Trabajo 1

Estudiantes

4,6

Paula Fernanda Bermeo Ruiz
Harold Smith Bolivar Reyes
David García Blandón
Danilo Giraldo López

Equipo 15

Docente

Julieth Veronica Guarin Escudero

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

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 brindada, en la cual hay 5 variables regresoras dadas por:

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

- Y: Riesgo de infección
- X_1: Duración de la estadía
- X_2: Rutina de cultivos
- X_3: Numero de camas
- X_4: Censo promedio diario
- X_5: Numero 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
β_0	2.9263
β_1	0.2263
β_2	-0.0409
β_3	0.0450
β_4	0.0043
β_5	0.0010

3pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 2.9263 + 0.2263 X_{1i} - 0.0409 X_{2i} + 0.045 X_{3i} + 0.0043 X_{4i} + 0.001 X_{5i};\ 1 \leqslant i \leqslant 69$$

1.2. Significancia de la regresión

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

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

Cuyo estadístico de prueba es:

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

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	53.4644	5	10.692884	13.1454	9.01504e-09
Error	51.2463	63	0.813433		

5pt

De la tabla Anova, se observa un valor P de 9.01504e-09, se rechaza la hipotesis nula sí el valor P es menor que α, el valor P resultante es menor a cualquier α dado, por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$, por lo tanto la regresión es significativa.

1.3. Significancia de los parámetros

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

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	2.9263	1.7599	1.6628	0.1013
β_1	0.2263	0.0684	3.3083	0.0016
β_2	-0.0409	0.0319	-1.2806	0.2050
β_3	0.0450	0.0143	3.1494	0.0025
β_4	0.0043	0.0067	0.6379	0.5259
β_5	0.0010	0.0007	1.5988	0.1149

6pt

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_1 y β_3 son significativos, pues sus P-valores son menores a α.

1.4. Interpretación de los parámetros

Identificando aquellos parámetros susceptibles de interpretación, solo se podrán interpretar parámetros que resultaron significativos individualmente, en este caso son: $\hat{\beta}_1$ y $\hat{\beta}_3$.

3pt

$\hat{\beta}_1 = 0.2263$: indica que por cada unidad (días) que se aumente la duración de la estadía (X_1) la probabilidad promedio de el riesgo de infección (Y) aumenta en 0.2263 unidades (porcentaje), cuando las demás predictoras se mantienen fijas.

$\hat{\beta}_3 = 0.0450$: indica que por cada unidad (camas) que se aumente el numero de camas (X_3) la probabilidad promedio de el riesgo de infección (Y) aumenta en 0.0450 unidades (porcentaje), cuando las demás predictoras se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2

3pt

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE} \tag{2}$$

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5106$, lo que significa que aproximadamente el 51 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

2. Pregunta 2

5pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Siendo la pregunta la siguiente: Use la tabla de todas las regresiones posibles, para probar la significancia simultánea del subconjunto de tres variables con los valores p más pequeños del punto anterior. Según el resultado de la prueba este subconjunto de parámetros son todos significativos? Explique su respuesta.

Se procede a hallar las covariable con el P-valor más bajo en el modelo resultando en X_1, X_3, X_5, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_3 = \beta_5 = 0 \\ \mathrm{H}_a : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 3, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	51.246	X1 X2 X3 X4 X5
Modelo reducido	85.588	X2 X4

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

$$Y_i = \beta_0 + \beta_2 X_{2i} + \beta_4 X_{4i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

2.2. Estadístico de prueba y conclusión

3pt

Se construye el estadístico de prueba como:

$$
\begin{aligned}
F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_4) - SSE(\beta_0,\ \cdots,\ \beta_5))/3}{MSE(\beta_0,\ \cdots,\ \beta_5)} \overset{H_0}{\sim} f_{3,63} \\
&= \frac{11.447}{0.813433} \\
&= 14.07287
\end{aligned} \tag{3}
$$

2pt

Ahora, comparando el $F_0 = 14.07287$ con $f_{0.95,3,63} = 2.7505$, se puede ver que $F_0 > f_{0.95,3,69}$ y por tanto qué se rechaza la hipotesis nula $H_0 : \beta_1 = \beta_3 = \beta_5 = 0$.

