Trabajo 1

4,8

Estudiantes

Yamid Andres Campo Gallego
Samuel Gutierrez Osorio
Juan Miguel Marquez Baron
Abraham David Miguel Cardenas

Equipo 32

Docente

Mateo Ochoa Medina

Asignatura

Estadística II

Sede Medellín

5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

19 pt

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

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

Donde la variable dependiente y las variables regresoras son:

- 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:

3 pt

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-0.7108
β_1	0.1666
β_2	0.0170
β_3	0.0489
β_4	0.0157
β_5	0.0019

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.7108 + 0.1666X_{1i} + 0.017X_{2i} + 0.0489X_{3i} + 0.0157X_{4i} + 0.0019X_{5i}\ 1 \leqslant i \leqslant 69$$

1.2. Significancia de la regresión

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

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

Cuyo estadístico de prueba es:

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

5 pt

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	66.7326	5	13.34652	13.0969	9.54197e-09
Error	64.2007	63	1.01906		

De la tabla ANOVA anterior se obtienen los valores del estadístico de prueba F_0=13.0969 y su correspondiente valor-P vp=9.54197e-09.

Con un α=0.05, se tiene que vp $<$ α , por lo cual se rechaza la hipótesis nula H_0 en la que β_j=0 con 1$\leq$ j $\leq$ 5, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$, por lo tanto el RLM propuesto es significativo, indicando que el riesgo de infección (Y) depende considerablemente de alguna de las variables predictoras.

1.3. Significancia de los parámetros

$$\begin{cases} \mathrm{H}_0 : \beta_j = 0 \\ \mathrm{H}_1 : \beta_j \neq 0 \text{ para } j = 0, 1, ..., 5 \end{cases}$$

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

6 pt

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-0.7108	1.6036	-0.4432	0.6591
β_1	0.1666	0.0804	2.0725	0.0423
β_2	0.0170	0.0300	0.5665	0.5730
β_3	0.0489	0.0159	3.0848	0.0030
β_4	0.0157	0.0090	1.7449	0.0859
β_5	0.0019	0.0008	2.3670	0.0210

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_1, β_3 y β_5 son significativos, debido a que, sus P-valores son menores a α.

1.4. Interpretación de los parámetros

Con un nivel de significancia $\alpha = 0.05$ concluimos que los parámetros individuales β_1, β_3 y β_5 son significativos en presencia de los demás parámetros del modelo, también concluimos que los parámetros β_0, β_2 y β_4 no son individualmente significativos en presencia de los demás parámetros del modelo.

-1 pt

Interpreten sólo los parámetros significativos, respecto a β_0 ya saben que se debe cumplir que el 0 esté en el intervalo

$\hat{\beta}_1 = 0.0423$ indica que por cada unidad que aumente la duración promedio de la estadía (X_1) el promedio del riesgo de infección aumenta en 0.1666 unidades, cuando las demás predictoras se mantienen fijas

$\hat{\beta}_3 = 0.0489$ nos indica que por cada unidad que aumente el número promedio de camas (X_3) en el hospital durante el periodo del estudio el promedio del riesgo de infección aumenta en 0,0489 unidades teniendo que las demás variables predictoras se mantienen fijas.

3 pt

$\hat{\beta}_5 = 0.0019$ indica que por cada unidad que aumente el número promedio de enfermeras (X_5), el riesgo de infección aumenta en 0.0019 unidades, esto cuando las demás variables se mantienen constantes.

1.5. Coeficiente de determinación múltiple R^2

3 pt

$$R^2 = \frac{SSR}{SST} = \frac{66.7326}{66.7326 + 64.2007} = 0.5096686634 \tag{2}$$

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.509668$, lo que significa que aproximadamente el 50.97 % de la variabilidad total del riesgo de infección es explicada por el modelo de regresión multiple propuesto.

