# Trabajo 1

4,9

Estudiantes


**Juan Jose Triana Higuita**
**Pablo Ochoa Palacio**
**Julian Alexander Ruiz Ocampo**
**Juan Andres Jimenez Velez**

Equipo 26


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 p+

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

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

A continuación se muestran el nombre de las variables:

- 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 |
|---|---|
| $\beta_0$ | 0.1720 |
| $\beta_1$ | 0.2446 |
| $\beta_2$ | -0.0054 |
| $\beta_3$ | 0.0475 |
| $\beta_4$ | 0.0105 |
| $\beta_5$ | 0.0023 |

3 p +

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 0.172 + 0.2446 X_{1i} - 0.0054 X_{2i} + 0.0475 X_{3i} + 0.0105 X_{4i} + 0.0023 X_{5i};\ 1 \leqslant i \leqslant 69$$

## 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:

5pt

Cuadro 2: Tabla ANOVA para el modelo

| | Sumas de cuadrados | g.l. | Cuadrado medio | $F_0$ | P-valor |
|---|---|---|---|---|---|
| Regresión | 61.5923 | 5 | 12.31847 | 12.1044 | 3.12145e-08 |
| Error | 64.1140 | 63 | 1.01768 | | |

De la tabla Anova, se observa un valor P aproximadamente igual a 0, 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. De esta manera, considerando la tabla Anova, notamos que hay una relación entre la variable respuesta y alguna o todas las variables predictorias. Así, el riesgo de infección, puede ser explicado: o por la duración de la estadía y/o la rutina de cultivos y/o el número de camas y/o el censo promedio diario y/o el número de enfermeras.

## 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 |
|---|---|---|---|---|
| $\beta_0$ | 0.1720 | 1.6012 | 0.1074 | 0.9148 |
| $\beta_1$ | 0.2446 | 0.1054 | 2.3199 | 0.0236 |
| $\beta_2$ | -0.0054 | 0.0299 | -0.1789 | 0.8586 |
| $\beta_3$ | 0.0475 | 0.0176 | 2.7008 | 0.0089 |
| $\beta_4$ | 0.0105 | 0.0071 | 1.4725 | 0.1459 |
| $\beta_5$ | 0.0023 | 0.0007 | 3.0842 | 0.0030 |

8pt

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros $\beta_1$, $\beta_3$ y $\beta_5$ son significativos, pues sus P-valores son menores a $\alpha$.

## 1.4. Interpretación de los parámetros

A continuación interpretamos los parámetros significativos. Respecto a $\beta_0$ ya sabemos que se debe cumplir que el 0 esté en el intervalo, por lo que en esta ocasión dicho parámetro no es significativo.

$\hat{\beta}_1$**: Por cada día que aumenta la duración promedio de la estadía de los pacientes en el hospital, la probabilidad promedio estimada de adquirir infección en el hospital aumenta en promedio 24.46 %, cuando las demás variables permanecen constantes.**

$\hat{\beta}_3$**: Por cada aumento en una unidad en el número promedio de camas en el hospital, la probabilidad promedio estimada de adquirir infección en el hospital aumenta en promedio 4.75 %, cuando las demás variables permanecen constantes.**

$\hat{\beta}_5$**: Por cada aumento en una unidad en el número promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio, la probabilidad promedio estimada de adquirir infección en el hospital aumenta en promedio 0.23 %, cuando las demás variables permanecen constantes.**

### 1.5. Coeficiente de determinación múltiple $R^2$

2pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.4899$, lo que significa que aproximadamente el 48.99 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión ajustado propuesto en el presente informe.

¿Cómo se calcula?