Esto nos indica que al menos uno de los parametros del subconjunto es distinto de cero, por lo tanto relevante para el modelo e impide su extracción del mismo.

3. Pregunta 3

4pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

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

$$
\begin{cases}
\mathrm{H}_0 : \beta_1 = 2\beta_3;\ \ \beta_2 = \beta_5 \\
\mathrm{H}_1 : \beta_1 \neq 2\beta_3;\ \ \beta_2 \neq \beta_5
\end{cases}
$$

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

2pt

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

El modelo reducido está dado por:

6pt

$$
Y_i = \beta_o + \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + \beta_4 X_{4i} + \varepsilon_i,\ \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ \ 1 \leqslant i \leqslant 69
$$

Donde $X_{1i}^* = X_{1i} + 2X_{3i}$ y $X_{2i}^* = X_{2i} + X_{5i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

2pt

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,63}$$
$$= \frac{(SSE(MR) - 51.2463)/2}{0.813433} \overset{H_0}{\sim} f_{2,63} \quad (4)$$

3.3. Región de rechazo

Luego, se rechaza $H_0 : \beta_1 = 2\beta_3;\ \beta_2 = \beta_5$ sí $F_0 > f_{\alpha,2,69}$

4. Pregunta 4

16,5 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Para la validación de este supuesto, se planteará la siguiente prueba de hipótesis que se realizará por medio de shapiro-wilk, acompañada de un gráfico cuantil-cuantil:

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

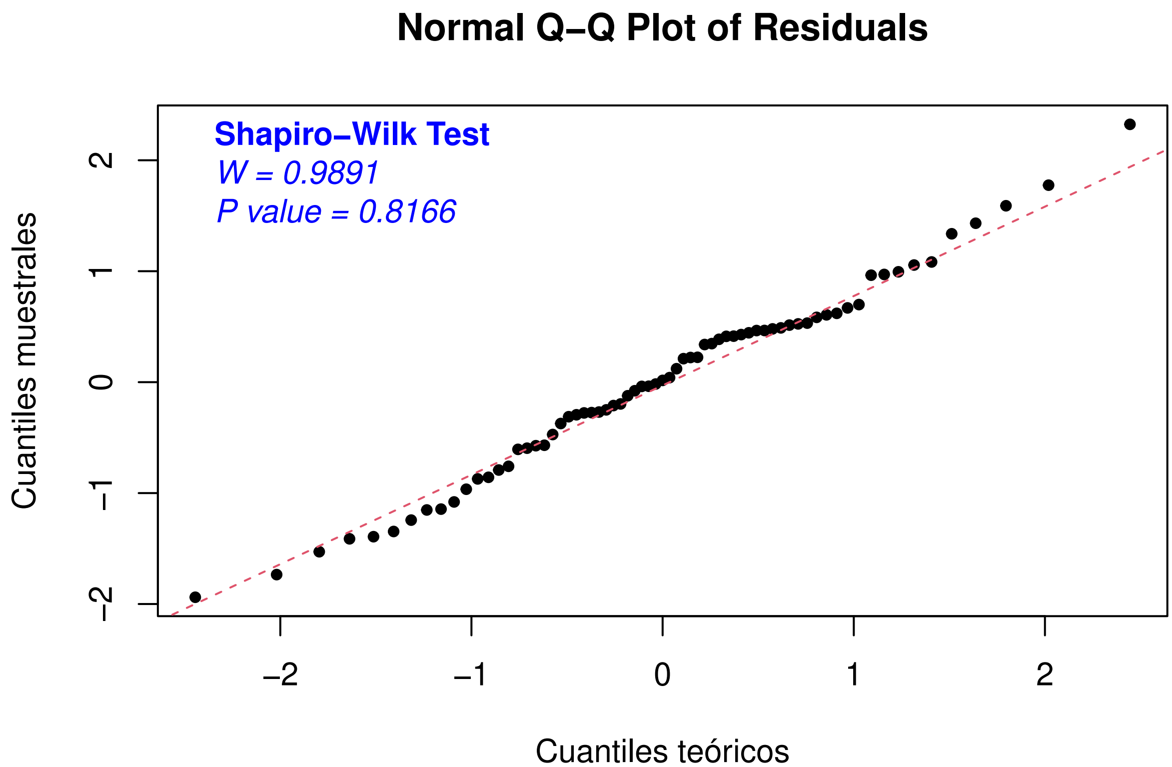


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

3,5 pt

Al ser el P-valor aproximadamente igual a 0.8166 y teniendo en cuenta que el nivel de significancia $\alpha = 0.05$, 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 μ y varianza σ^2, a pesar de que en la gráfica de comparación de cuantiles se observan un par de datos potencialmente atipicos en la cola superior.