Por otro lado, se puede calcular el R2 ajustado como una medida de bondad de ajuste, así:

$$R^2_{adj} = 1 - \frac{(n-1)MSE}{SST} = 1 - \frac{(69-1)1.01906}{66.7326 + 64.2007} = 0.4707528184 \tag{3}$$

El valor de R^2_{adj}=0.470752 es menor que R^2=0.5097, lo que indica que en el modelo pueden haber variables que no aporten significativamente. En otras palabras, se puede depurar el modelo.

2. Pregunta 2 5pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más pequeño en el modelo fueron 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:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_3 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 3, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	64.201	X1 X2 X3 X4 X5
Modelo reducido	103.263	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

Se construye el estadístico de prueba como:

3pt

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_4) - SSE(\beta_0, \ \cdots, \ \beta_5))/3}{MSE(\beta_0, \ \cdots, \ \beta_5)} \overset{H_0}{\sim} f_{3,63} \\ &= \frac{103.263 - 64.201}{1.01906} \\ &= 38.33140345 \end{aligned} \tag{4}$$

2pt

Ahora, comparando el F_0 con $f_{0.95,3,63} = 2.7505$, se puede ver que $F_0 > f_{0.95,3,63}$

Se rechaza H_0 y se concluye que el riesgo de infección depende de al menos una variable del subconjunto.

3. Pregunta 3

5pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se quiere probar $H_0 : \beta_1 = \beta_2,\ \beta_3 = \beta_4$ versus una hipotesis alternativa, por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2;\ \beta_3 = \beta_4 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

reescribiendo matricialmente:

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

2pt

Con $\mathbf{L}$ dada por

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

El modelo reducido está dado por:

RM:

$$Y_i = \beta_o + \beta_2 X_{2i}^* + \beta_4 X_{4i}^* + \beta_5 X_{5i} + \varepsilon_i \ 1 \leqslant i \leqslant 69$$

Donde $X_{2i}^* = X_{1i} + X_{2i}$ y $X_{4i}^* = X_{3i} + X_{4i}$

1pt

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$\begin{aligned} F_0 &= \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,63} \\ F_0 &= \frac{(SSE(MR) - 64.2007)/2}{1.01906} \overset{H_0}{\sim} f_{2,63} \end{aligned} \tag{5}$$

2pt

SSE(RM) no se puede obtener de la tabla de todas las regresiones posibles, ya que ésta no admite sumas de variables entre sus opciones.

4. Pregunta 4

4 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:

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

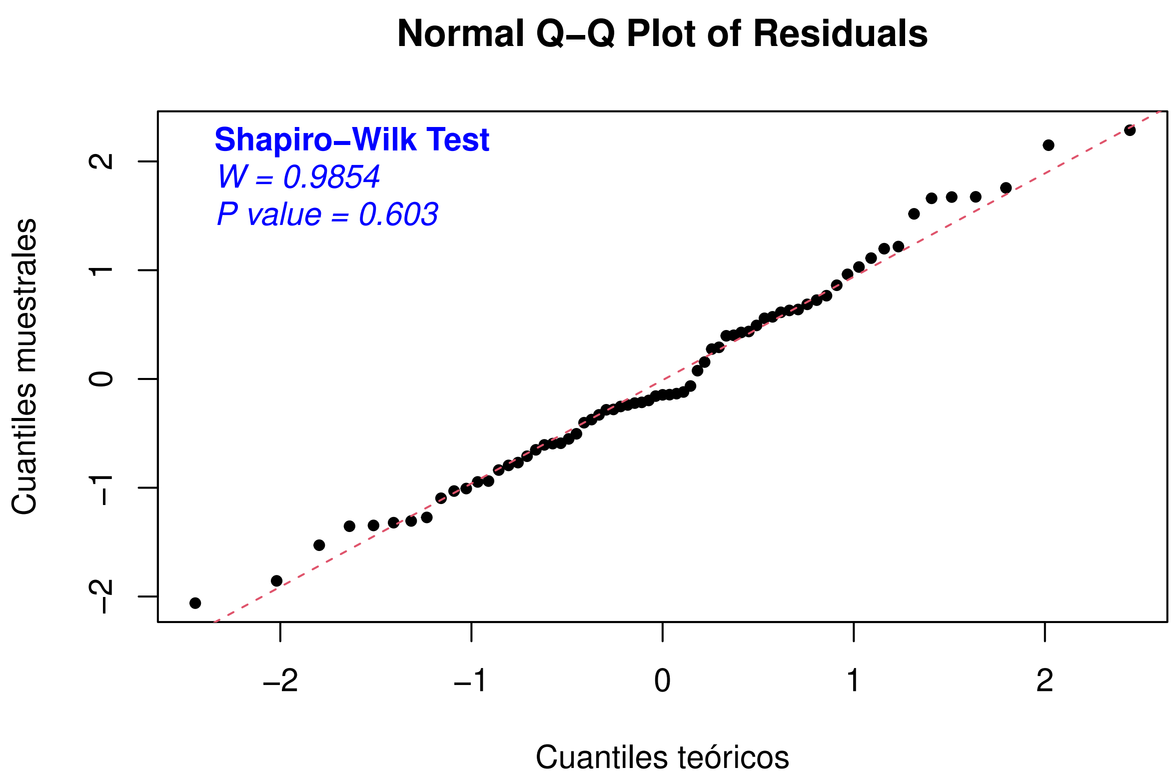


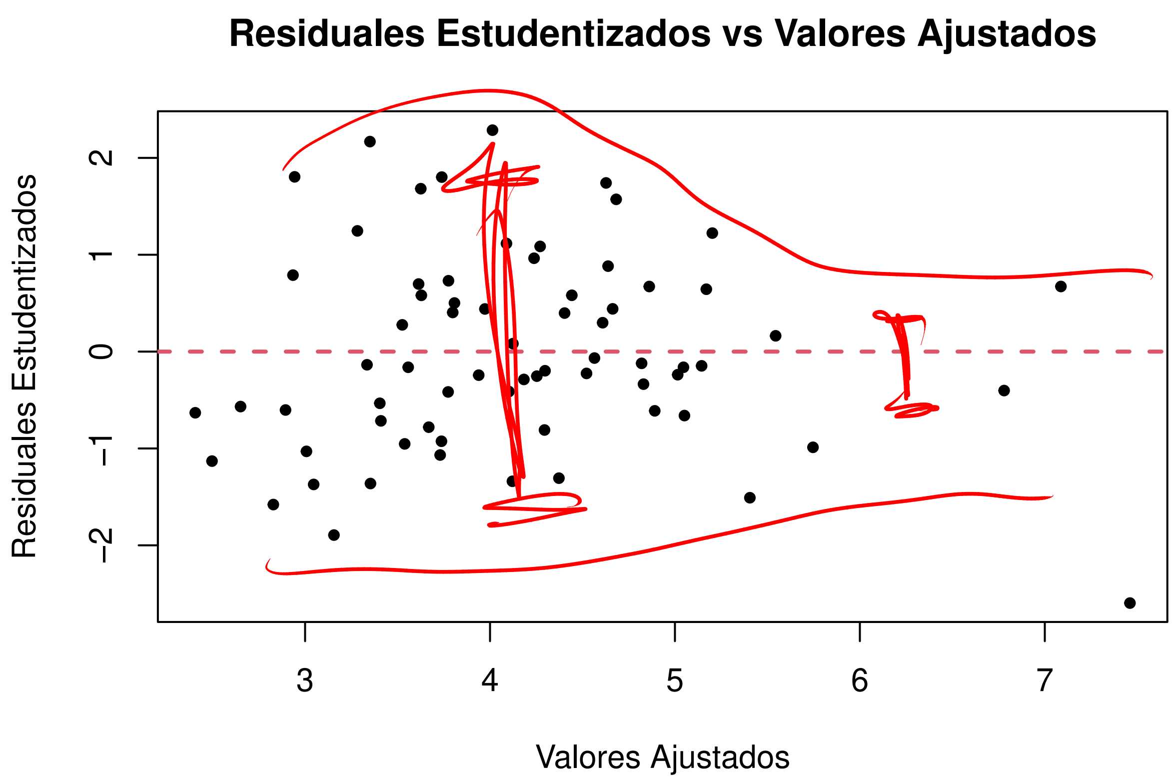
Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