## 2. Pregunta 2

5pt

### 2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más bajo 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 | 64.114 | X1 X2 X3 X4 X5 |
| Modelo reducido | 111.105 | X2 X4 |

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

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

3pt

## 2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_4) - SSE(\beta_0,\ \cdots,\ \beta_5))/3}{MSE(\beta_0,\ \cdots,\ \beta_5)} \overset{H_0}{\sim} f_{3,63} \\ &= \frac{15.66367}{1.01768} \\ &= 15.39154 \end{aligned} \tag{2}$$

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 tanto se rechaza la hipótesis nula en la que que $\beta_j = 0$ para $j = 1, 3, 5$, aceptando la hipótesis alternativa en la que algún $\beta_j$ distinto de 0 para $j = 1, 3, 5$, lo cual indica que el subconjunto de parámetros es significativo, por lo tanto, no es posible descartar estas variables.

# 3. Pregunta 3

4pt

## 3.1. Prueba de hipótesis y prueba de hipótesis matricial

Nos hacemos las siguientes preguntas: ¿El efecto de la duración de la estadía sobre el riesgo de infección es igual al efecto del censo promedio diario sobre el riesgo de infección? y ¿El efecto de la rutina de cultivos sobre el riesgo de infección es igual a 2 veces el efecto del número de camas sobre el riesgo de infección? A partir de esto planteamos la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_4;\ \ \beta_2 = 2\beta_3 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

reescribiendo matricialmente:

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

Con $\mathbf{L}$ dada por

2pt

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

El modelo reducido está dado por:

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

0pt

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

$X_{2i}^* = 2X_{2i} + X_{3i}$

### 3.2. Estadístico de prueba

El estadístico de prueba $F_0$ está dado por:

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

2pt

A continuación se reemplazan lo valores conocidos SSE(MF) y MSE(MF) en el estadístico de prueba

$$F_0 = \frac{(SSE(MR) - 64.1140)/2}{1.01768} \overset{H_0}{\sim} f_{2,63} \tag{4}$$