↳ Faltó más análisis gráfico

4.1.2. Varianza constante

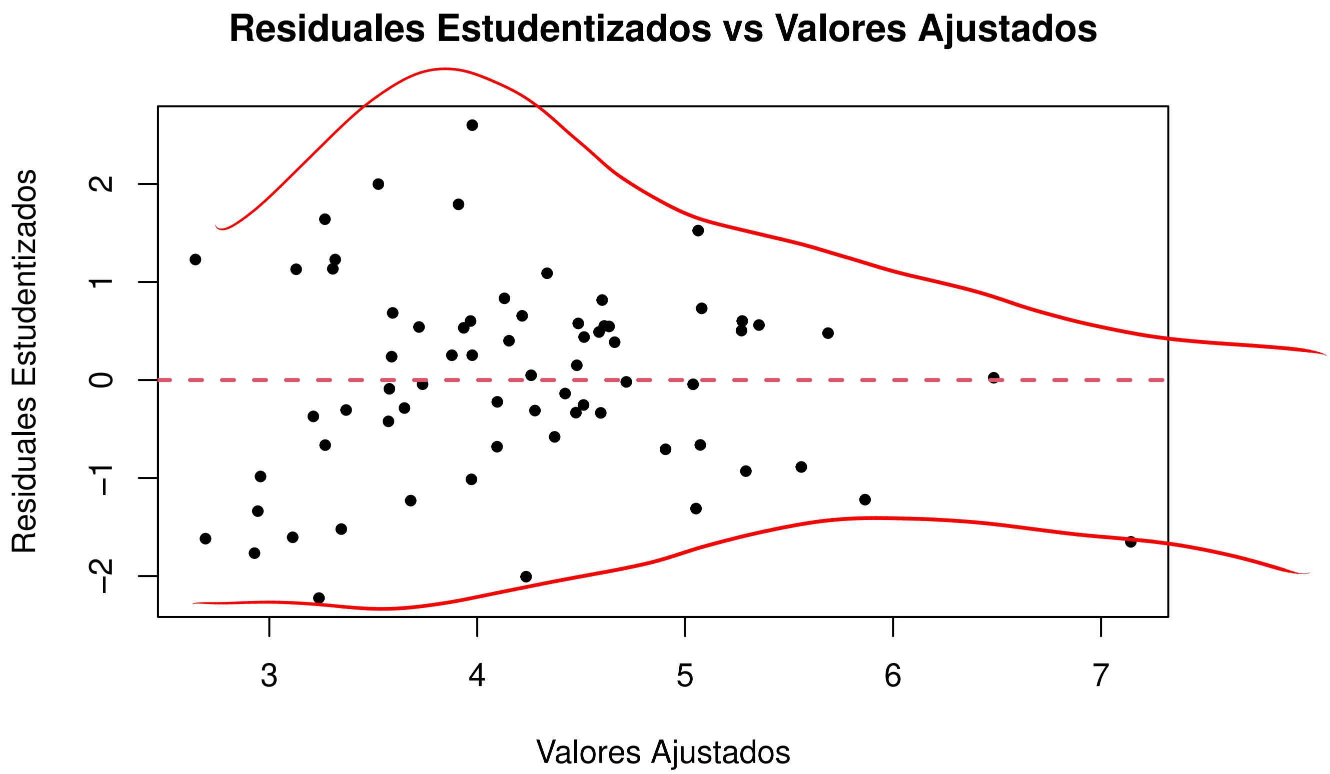


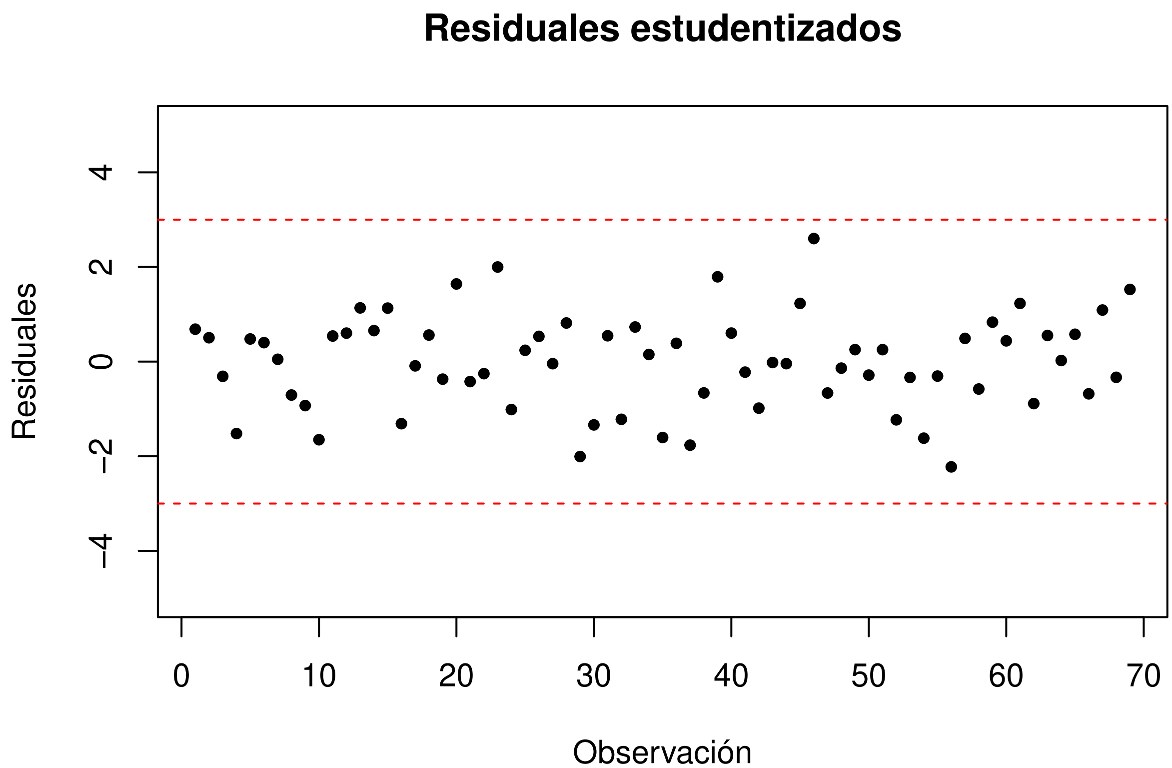
Figura 2: Gráfico residuales estudentizados vs valores ajustados