3 pt

Al ser el P-valor aproximadamente igual a 0.603 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. la grafica de comparación de cuantiles permite ver un patrón de los residuales donde siguen la linea roja sin tener colas o influencias muy abruptas, entonces se concluye que, por este motivo el suspuesto de normalidad se cumple.

→ tenía patrón irregular

4.1.2. Varianza constante



3pt

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que hay un patrón de aumento y decrecimiento de la varianza, ya que al inicio de las observaciones se muestran datos más dispersos, pero a medida que se acerca al centro la dispersión de los datos disminuye, por lo cual se puede concluir que no se cumple el supuesto de varianza constante, además se visualiza la presencia de observaciones extremas.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

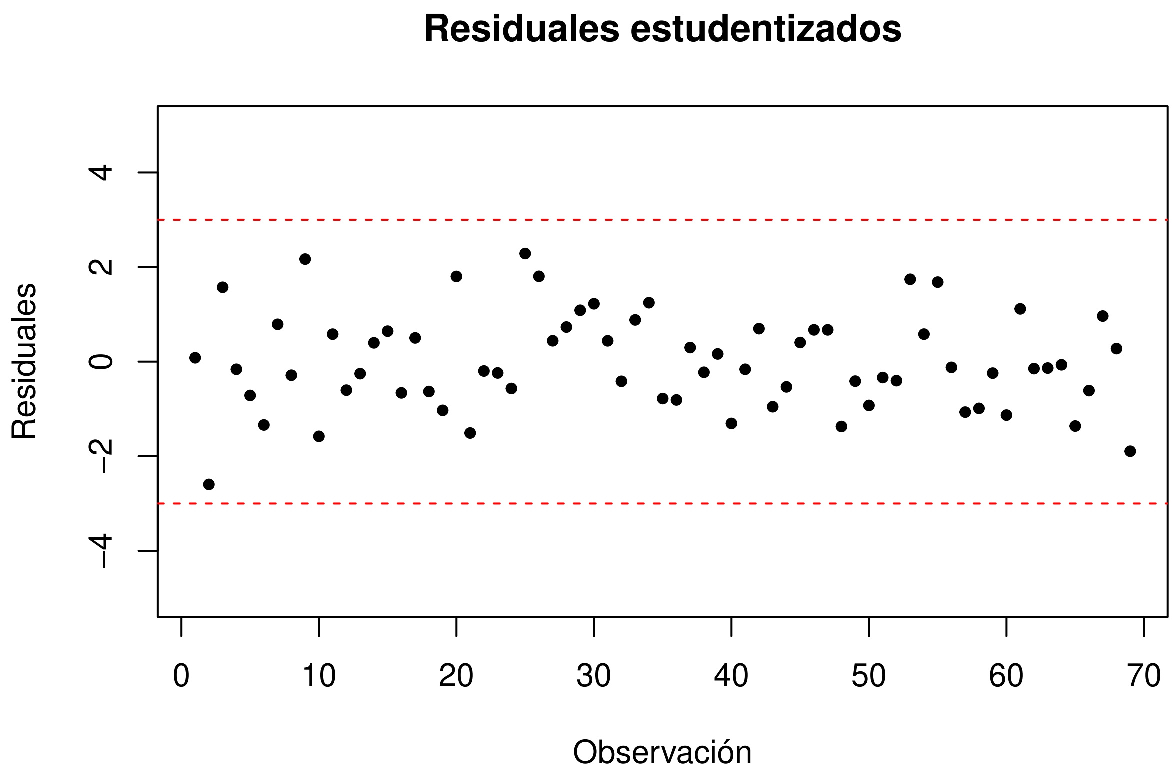


Figura 3: Identificación de datos atípicos

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 $|r_{estud}| > 3$.

