4,9

# Trabajo 1

Estudiantes


**Tomas Castilla Barrero**
**Angel David Lopez Gamero**
**Julian Antonio Narvaez Romo**


Equipo: 20

Docente

**Julieth Vernoica Guarín 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

19pt

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ \ 1 \leqslant i \leqslant 74$$

Donde:

- Y:Probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje).
- $X_1$:Duración promedio de la estadía de todos los pacientes en el hospital (en días).
- $X_2$:Razón del número de cultivos realizados en pacientes sin síntomas de infección hospitalaria, por cada 100.
- $X_3$:Número promedio de camas en el hospital durante el periodo del estudio.
- $X_4$:Número promedio de pacientes en el hospital por día durante el periodo del estudio.
- $X_5$:Número promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio.

## 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 |
|---|---|
| $\beta_0$ | -0.2233 |
| $\beta_1$ | 0.1096 |
| $\beta_2$ | 0.0196 |
| $\beta_3$ | 0.0454 |
| $\beta_4$ | 0.0182 |
| $\beta_5$ | 0.0016 |

2pt

Supuestos y $\varepsilon_i$ no va en ec. ajustada

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.2233 + 0.1096X_{1i} + 0.0196X_{2i} + 0.0454X_{3i} + 0.0182X_{4i} + 0.0016X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 74$$

## 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:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,68} \tag{1}$$

Ahora, se presenta la tabla Anova:

5 pt

Cuadro 2: Tabla ANOVA para el modelo

| | Sumas de cuadrados | g.l. | Cuadrado medio | $F_0$ | P-valor |
|---|---|---|---|---|---|
| Regresión | 60.1213 | 5 | 12.024267 | 12.6406 | 1.08722e-08 |
| Error | 64.6842 | 68 | 0.951238 | | |

De la tabla Anova, se observa un valor P es demasiado pequeño, por lo que a un nivel de significancia de 0.05 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 |
|---|---|---|---|---|
| $\beta_0$ | -0.2233 | 1.5072 | -0.1482 | 0.8826 |
| $\beta_1$ | 0.1096 | 0.0841 | 1.3036 | 0.1968 |
| $\beta_2$ | 0.0196 | 0.0271 | 0.7218 | 0.4729 |
| $\beta_3$ | 0.0454 | 0.0127 | 3.5839 | 0.0006 |
| $\beta_4$ | 0.0182 | 0.0069 | 2.6365 | 0.0104 |
| $\beta_5$ | 0.0016 | 0.0007 | 2.2469 | 0.0279 |

6 pt

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros $\beta_3$ ,$\beta_4$ y $\beta_5$ son significativos, pues sus P-valores son menores a $\alpha$.

## 1.4. Interpretación de los parámetros

3pt

$\hat{\beta}_3$:Indica que por cada unidad que se aumenta el número promedio de camas en el hospital durante el periodo del estudio (X3), el porcentaje promedio de la eficacia en el control de infecciones hospitalarias aumenta en 0.0454 unidades, cuando las demás predictoras se mantienen fijas.

$\hat{\beta}_4$:Indica que por cada unidad que se aumenta el número promedio de pacientes en el hospital por día durante el periodo del estudio (X4), el porcentaje promedio de la eficacia en el control de infecciones hospitalarias aumenta en 0.0182 unidades, cuando las demás predictoras se mantienen fijas.

$\hat{\beta}_5$:Indica que por cada unidad que se aumenta el número promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio (X5), el porcentaje promedio de la eficacia en el control de infecciones hospitalarias aumenta en 0.0016 unidades, cuando las demás predictoras se mantienen fijas.

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

3pt

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

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.929458$, lo que significa que aproximadamente el 92.9458 % de la variabilidad total observada en el riesgo de infección es explicada por el modelo propuesto RLM y un $0,7054$% de la variabilidad total es explicada por el error obtenido. sin embargo, esto no implica que el modelo sea adecuado. Para seleccionar un modelo ajustado, se recomienda realizar el $R^2_{adj}$ con un valor obtenido de 44.36 %, como $R^2_{adj}$ es menor que $R^2$, se sugiere que en el modelo pueden haber variables que no contribuyan de manera significativa. Se destaca que esta prueba no indica la validez de un modelo, simplemente es un parámetro de comparación entre ellos.

