4,5

Trabajo 1

Estudiantes

Stefany Cantero Cárdenas
Daniel Alejandro Sánchez Villota
Johan Camilo Amado Sabbagh
Samuel Higuita Pulgarin

Equipo

25

Docente

Mateo Ochoa Medina

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín
5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1 17 pt

Con base en la base de datos brindada, se tiene una muestra de 74 hospitales de EE.UU. en los cuales se realizo un estudio sobre la eficiencia en el control de infecciones hospitalarias a partir de 5 variables regresoras:

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

Donde:

- Y: Riesgo de infección hospitalaria en EE.UU (porcentaje)
- X_1: Duración de la estadía de los pacientes en el hospital (Promedio)
- X_2: Razon cultivos de pacientes sin sintomas por cada 100
- X_3: Número de camas en el Hospital (Promedio)
- X_4: Pacientes por dia en el Hospital (Promedio)
- X_5: Enfermeras FTE en el periodo de estudio (Promedio)

1.1. Modelo de regresión

3 pt

Al ajustar el modelo, se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-0.7643
β_1	0.1840
β_2	0.0116
β_3	0.0481
β_4	0.0201
β_5	0.0014

Despues de estimar los parámetros, se obtiene la ecuación de regresión ajustada:

$$\hat{Y}_i = -0.7643 + 0.184X_{1i} + 0.0116X_{2i} + 0.0481X_{3i} + 0.0201X_{4i} + 0.0014X_{5i}$$

1.2. Significancia de la regresión

β_0 no va acá

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

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2,..., 5} \end{cases}$$

0

¿ y β_0 ?

3pt

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:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	68.7005	5	13.740096	15.1837	5.32469e-10
Error	61.5347	68	0.904921		

De la tabla Anova,se tiene que dado que el valor-p es muy cercano a cero, por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $0 \leqslant j \leqslant 5$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$, por tanto podemos concluir que 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	-0.7643	1.5728	-0.4859	0.6286
β_1	0.1840	0.0921	1.9981	0.0497
β_2	0.0116	0.0277	0.4183	0.6770
β_3	0.0481	0.0128	3.7500	0.0004
β_4	0.0201	0.0069	2.9053	0.0049
β_5	0.0014	0.0006	2.2071	0.0307

6pt

Los valores-p presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_1, β_3, β_4, β_5 son significativos, pues sus valores-p son menores a α, lo que significa que tienen un impacto significativo en el modelo. Por otro lado, el parámetro β_2 no es significativo, ya que su valor-p es mayor que α, lo que sugiere que no contribuye de manera significativa a la respuesta estudiada.

1.4. Interpretación de los parámetros

Una vez realizadas las respectivas pruebas de significancia, se analiza cada covariable y se tiene que:

$\hat{\beta}_1$**:** Por cada unidad de incremento en la duración de estadía de los pacientes en el hospital,el promedio del porcentaje aumenta en 0.1839792 unidades en el riesgo de infección hospitalaria en EE.UU, cuando las otras covariables se mantienen constantes

$\hat{\beta}_3$**:** Por cada unidad adicional en el número de camas en el hospital, el promedio del porcentaje aumenta en 0.0480792 unidades en el riesgo de infección hospitalaria en EE.UU, cuando las otras covariables se mantienen constantes 3 pt

$\hat{\beta}_4$**:** Por cada incremento de día por cada paciente en el Hospital, el promedio del porcentaje aumenta en 0.0201106 unidades en el riesgo de infección hospitalaria en EE.UU, cuando las otras covariables se mantienen constantes

$\hat{\beta}_5$**:** Por cada de incremeto de enfermeras FTE en el periodo de estudio , el promedio del porcentaje aumenta en 0.0013782 unidades en el riesgo de infección hospitalaria en EE.UU, cuando las otras covariables se mantienen constantes

1.5. Coeficiente de determinación múltiple R^2

2 pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5275$, lo que significa que aproximadamente el 52.75 de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

¿cómo se calcula?