## 4. Pregunta 4

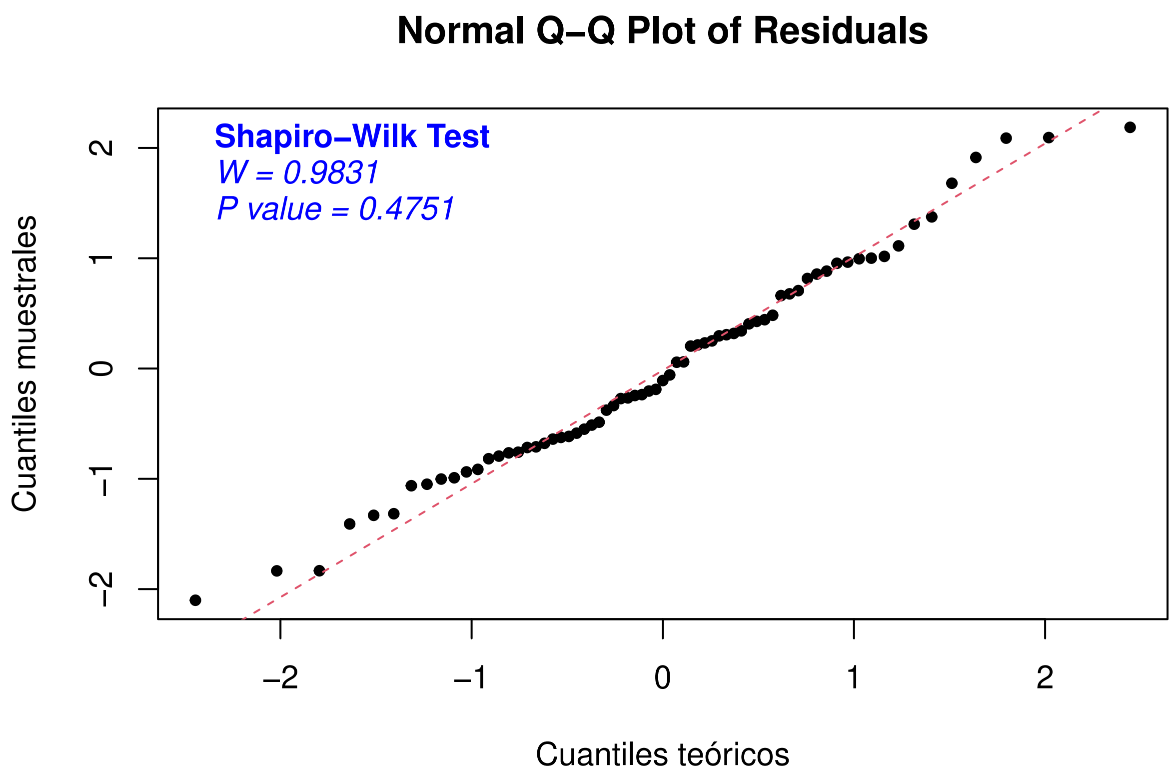
16pt

### 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:

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




9pt

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

El P-valor arrojado por el Shapiro-Wilk test es de 0.4751, fijando el nivel de significancia en $\alpha = 0.05$, tenemos que el P-valor es mayor, esto nos indica que la hipótesis nula no debe ser rechzada y con esto, asumir que en el modelo los datos se distribuyen normalmente con una media $\mu$ y una varianza $\sigma^2$, sin embargo, en la gráfica Cuantiles teóricos vs Cuantiles muestrales, se observa que tanto la cola inferior como la cola superior presentan patrones irregulares bastante notorios, mientras que en el resto de puntos también se pueden observar ciertas irregularidades no tan pronunciadas. A partir de esto y del hecho de que el análisis gráfico tiene más peso sobre la determinación de la normalidad de un modelo, podemos rechazar la hipótesis nula y con esta el cumplimiento del supuesto de normalidad.

Ahora se validará si la varianza cumple con el supuesto de ser constante.