# 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 $X_3, X_4, X_5$, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

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

Cuadro 4: Resumen tabla de todas las regresiones

| | $SSE$ | Covariables en el modelo |
|---|---|---|
| Modelo completo | 64.684 | X1 X2 X3 X4 X5 |
| Modelo reducido | 94.054 | X1 X2 |

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 74$$

## 2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

3pt

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_2) - SSE(\beta_0,\ \cdots,\ \beta_5))/3}{MSE(\beta_0,\ \cdots,\ \beta_5)} \overset{H_0}{\sim} f_{3,68} \\ &= \frac{3.26333}{0.9512} \\ &= 3.43071 \end{aligned} \tag{3}$$

2pt

Ahora, comparando el $F_0$ con $f_{0.95,3,68} = 2.7395$, se puede ver que $F_0 > f_{0.95,3,68}$ y por tanto rechazamos la hipotesis nula ya que el estadisto de prueba se encuentra en la region de rechazo. No es posible descartar las variables del subconjunto, ya que, se comprobo que al menos uno los coeficientes de las variables del subconjunto con el que se trabajo en el planteamiento de prueba es significativa y distinta de cero. Los resultados son coherentes con las expectativas derivadas de la prueba de significancia de los parámetros individuales, ya que las variables $\beta_3$, $\beta_4$, y $\beta_5$ son significativas para el modelo RLM..

# 3. Pregunta 3

5pt

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

Supunga que se quiere probar que

$$\begin{cases} \text{H}_0 & : \beta_1 = \beta_2;\ \beta_3 = 2\beta_2;\ \beta_4 = \beta_5 \end{cases}$$

versus una hipotesis alternativa.

Se plantea la siguiente prueba de hipotesis:

$$\begin{cases} \text{H}_0 : \beta_1 = \beta_2;\ \beta_3 = 2\beta_2;\ \beta_4 = \beta_5 \\ \text{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

reescribiendo matricialmente:

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

Con $\mathbf{L}$ dada por

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

3pt

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_2 X_{2i}^* + \beta_4 X_{4i}^* + \beta_5 X_{5i} + \varepsilon_i,\ \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 74$$

Donde $X_{2i}^* = (X_{1i} + X_{2i} + 2X_{3i})$ y por otra parte $X_{4i}^* = (X_{4i} + X_{5i})$

## 3.2. Estadístico de prueba

El estadístico de prueba $F_0$ está dado por:

2pt

$$F_0 = \frac{(SSE(MR)^* - 64.684)/3}{0.95123} \overset{H_0}{\sim} f_{3,68} \tag{4}$$