2. Pregunta 2

3 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Se desea probar la significancia simultánea del conjunto de tres variables con lo p-valores ~~mayores~~. menores

Despues de estimar los coeficientes de regresión, se observa que las variables predictoras con los p valores ~~mayores~~ corresponden a X_1, X_2, X_5,
menores

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 2, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	61.535	X1 X2 X3 X4 X5
Modelo reducido	74.974	X3 X4

En consecuencia con la prueba de hipotesis basada en los 3 p-valores mayores, un modelo reducido para laprueba de significancia de subconjunto es:

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

1pt

2.2. Estadístico de prueba y conclusión

establecemos el estadístico de prueba como:

Al menos son consecuentes con el error

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_4) - SSE(\beta_0,\ \cdots,\ \beta_5))/3}{MSE(\beta_0,\ \cdots,\ \beta_4)} \overset{H_0}{\sim} f_{3,68} \\ &= \frac{74.974 - 61.535/3}{0.904921} \\ &= \frac{4.479667}{0.904921} \\ &= 4.95034005 \end{aligned} \tag{2}$$

2pt

Para el criterio de decisión se requiere obtener el valor crítico de una distribución $f_{3,68} = 2.7395$ a un nivel de significancia $\alpha = 0.05$, esto es, $f_{0.05,3,68}$=2.739502.

Como $F_0 = 4.95034005 > f_{0.05,3,68} = 2.739502$, entonces se rechaza H_0 y se concluye que la probabilidad promedio de adquirir infección en el hospital de EE.UU. depende la duracion de estadia, las rutinas de cultivos y/o el número de enfermeras.

Por lo tanto, no es posible descartar del modelo las variables del subconjunto examinado.

3. Pregunta 3

5pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Haciendo la observación del problema y de las variables contenidas en él podemos realizarnos algunas preguntas, como por ejemplo: ¿El efecto de la Duración de la estadía sobre el Riesgo de infección es igual al Censo promedio diario sobre el Riesgo de infección? ó

¿El Número de camas sobre el Riesgo de infección es igual a 5 veces el Número de enfermeras sobre el Riesgo de infección? Usando estas preguntas podemos plantear la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_4;\ \beta_3 = 5\beta_5 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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 **L** dada por

$$L = \begin{bmatrix} 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -5 \end{bmatrix}$$

1pt

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 $X_{4i}^* = X_{1i} + X_{4i}$ y $X_{5i}^* = 5X_{3i} + X_{5i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

2pt

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,68} \tag{3}$$

Ahora, reemplazamos los valores que conocemos en la ecuación. De la tabla ANOVA de la pregunta número 1 conocemos el SSE(MF) y el MSE(MF), entonces:

$$F_0 = \frac{(SSE(MR) - 61.5347)/2}{0.904921} \overset{H_0}{\sim} f_{2,68} \tag{4}$$

4. Pregunta 4

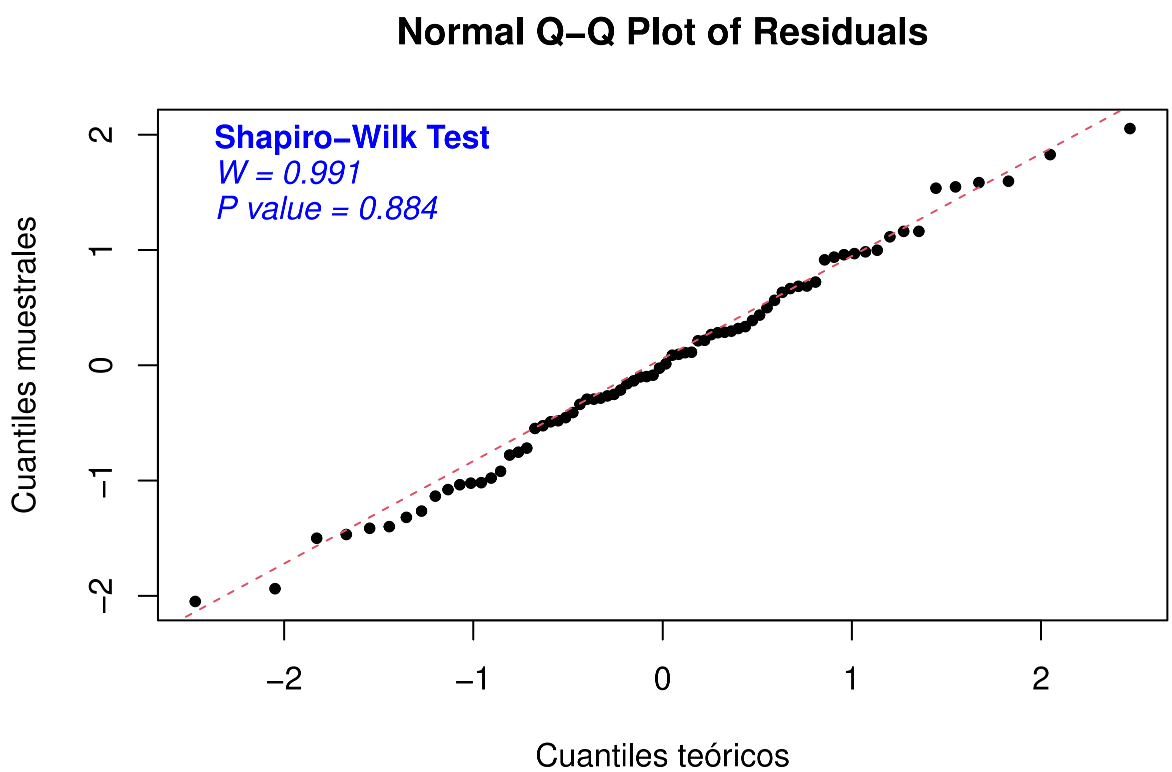
20pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Con el fin de validar el supuesto de normalidad del modelo, y teniendo en cuenta que en el curso asumimos la independencia de los errores, se planteará una prueba de hipótesis con nivel de significancia $\alpha = 0.05$ que se realizará por medio de la prueba de normalidad Shapiro-Wilk y se tendrá en cuenta un análisis gráfico de probabilidad normal de los residuales cuantil-cuantil:

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



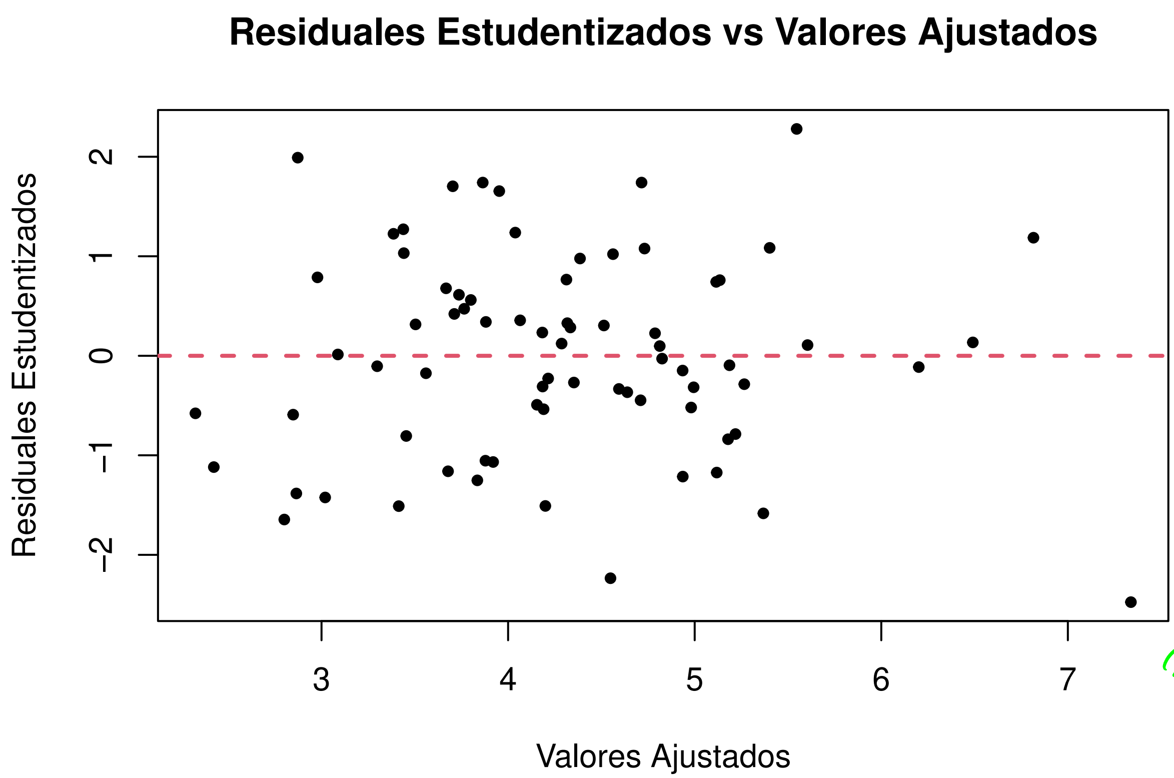
4 Pt

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

El resultado de la prueba de normalidad arroja un valor p de aproximadamente 0.884, el cual supera considerablemente el nivel de significancia establecido para la prueba. Por lo tanto, no se rechazaría la hipótesis nula, lo que implicaría que el modelo cumple con el supuesto de normalidad, es decir, que los datos siguen una distribución normal con una media μ y varianza σ^2. De igual manera, al analizar la gráfica de comparación de cuantiles, podemos apreciar que la mayor cantidad de datos se distribuyen a lo largo de la línea recta. Así que, en consideración del análisis gráfico y el resultado del análisis teórico, decidimos no rechazar la hipótesis nula, concluyendo que el modelo cumple con el supuesto de normalidad de los residuales.