3pt

4.2.2. Puntos de balanceo

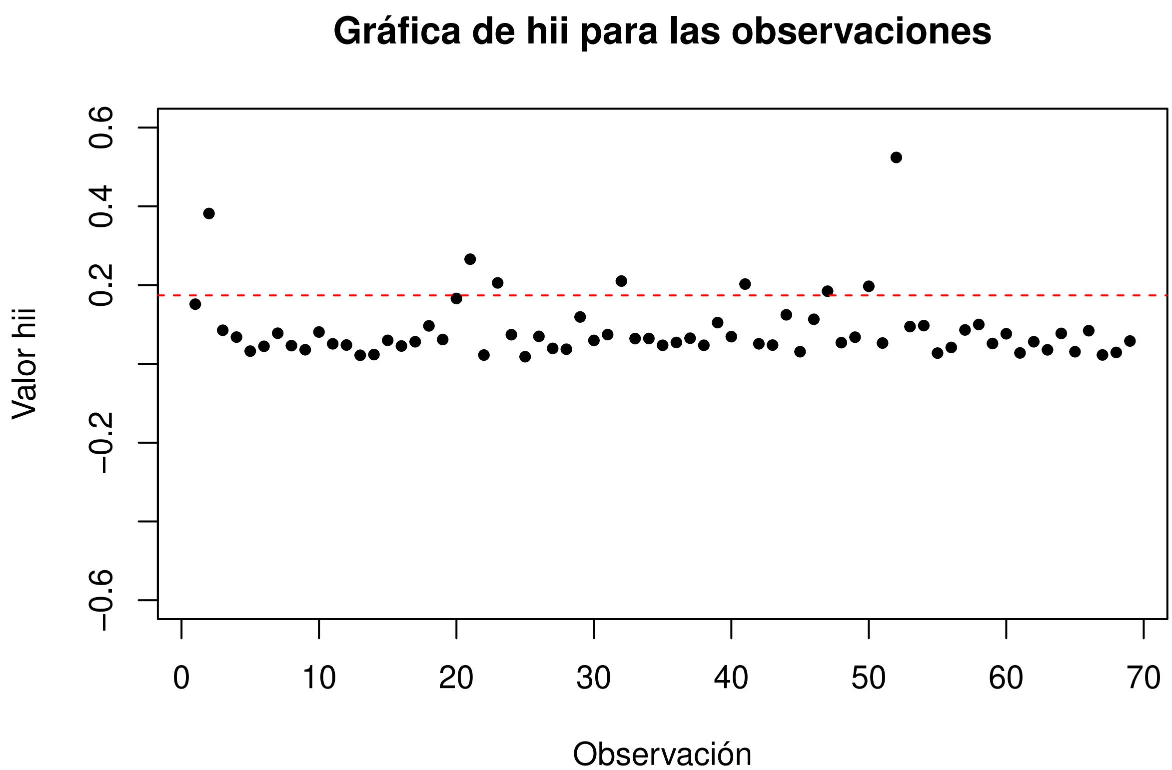


Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 2   -2.5964  0.6942    0.3819 -2.1425
## 21  -1.5090  0.1375    0.2659 -0.9177
## 23  -0.2389  0.0025    0.2058 -0.1207
## 32  -0.4163  0.0077    0.2104 -0.2135
## 41  -0.1620  0.0011    0.2027 -0.0810
## 47   0.6723  0.0171    0.1846  0.3185
## 50  -0.9258  0.0351    0.1971 -0.4582
## 52  -0.4023  0.0297    0.5242 -0.4194
```

3p7

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 2\frac{6}{69} = 0.1739$, se puede apreciar que existen 8 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{p}{n}$, los cuales son los presentados en la tabla.