Nota: El $SSE(RM)^*$ 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

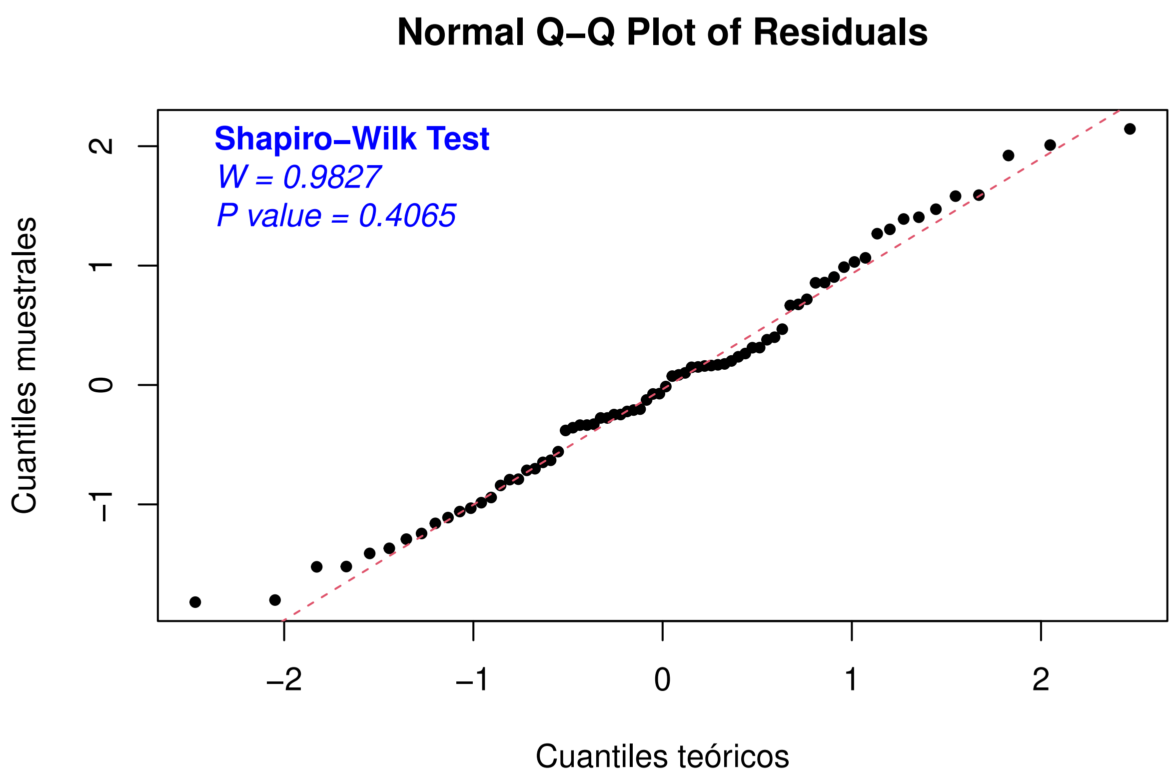
20pt

## 4.1. Supuestos del modelo

### 4.1.1. Normalidad de los residuales

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

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

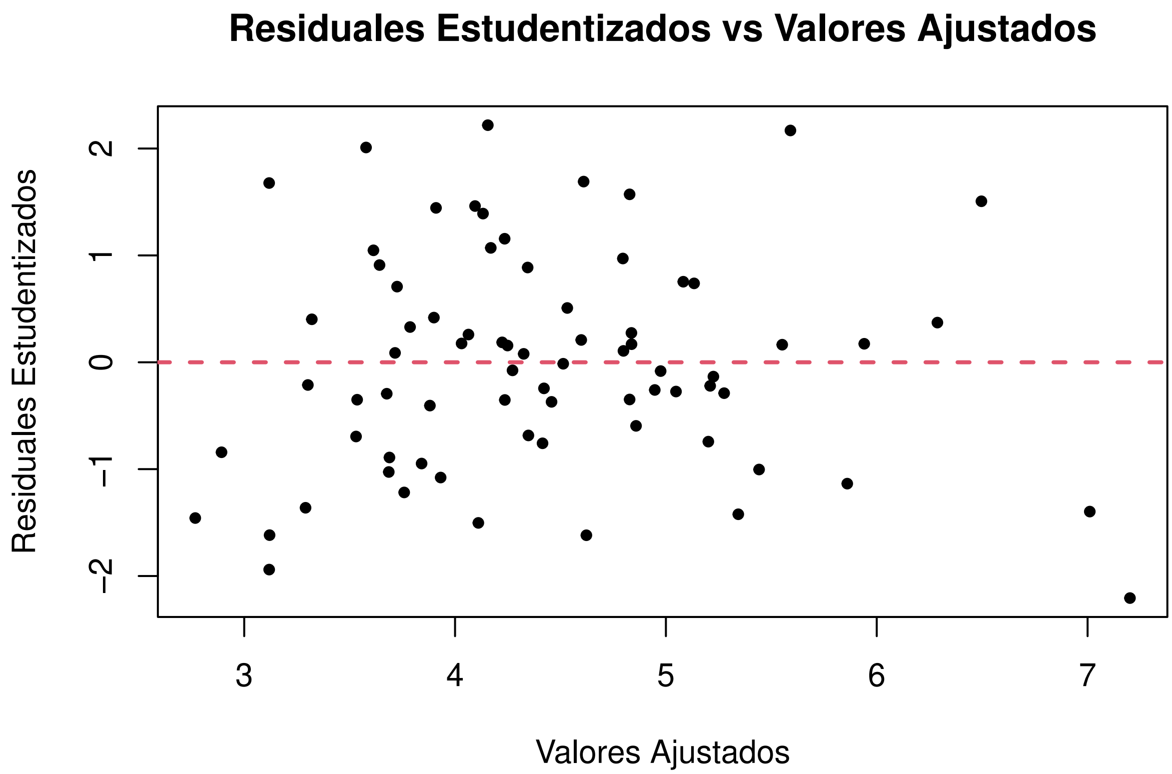



4pt

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

En un análisis gráfico se observa que el patrón de residuales se acomoda a la línea punteada roja que representa el ajuste de los residuales a una distribución normal, sin embargo, hay presencia de dispersión de puntos con respecto a esta recta, las cuales podrian haber sido causadas por puntos de balanceo o observaciones atipicas que se presenten en la muestra. A pesar de estos hallazgos, la manifestación de estos patrones es mínima y se puede concluir que los residuales se ajustan satisfactoriamente a la distribución normal. Este resultado se respalda con la prueba hipótesis de normalidad por medio de Shapiro -Wilk, con P-valor aproximadamente igual a 0.4065. Considerando el nivel de significancia de $\alpha = 0.05$, se resuelve que el P-valor es significativamente mayor, lo que implica que no se rechazaría la hipótesis nula que sostiene la distribución normal para residuales en el modelo.