4.1.2. Varianza constante

Para validar la varianza constante de los datos, se realizará un gráfico de Residuales estudentizados vs Valores ajustados:



3pt

Figura 2: Gráfico residuales estudentizados vs valores ajustados

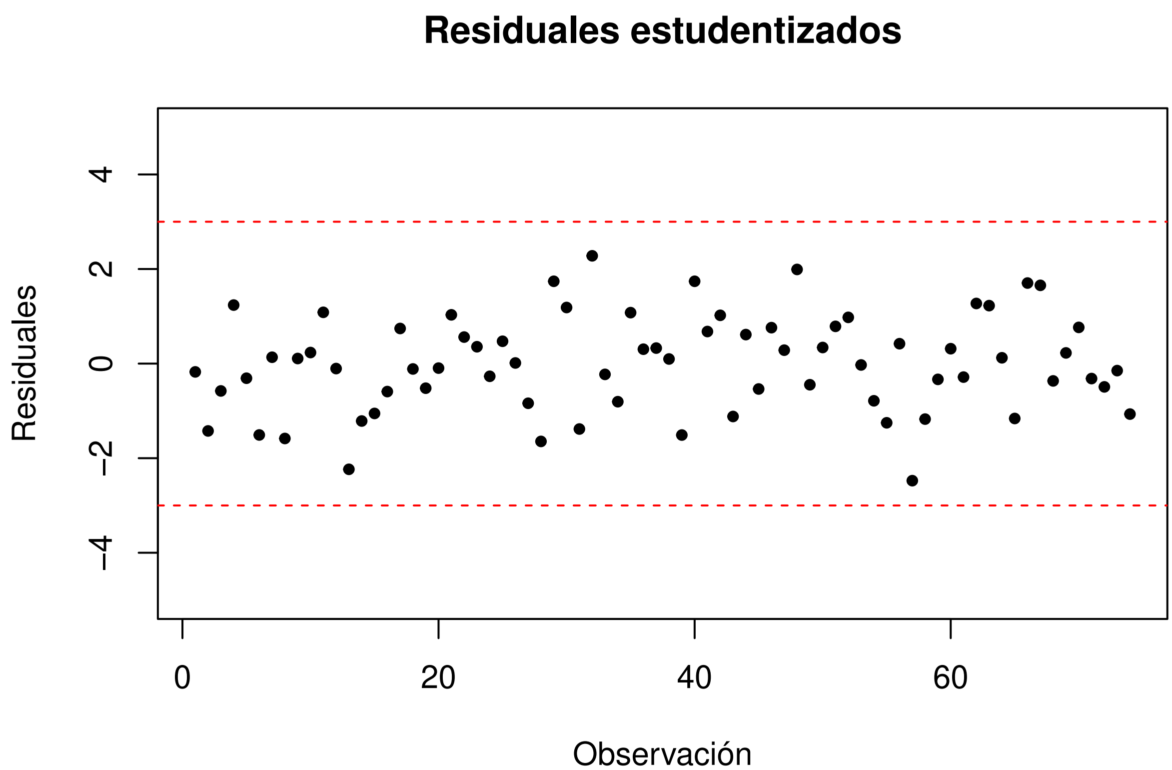
En el gráfico resultante podemos observar que los datos están dispersos de manera aleatoria alrededor de la línea recta y sin patrones evidentes. Además, notamos que algunas observaciones tienen una media igual a cero o aproximadamente cero. Este comportamiento sugiere que la variabilidad de los residuales es constante en todos los niveles de los valores ajustados y nos permite afirmar que el supuesto de varianza constante se cumple para el modelo.