Estos puntos de balanceo posiblemente afecten al R^2, ya que los puntos de balanceo aumentan el valor de este, al encontrarse por encima de la tendencia central de las observaciones, esto se debe a que el modelo intentará ajustarse a estos puntos atípicos, lo que resulta en un mayor coeficiente de determinación. Además también se pueden ver afectados los coeficientes estimados de los errores estándar

4.2.3. Puntos influenciales

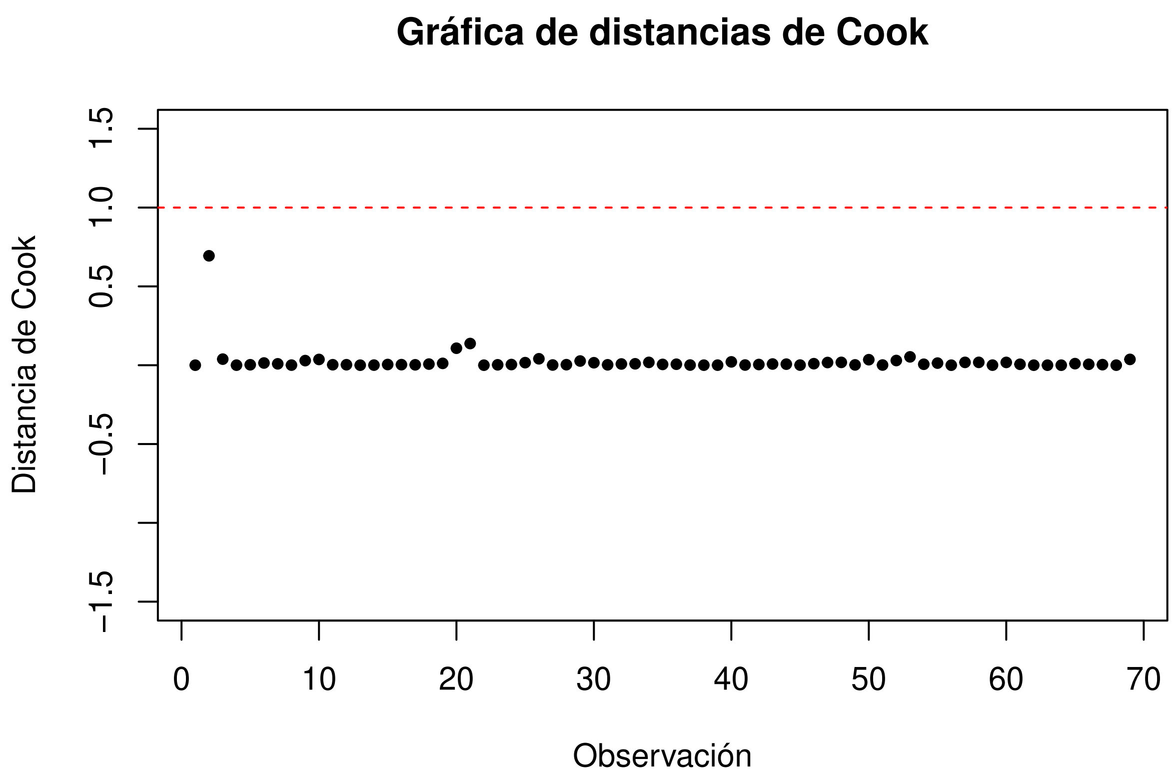


Figura 5: Criterio distancias de Cook para puntos influenciales

Una observación es influencial si cumple el criterio $|D_i| > 1$

Por la grafica de distancias de Cooks podemos concluir que no hay ningún punto influencial, debido a que ninguno cumple el criterio de $|D_i| > 1$

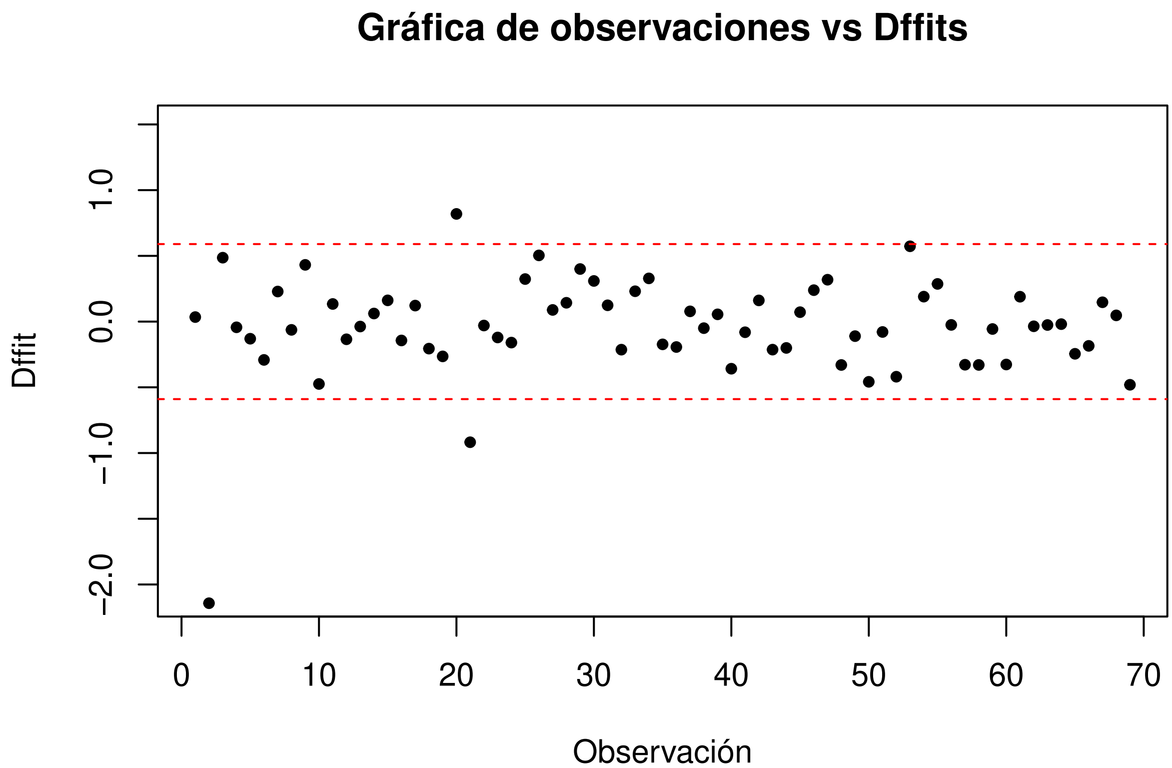


Figura 6: Criterio Dffits para puntos influenciales