#### 4.1.2. Varianza constante




3pt

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados $r_i$ vs valores ajustados $\hat{Y}_i$ se puede observar que el patrón de la nube de puntos mantiene una dispersión uniforme, del mismo modo, no se observa evidencia de la necesidad de falta de ajuste en este modelo, esto nos permite concluir que el supuesto de varianza constante se cumple y el modelo se adapta al lineal múltiple.

## 4.2. Verificación de las observaciones

### 4.2.1. Observaciones atípicas

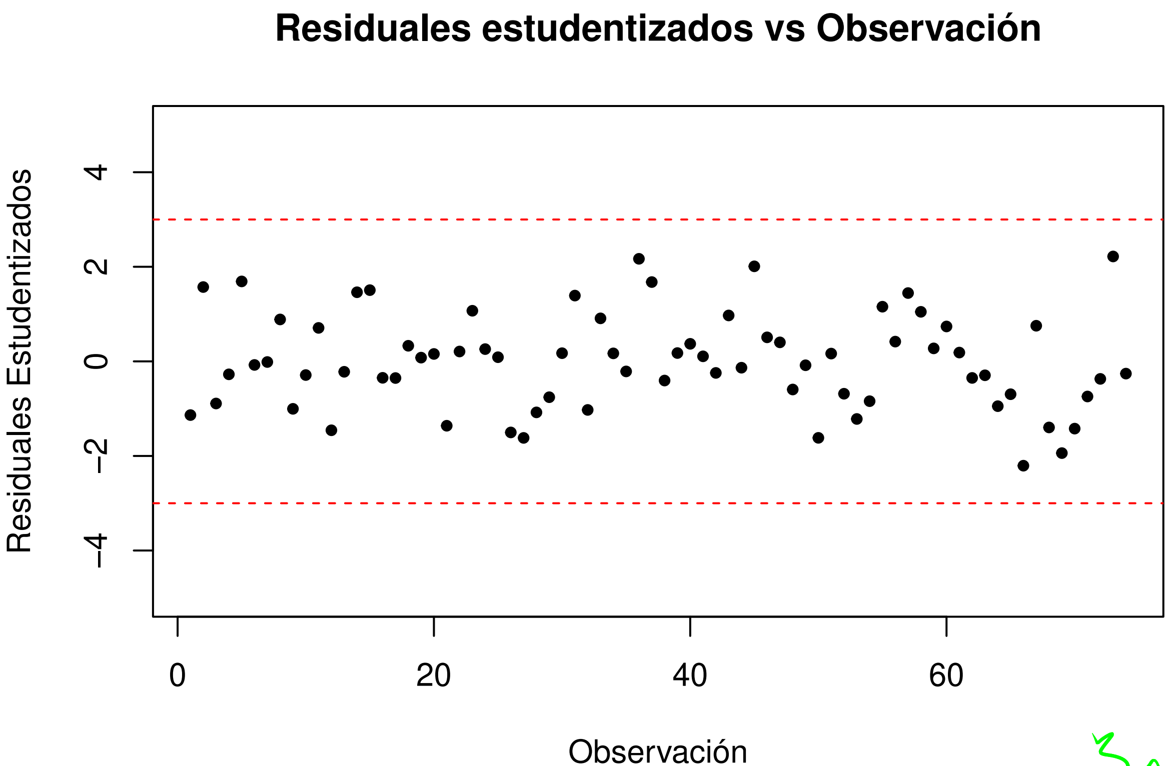



Figura 3: Identificación de datos atípicos

Según lo evidenciado en el gráfico anterior, no se encuentran observaciones atípicas que se separen de manera significativa en términos del riesgo de infección (Valor Y) con respecto al resto de observaciones obtenidas en el conjunto de datos. Esta afirmación se respalda por la falta de residuales estudentizados que superen el umbral de $|r_{estud}| > 3$. Además, se confirma la ausencia de observaciones influyentes respecto al valor Y, debido a que el riesgo de infección no ejerce un impacto sustancial en los coeficientes de regresión ajustada y por lo tanto, podemos concluir que la variable de respuesta no sesga el analisis de datos de manera importante.