3pt

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que hay patrones que permiten suponer que la varianza no es constante ya que en la parte superior se observa un patron de acento circlunfejo ˆ y en la parte inferior un patron ascendente, por lo que se rechaza la hipotesis de la varianza constante.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos



3pt

Figura 3: Identificación de datos atípicos

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

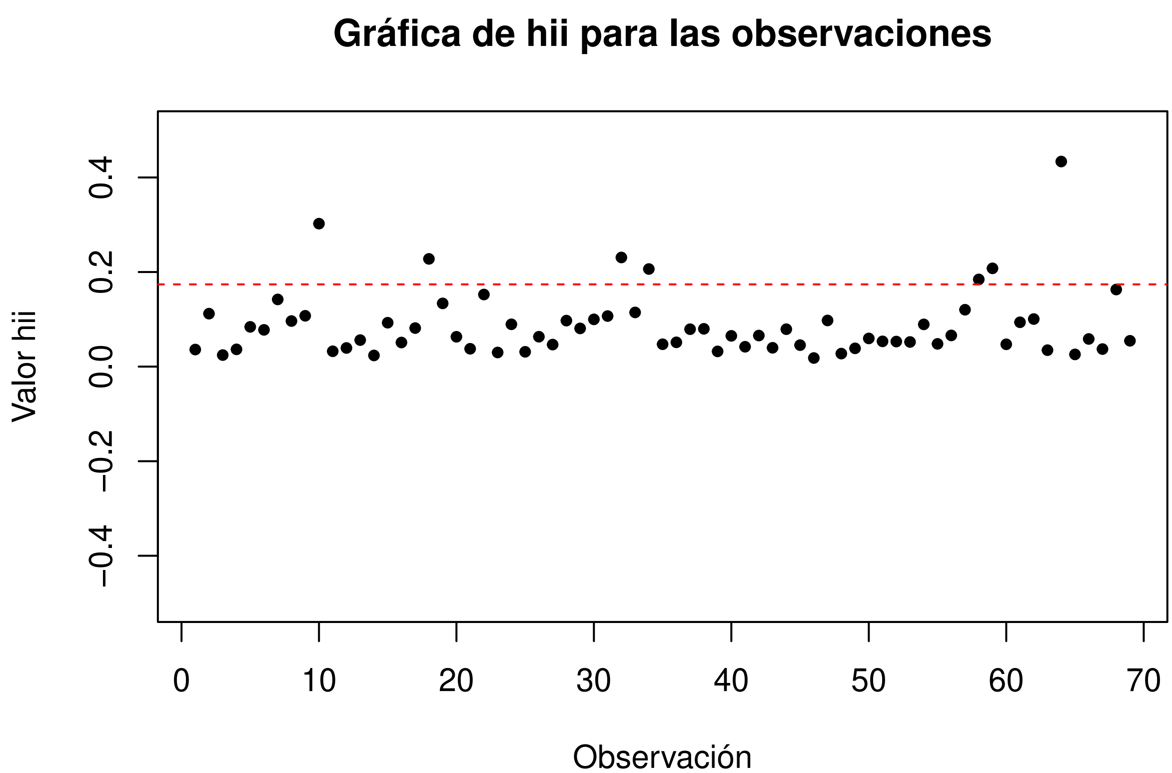


Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 10  -1.6509  0.1968    0.3023 -1.1022
## 18   0.5613  0.0155    0.2278  0.3032
## 32  -1.2203  0.0745    0.2308 -0.6711
## 34   0.1506  0.0010    0.2064  0.0762
## 58  -0.5799  0.0127    0.1844 -0.2743
## 59   0.8335  0.0304    0.2078  0.4258
## 64   0.0232  0.0001    0.4337  0.0201
```