```
##     res.stud Cooks.D hii.value  Dffits
## 2    -2.5964  0.6942    0.3819 -2.1425
## 20    1.8017  0.1078    0.1661  0.8191
## 21   -1.5090  0.1375    0.2659 -0.9177
```

4pt

Como se puede ver,las observaciones 2, 20 y 21 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{6}{69}} = 0.5897678$, 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.

Se debe realizar un analisís sobre estos puntos influenciales, debido a que, estos tienen impacto notable sobre los coeficientes de regresión ajustados, lo cual conlleva a resultados engañosos en el modelo de regresión, ya que dichos datos empujan la línea de regresión en una dirección particular, lo que altera el ajuste del modelo de datos. El modelo se vuelve altamente influenciable por estos datos y si hay pequeños cambios en estos puede resultar en cambios significativos en los coeficientes de la regresión, por tanto en el ajuste y confiablidad del modelo.

4.3. Conclusión

3pt

A pesar de que el modelo cumple el supuesto de distribución normal de los errores, vemos que no cumple el de varianza constante, por lo qué podemos afirmar que el modelo no es válido y probablemente se esté viendo afectado por datos extremos dentro del mismo, se debe hacer una investigación sobre los 3 datos influenciales y los 8 de balanceo, y determinar que hacer con ellos si se planea la formulación de otro modelo de regresión.