### 4.2.2. Puntos de balanceo

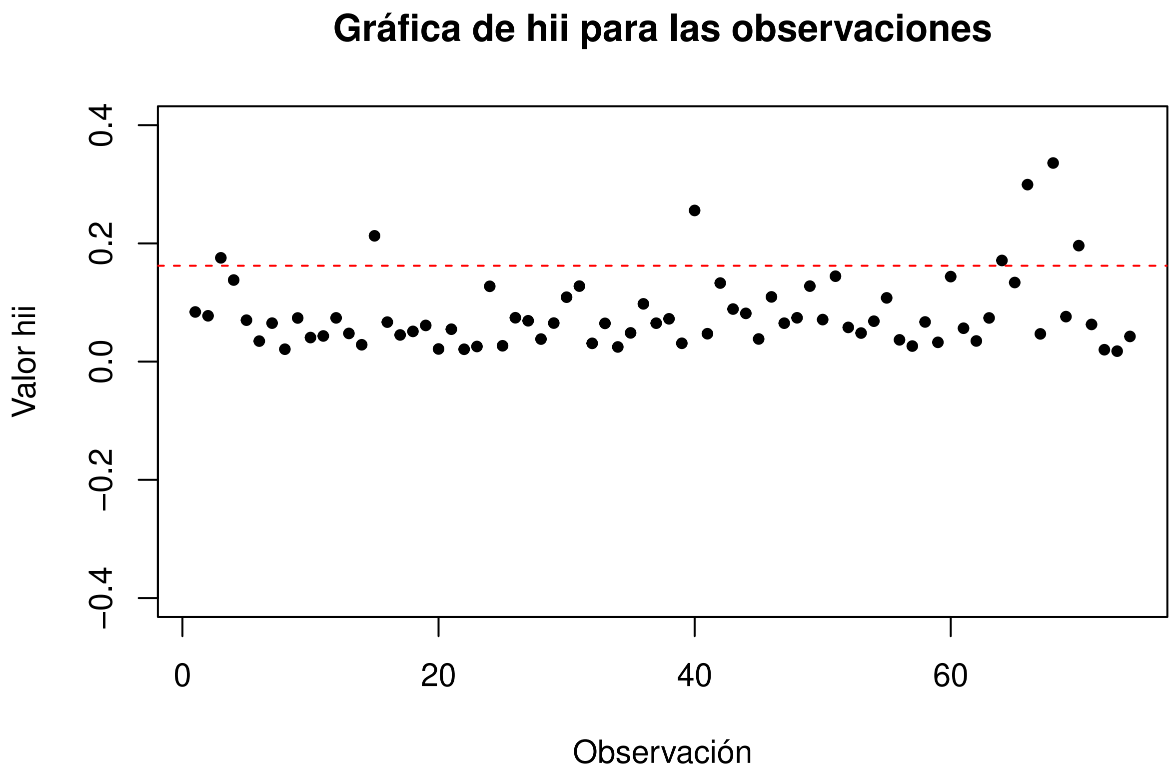



Figura 4: Identificación de puntos de balanceo

Cuadro 5: Tabla de puntos de balanceo

| | Residuales estudentizados | Cook.D | hii | Dffits |
|---|---|---|---|---|
| 3 | -0.8909 | 0.0282 | 0.1756 | -0.4105 |
| 15 | 1.5064 | 0.1022 | 0.2128 | 0.7907 |
| 40 | 0.3712 | 0.0079 | 0.2557 | 0.2162 |
| 64 | -0.9475 | 0.0309 | 0.1711 | -0.4301 |
| 66 | -2.2062 | 0.3467 | 0.2994 | -1.4857 |
| 68 | -1.3972 | 0.1646 | 0.3360 | -1.0009 |
| 70 | -1.4220 | 0.0823 | 0.1962 | -0.7080 |

Al analizar la gráfica de valores $h_{ii}$ vs observaciones, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$ que es igual a 0.1621, se puede notar que, según el criterio $h_{ii} > 2\frac{p}{n}$, hay evidencia de la existencia de 7 puntos de balanceo presentandos en el cuadro

3pt

5. Estos puntos se distinguen porque se encuentran alejados del resto de datos en el espacio de las predictoras (Valores X) y potencialmente puedan causar observaciones influénciales. El impacto de estos puntos de balanceo es notable sobre ciertas propiedades del modelo de regresión ajustada, Aunque su efecto no se refleja de manera directa sobre las estimaciones de los coeficientes, pueden tener impacto marcado sobre las estadisticas de resumen como el $R^2$ y el error estandar de los coeficientes estimados.

