que ninguna observación es influencial.
- De lo contrario de acuerdo por el criterio **Dffits**, de los valores DFFITS tenemos que las observaciones 12, 13, 45 y 57 son influenciales.

Podemos verlo a continuación graficamente.

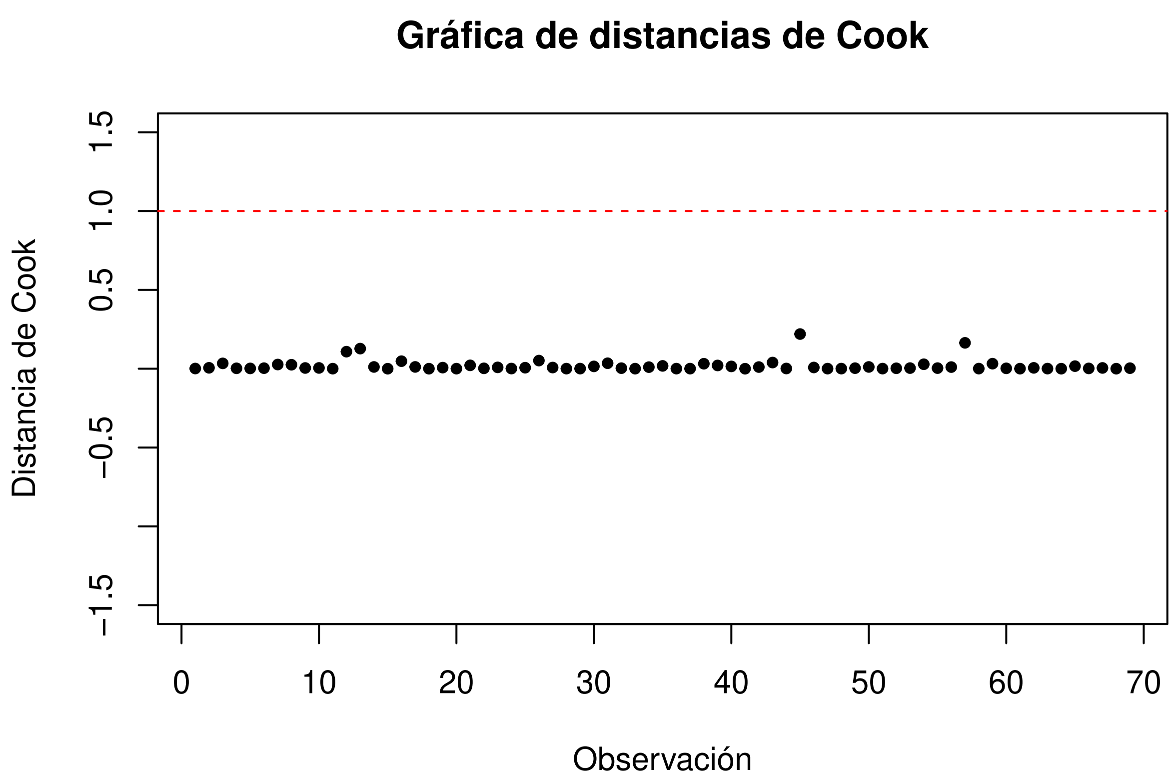


Figura 5: Criterio distancias de Cook para puntos influenciales

Podemos ver en la gráfica como vimos anteriormente en la tabla por el criterio de Cook no tenemos ninguna observación influenciable.

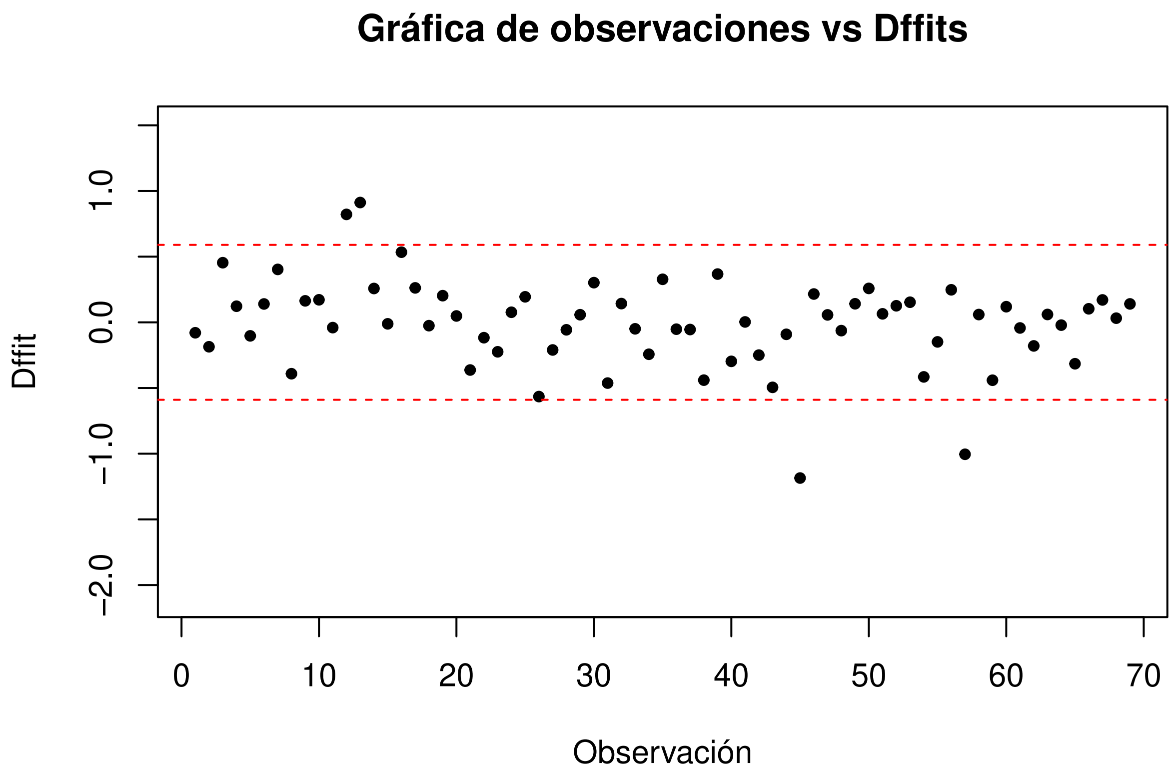


Figura 6: Criterio Dffits para puntos influenciales

4pt

Notamos que por el criterio Dffits si encontramos 4 observaciones influenciales que vimos en la tabla anteriormente, con la grafica queda confirmado.

4.3. Conclusión

3pt

Como vimos anteriormente el supuesto de normalidad en los errores del modelo no se cumple por lo tanto el modelo no es valido para hacer estimaciones y predicciones (inferencia sobre el modelo), Lo contrario al supuesto de varianza constate que vimos que si se cumplía. Es importante resaltar que estos procesos de prueba se hicieron teniendo en cuenta las observaciones de balance e influenciales las cuales pueden afectar las pruebas.

- Ninguna observación atípica
- Las obsevaciones 4,16,19,45,57,62 y 66 son puntos de balanceo los cuales afectan al R^2 y cambian un poco la inclinación del modelo estimado y la calidad del modelo.
- Las observaciones 12, 13, 45 y 57 son influenciales:

Los puntos influenciales pueden afectar de manera significativa la estimación de coeficientes, la precisión de las predicciones y la validez de las inferencias estadísticas en un modelo de regresión múltiple. Por lo tanto, es importante identificar y evaluar la influencia de estas observaciones durante el análisis de regresión para tomar decisiones informadas sobre la inclusión o exclusión de observaciones en el modelo.