2pt

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 0.1739$, se puede apreciar que existen 7 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 0.1739$, los cuales son los presentados en la tabla.

¿Qué causan?

4.2.3. Puntos influenciales

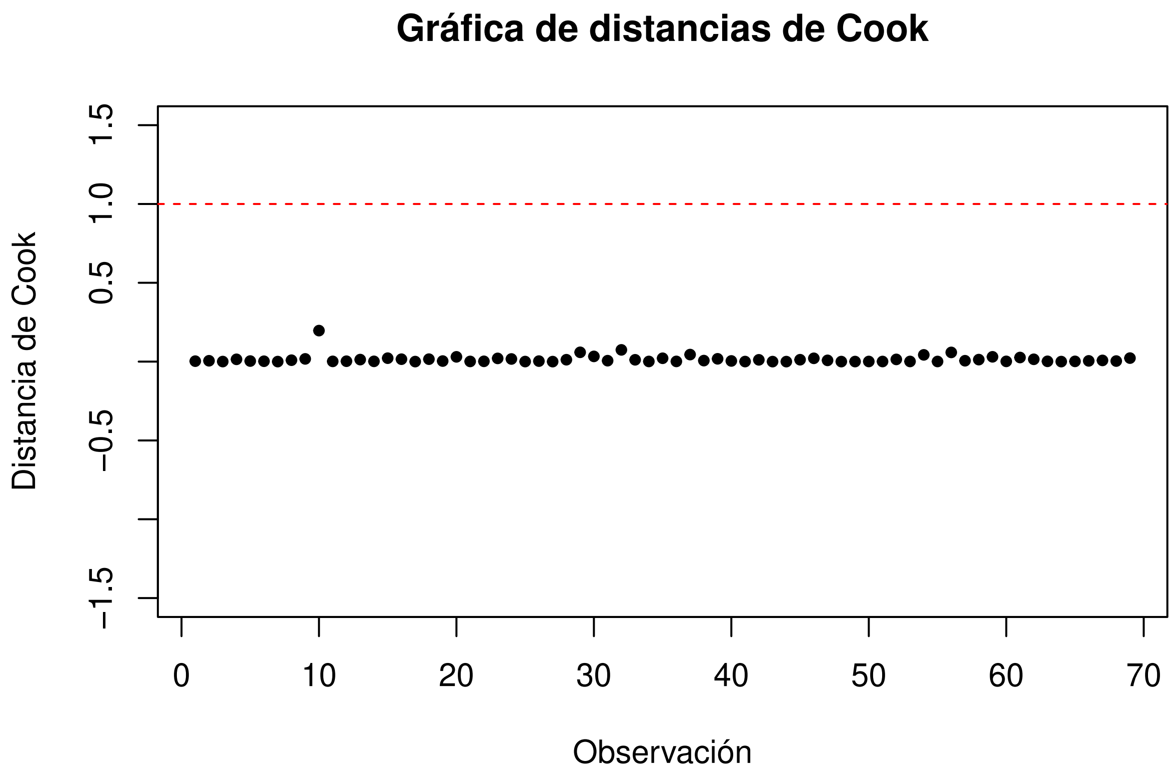


Figura 5: Criterio distancias de Cook para puntos influenciales

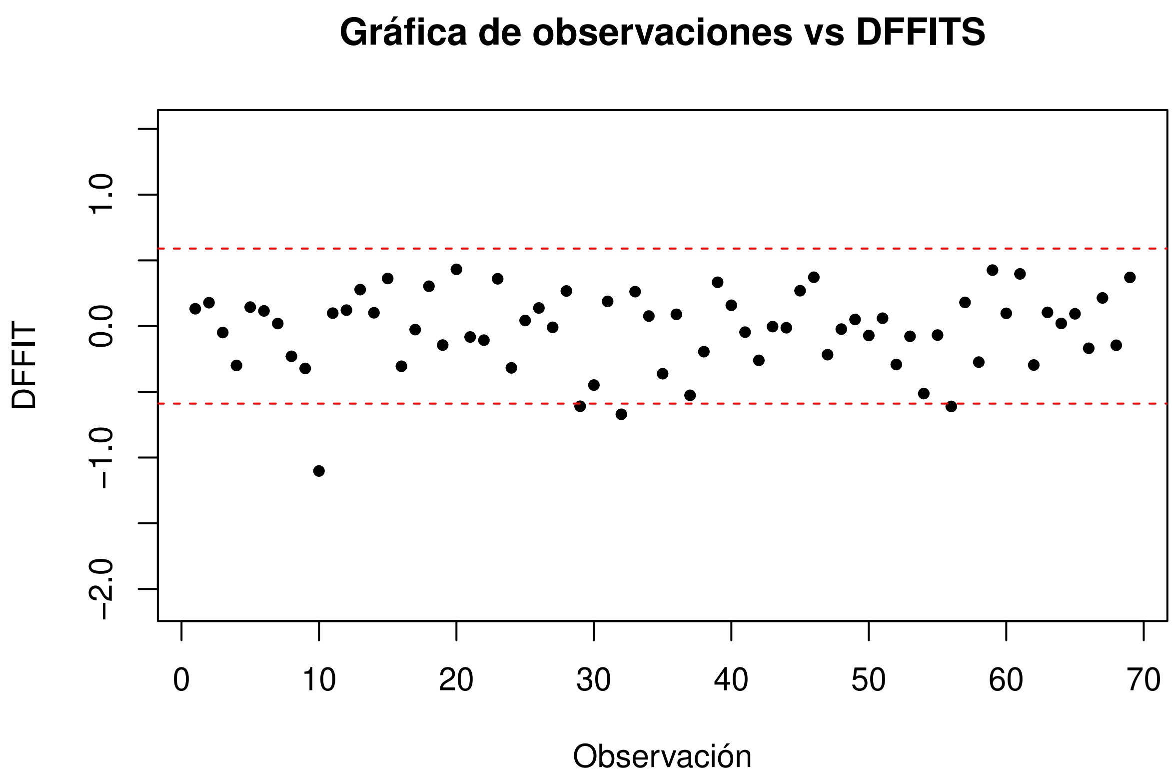


Figura 6: Criterio DFFITS para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 10  -1.6509  0.1968    0.3023 -1.1022
## 29  -2.0065  0.0590    0.0808 -0.6100
## 32  -1.2203  0.0745    0.2308 -0.6711
## 56  -2.2245  0.0583    0.0660 -0.6112
```