#### 4.2.3. Puntos influenciales

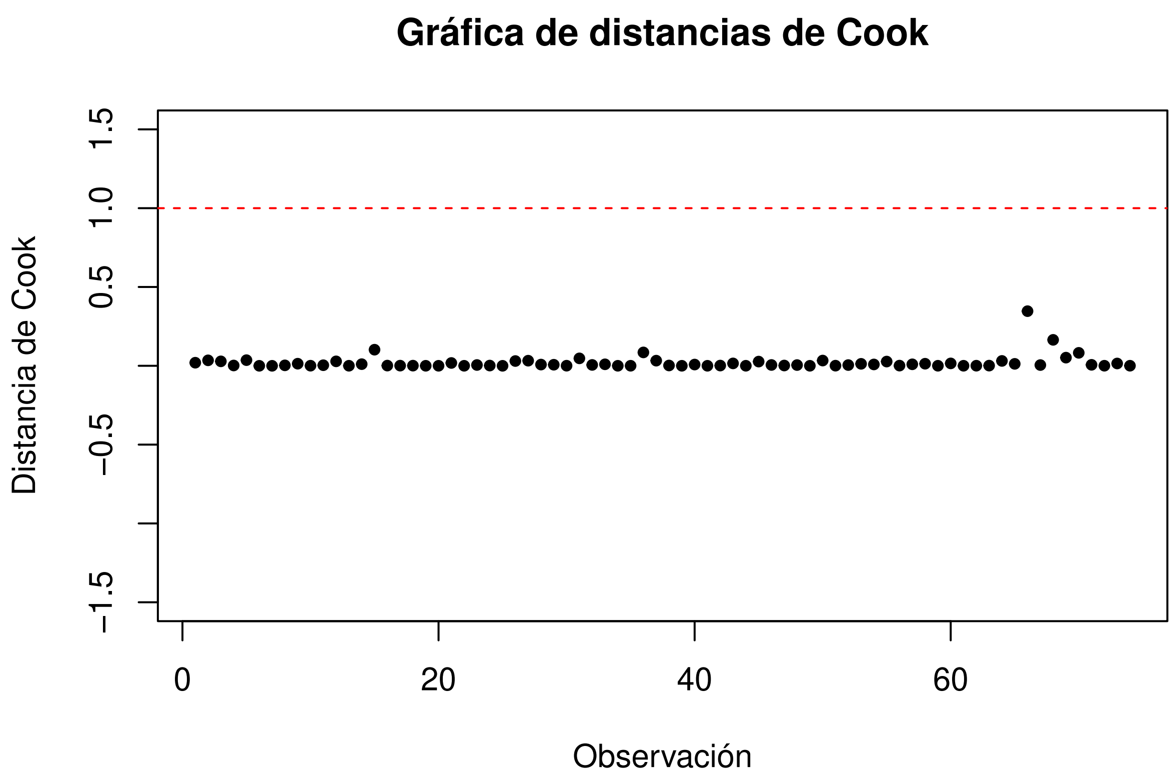



Figura 5: Criterio distancias de Cook para puntos influenciales

De acuerdo con el criterio de Cook, se afirma que la observación $i$ será influencial si $Cook.D_i > 1$. Sin embargo, en la figura anterior, no se identifica ninguna evidencia de puntos influénciales según este criterio.