#### 4.1.2. Varianza constante

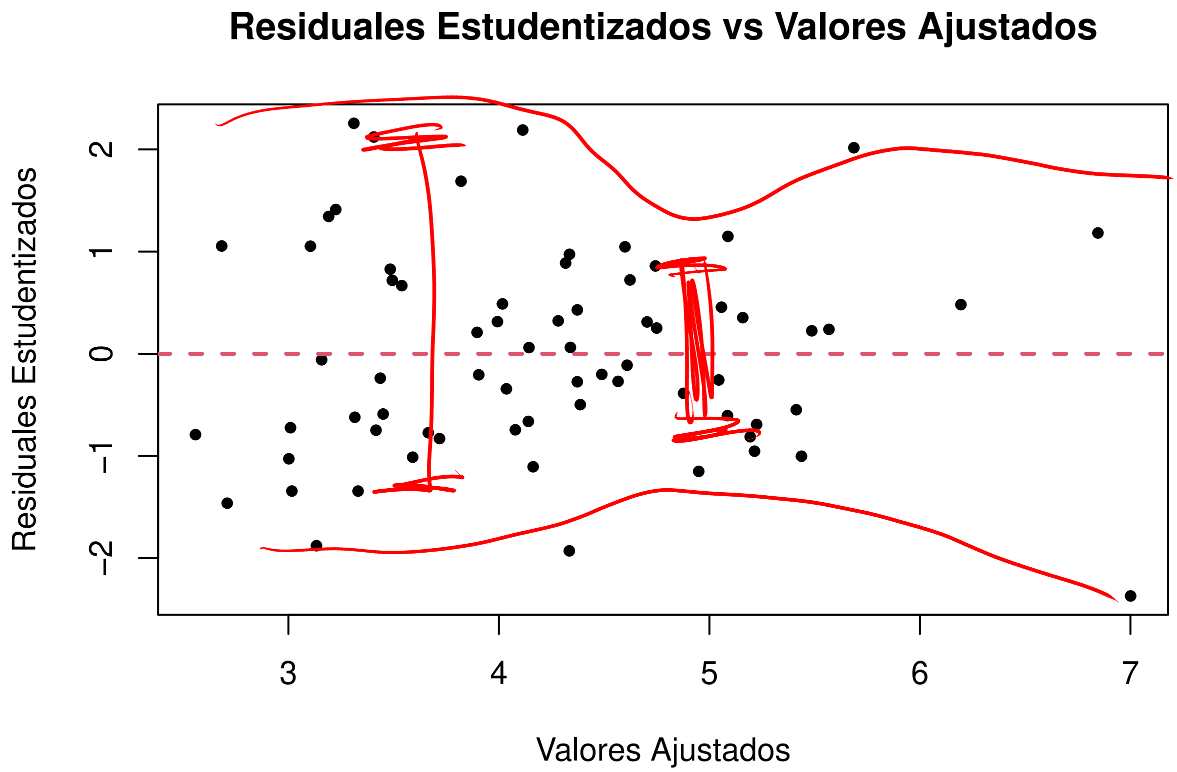



Figura 2: Gráfico residuales estudentizados vs valores ajustados

1,5 pt

En el gráfico de Valores Ajustados vs Residuales Estudentizados se logra apreciar que la varianza es constante, ya que presenta patrones regulares, es decir homocedasticidad. Este análisis gráfico permite deducir que la varianza en el modelo es constante y por lo tanto, cumple con tal supuesto. En el gráfico también se observa que la media del modelo es 0.