3pt

Como se puede ver,las observaciones 10, 29, 32 y 56 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$, con $2\sqrt{\frac{p}{n}} = 0.5897678$, es un punto influencial. Cabe destacar también que con el criterio de distancias de Cook, en el cual para cualquier punto cuya $D_i > 1$, es un punto influencial, ninguno de los datos cumple con serlo.

¿Qué causan?

4.3. Conclusión

2pt

El modelo de regresión presenta ~~preocupaciones~~ significativas en cuanto a si la variabilidad es constante. Estos problemas comprometen la validez de las inferencias realizadas y sugieren que el modelo no se ajusta adecuadamente a los datos. - Una cosa es bondad de ajuste y otra es validez.

Se identificaron observaciones influenciales y puntos de balanceo que tienen un impacto desproporcionado en el modelo. Estas observaciones podrían estar sesgando las estimaciones y afectando la precisión de las predicciones, lo que indica la necesidad de un análisis más detenido y posiblemente su exclusión del modelo para mejorar su validez.

En resumen, debido a las violaciones de los supuestos y la presencia de observaciones extremas, se hace imperativo re-evaluar exhaustivamente el modelo de regresión. Se sugiere explorar transformaciones de datos, evaluar modelos alternativos y, si es posible, recopilar más datos. Actualmente, el modelo carece de validez y confiabilidad para realizar predicciones precisas. Es esencial considerar enfoques alternativos y estrategias de mejora, como la inclusión de datos adicionales o transformaciones adecuadas, para optimizar la calidad y la precisión del modelo de regresión.