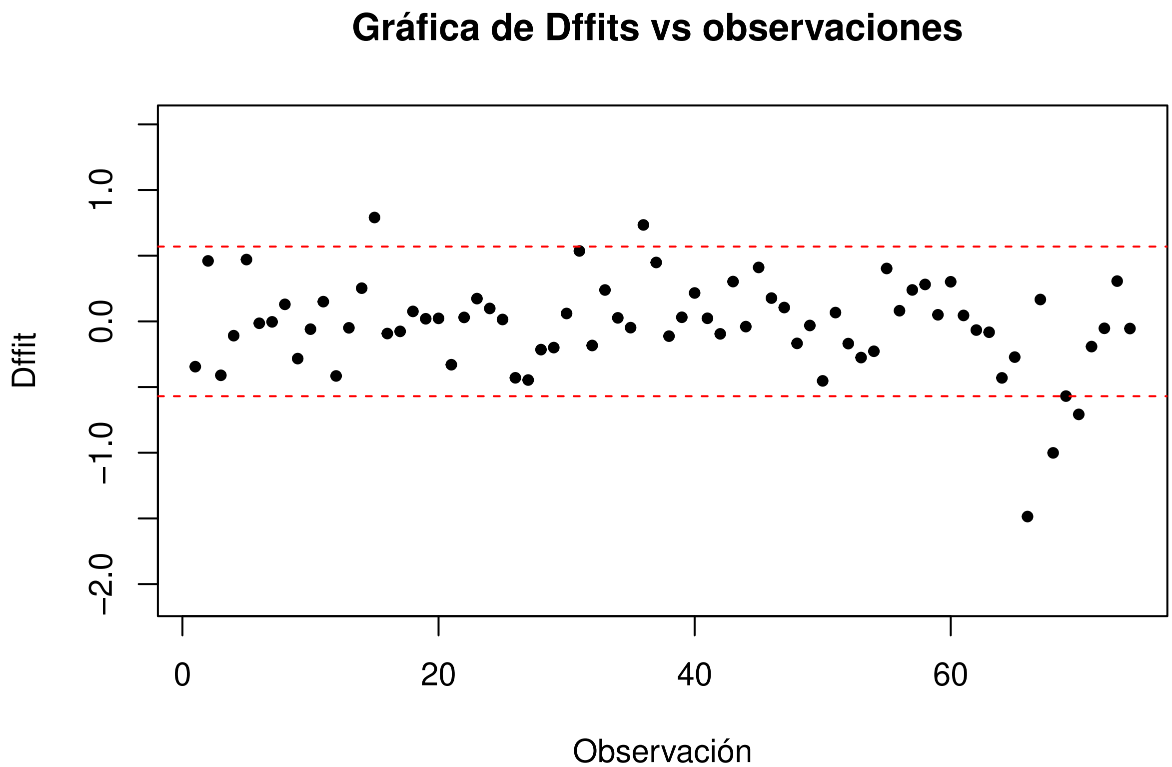



Figura 6: Criterio Dffits para puntos influenciales

Cuadro 6: Tabla de puntos de influénciales

| | Residuales estudentizados | Cook.D | hii | Dffits |
|---|---|---|---|---|
| 15 | 1.5064 | 0.1022 | 0.2128 | 0.7907 |
| 36 | 2.1690 | 0.0849 | 0.0977 | 0.7343 |
| 66 | -2.2062 | 0.3467 | 0.2994 | -1.4857 |
| 68 | -1.3972 | 0.1646 | 0.3360 | -1.0009 |
| 70 | -1.4220 | 0.0823 | 0.1962 | -0.7080 |

Como se puede ver, las observaciones ubicadas en el cuadro 6 cumplen con el criterio de Dffits, el cual establece que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, es un punto influencial. Cabe destacar que las observaciones 15,66,68,70 corresponden tambien a puntos de balanceo, por lo tanto , estos puntos ejercen un impacto considerable sobre los coeficientes de regresión ajustados en este modelo, en otras palabras, halan el modelo ajustado en dirección de estas observaciones influénciales y podrían provocar cambios sustanciales en el ajuste de la regresión lineal si se excluyeran del análisis, es decir, el impacto de estos puntos revela

4 pt

cuanto se desvia el valor de la observación en la variable predictora de la media de la variable predictiva.

## 4.3. Conclusión

3pt

Basándonos en los resultados obtenidos de la validación de los supuestos sobre los residuales estudentizados, podemos afirmar que se cumplen con el criterio de normalidad, una aparante varianza constante, la media de los errores demostrada matematicamente como igual a 0 y el supuesto de independencia que siempre se cumple para este curso. Se Concluye en un modelo RLM valido. Sin embargo, es relevante destacar la existencia de puntos de balanceo e influencia en los datos. Aunque su presencia pueda afectar la normalidad y el ajuste en el modelo de regresión lineal múltiple, no constituirá un factor determinante que lleve a desechar o aprobar por completo la validez del modelo RLM.