X

## 4.2. Verificación de las observaciones

### 4.2.1. Datos atípicos

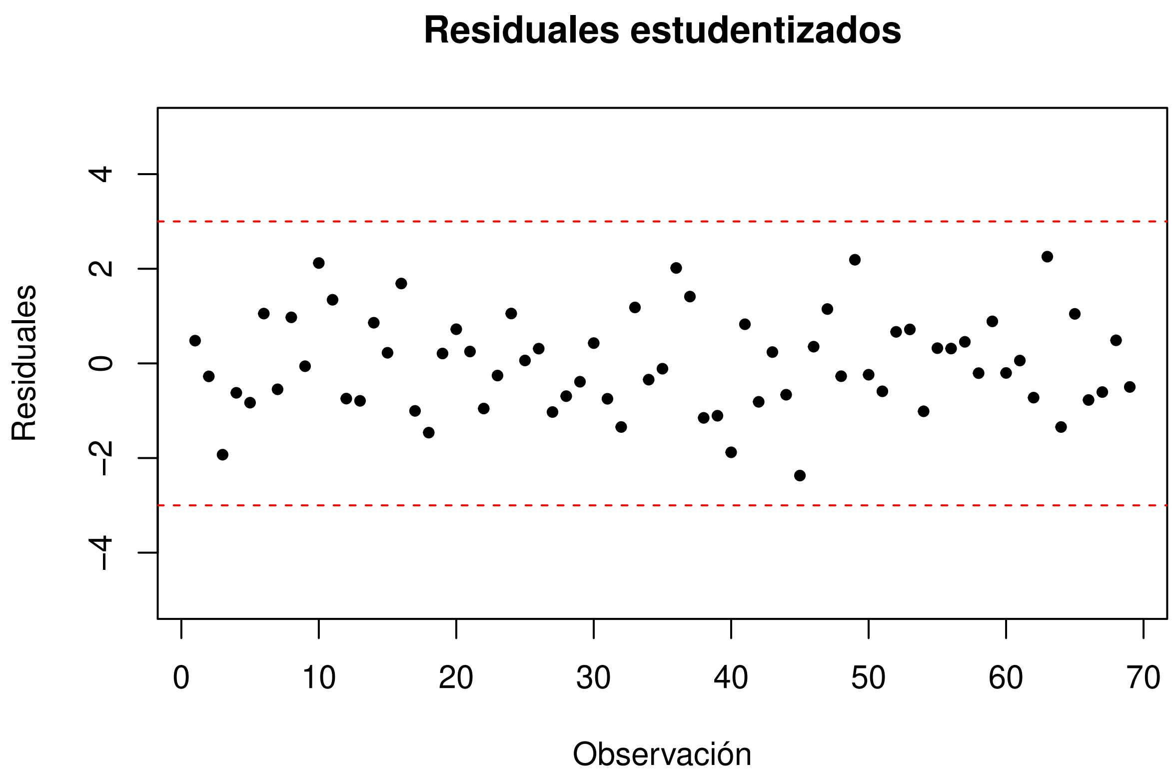



Figura 3: Identificación de datos atípicos

3pt

A partir del análisis de la gŕafica anterior se puede notar que no hay datos atípicos en el conjunto de datos, debido a que ninguno de los residuales estandarizados sobrepasa el criterio $|r_{estud}| > 3$. La ausencia de datos atípicos indica que no hay puntos de datos que se desvíen significativamente de la tendencia general de los datos. Esto sugiere que nuestros datos son relativamente coherentes y que no hay observaciones extremas que puedan sesgar de manera significativa las estimaciones de los coeficientes del modelo.

### 4.2.2. Puntos de balanceo

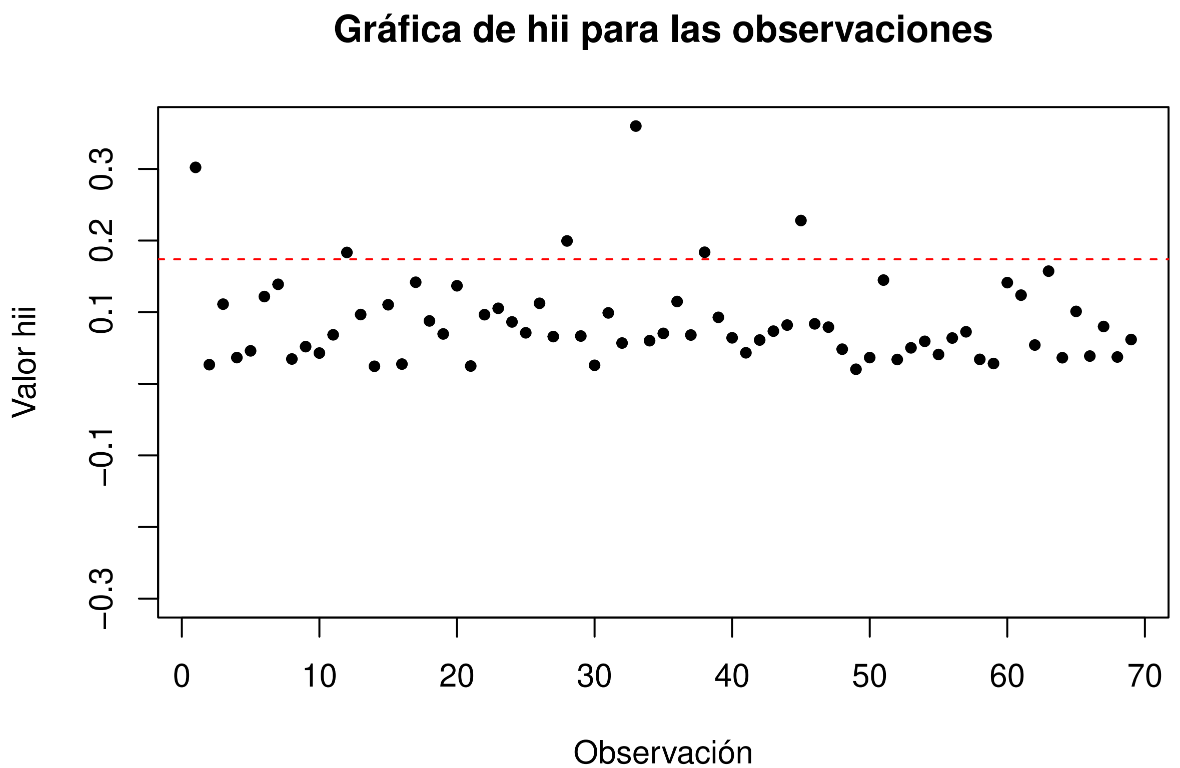



Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 1    0.4808  0.0167    0.3022  0.3144
## 12  -0.7439  0.0207    0.1834 -0.3513
## 28  -0.6918  0.0199    0.1995 -0.3439
## 33   1.1823  0.1310    0.3599  0.8895
## 38  -1.1511  0.0497    0.1837 -0.5475
## 45  -2.3703  0.2765    0.2280 -1.3389
```

1,5 pt

Al observar la gráfica de observaciones vs valores $h_{ii}$, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$, se puede apreciar que existen 6 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. La existencia de estos puntos indica ~~que por lo menos dos variables independientes en el modelo están altamente correlacionadas entre sí.~~

eso no indican

### 4.2.3. Puntos influenciales

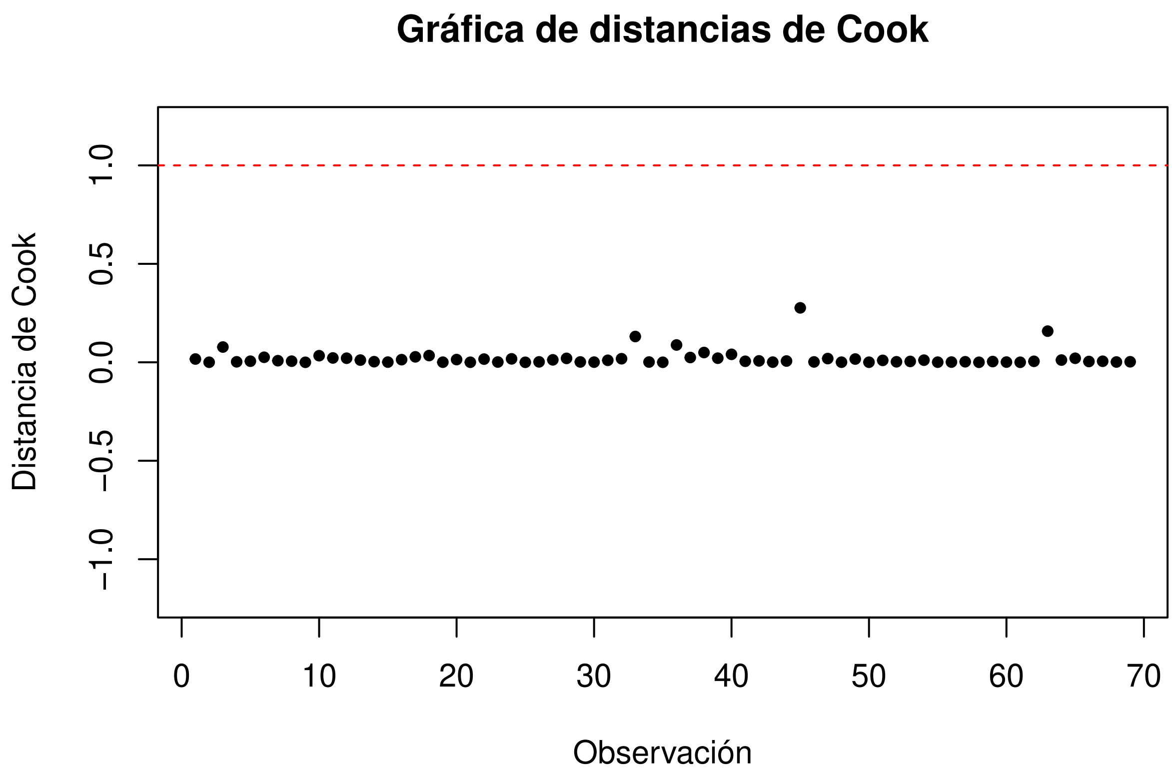



Figura 5: Criterio distancias de Cook para puntos influenciales

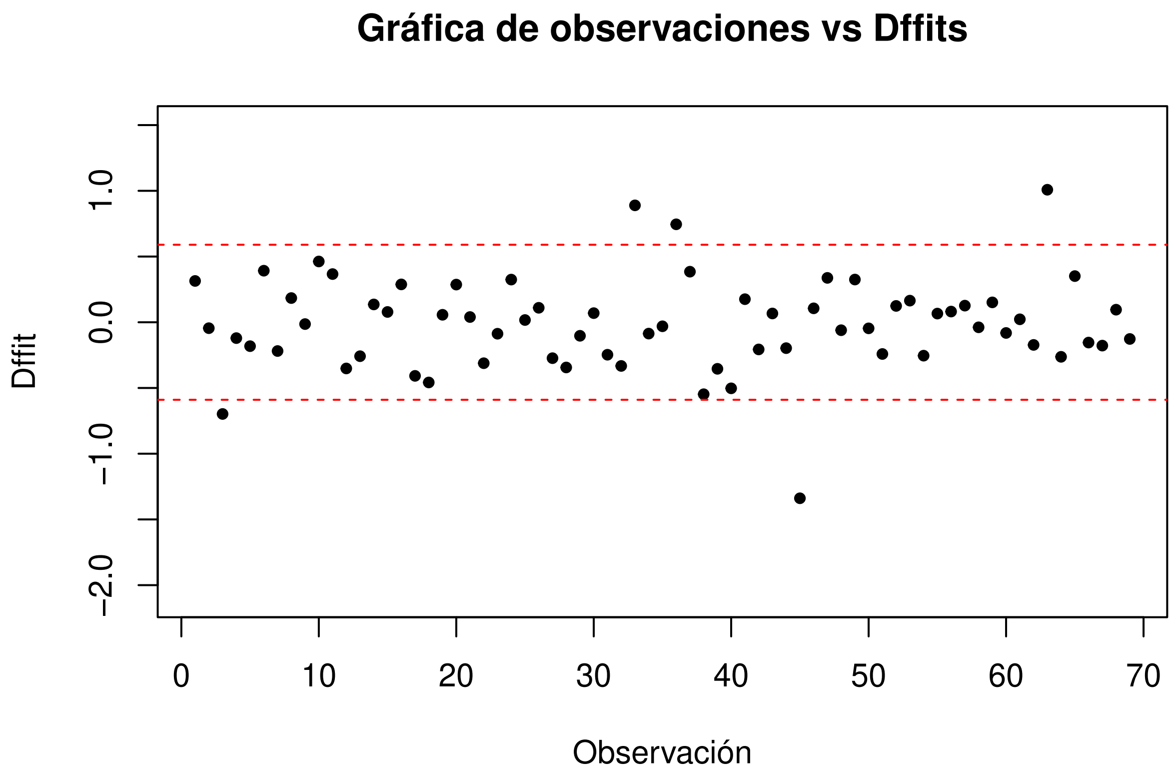



Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 3   -1.9289  0.0776    0.1112 -0.6977
## 33   1.1823  0.1310    0.3599  0.8895
## 36   2.0165  0.0880    0.1149  0.7453
## 45  -2.3703  0.2765    0.2280 -1.3389
## 63   2.2556  0.1582    0.1573  1.0082
```

3pt

A partir del grafico y la información suministrada por la tabla, las observaciones 3, 33, 36, 45 y 63 son puntos influenciales según el criterio de Dffits, dicho critero dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, es un punto influencial, sin embargo, de acuerdo al 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 esta característica.

¿Qué causan?

## 4.3. Conclusión

3pt

Concluimos inicialmente que nuestro modelo no es valido dado que no cumple con el supuesto de normalidad a partir del analisis grafico que se realizó.

En resumen, es importante destacar que los puntos atípicos relacionados con el balanceo no tienen un efecto directo en la normalidad del modelo. Sin embargo, afirmar categóricamente que los puntos de balanceo impactan la normalidad sería impreciso, ya que su influencia puede variar. Lo correcto es indicar que estos puntos podrían estar incidiendo en la normalidad. Además, es importante señalar que evaluar exhaustivamente si estos puntos están afectando la normalidad resultaría complicado. Por lo tanto, es suficiente reconocer que existe la posibilidad de que estén teniendo un efecto.