4.2. Verificación de las observaciones

Una vez realizadas las validaciones de los supuestos en el modelo de regresión lineal múltiple, se realizará la verificación de observaciones atípicas, puntos de balanceo y observaciones influenciales.

4.2.1. Datos atípicos

Para la identificación de datos atípicos, se realizará un análisis gráfico de Residuales estudentizados vs Observaciones:



3 pt

Figura 3: Identificación de datos atípicos

En la gráfica resultante, podemos observar que no se identifican observaciones atípicas en la muestra basados en el criterio de $|r_{estud}| > 3$. Podríamos considerar esto como algo positivo, ya que sería un indicativo de que no hubo errores de registro de la información a la hora de tomar la muestra y que podemos confiar en las inferencias realizadas anteriormente. Sin embargo, es importante destacar que esta ausencia no garantiza por completo que el modelo sea perfecto o que no existan otros tipos de inconvenientes con las mediciones.

4.2.2. Puntos de balanceo

Se continuará esta verificación con una identificación de puntos de balanceo. Para este análisis se hará uso de un gráfico Observaciones vs Valores h_{ii}:

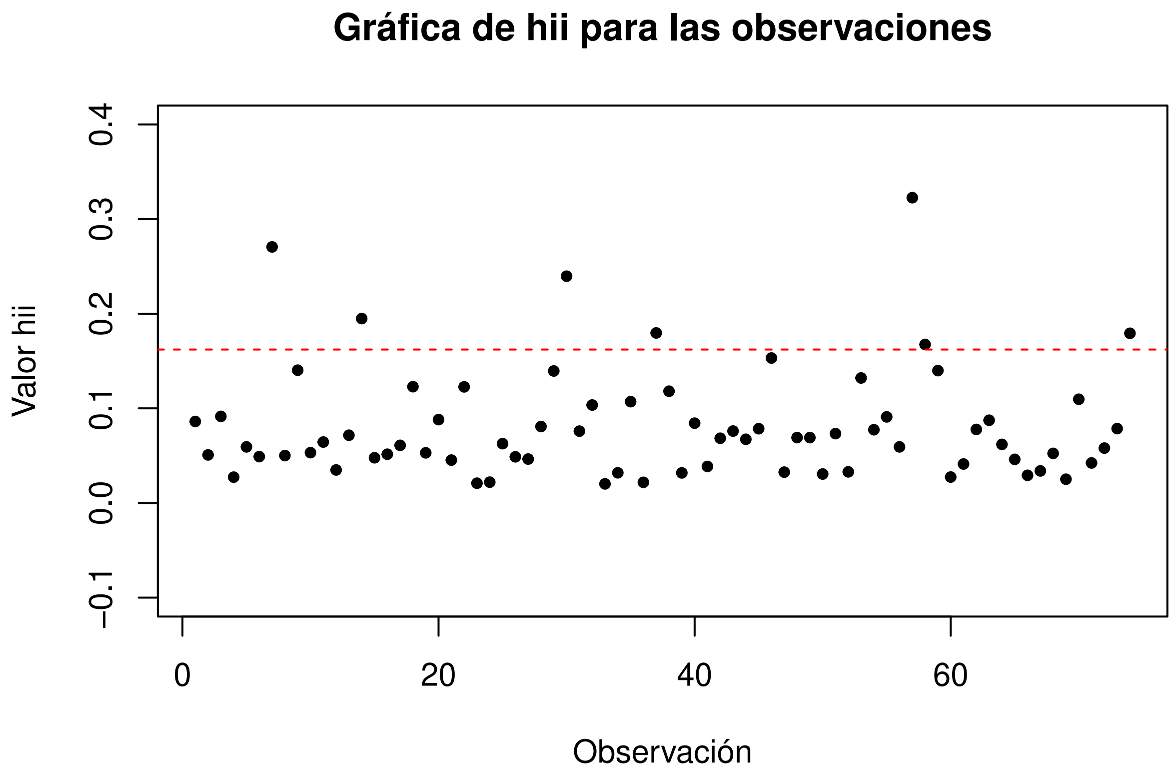


Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 7    0.1345  0.0011    0.2706  0.0813
## 14  -1.2139  0.0594    0.1949 -0.5993
## 30   1.1865  0.0739    0.2396  0.6680
## 37   0.3280  0.0039    0.1797  0.1525
## 57  -2.4754  0.4863    0.3226 -1.7776
## 58  -1.1733  0.0462    0.1675 -0.5278
## 74  -1.0673  0.0415    0.1793 -0.4994
```

3pt

En la gráfica presentada, la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$ y se puede apreciar que existen siete 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 presencia de estos siete puntos de balanceo puede estar relacionada con el resultado del análisis de la gráfica cuantil-cuantil en donde podemos evidenciar que algunas observaciones en las colas se encuentran un poco dispersas. Esto se debe a que los puntos de balanceo, al ser observaciones alejadas de la mayoría de la muestra, podrían contribuir a una falta de normalidad en los residuales. Sin embargo, en el caso de nuestro modelo, estos puntos no tienen la predominancia suficiente para afectar el supuesto de normalidad.

4.2.3. Puntos influenciales

Para finalizar la verificación de observaciones, evaluaremos los posibles puntos influenciales en nuestro modelo de regresión lineal múltiple basándonos en los criterios de distancias de Cook y diagnóstico DFFITS. A continuación se presentan las respectivas gráficas:

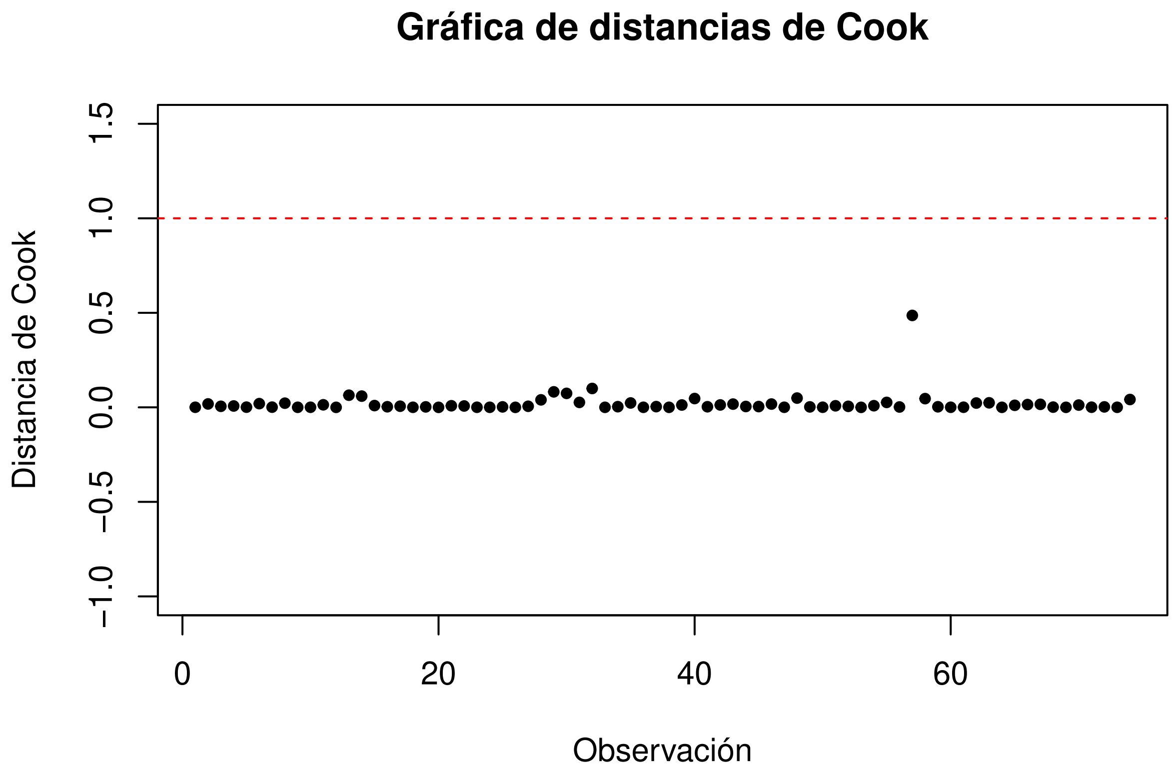


Figura 5: Criterio distancias de Cook para puntos influenciales

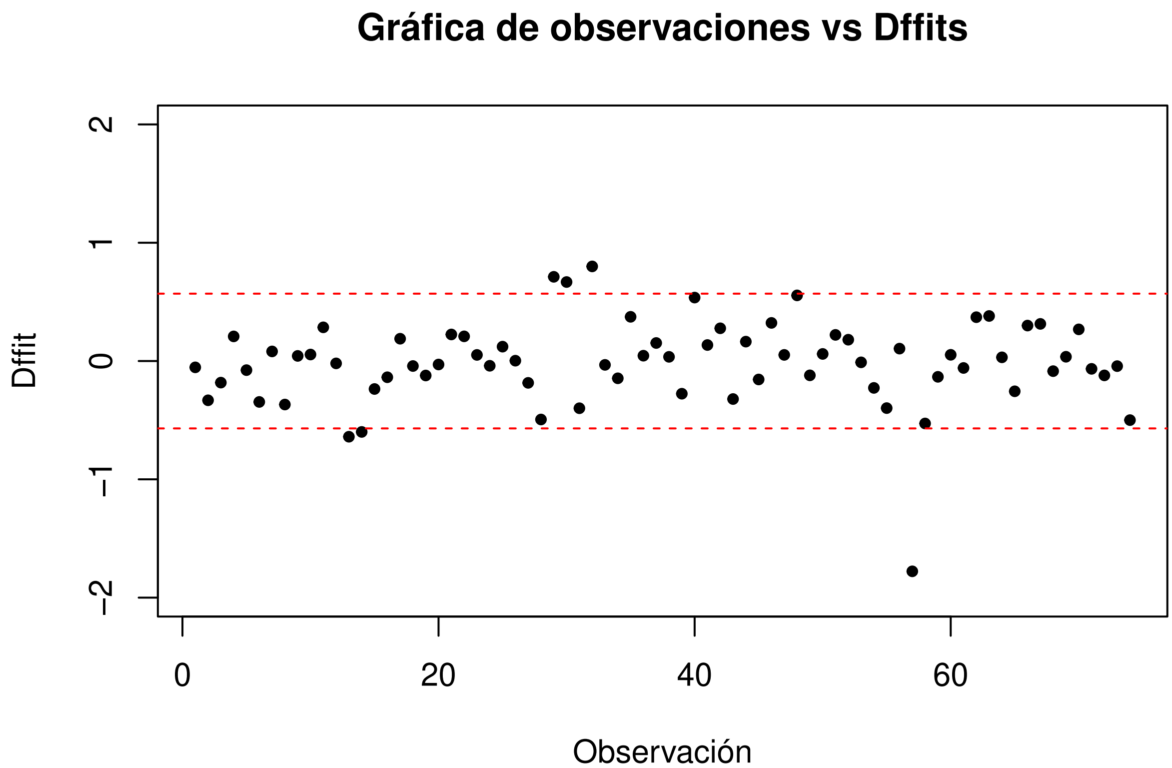


Figura 6: Criterio Dffits para puntos influenciales

```
##     res.stud Cooks.D hii.value  Dffits
## 13  -2.2352  0.0642    0.0716 -0.6399
## 14  -1.2139  0.0594    0.1949 -0.5993
## 29   1.7406  0.0819    0.1395  0.7118
## 30   1.1865  0.0739    0.2396  0.6680
## 32   2.2796  0.0999    0.1035  0.7998
## 57  -2.4754  0.4863    0.3226 -1.7776
```

4pt

Como podemos observar en la gráfica de distancias de Cook, no se identificaron datos que superaran el valor establecido por el criterio $D_i > 1$, lo que indica que no se encontraron observaciones influyentes sobre el vector de parámetros estimados $\hat{\beta}$.

Por otro lado, en la gráfica del diagnóstico DFFITS, se observaron seis puntos que superaron el criterio de influencia $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$ y dos de ellos son a su vez puntos de balanceo. Así que, estos se pueden considerar observaciones influenciales los cuales tienen la capacidad para modificar significativamente los valores ajustados del modelo y, por lo tanto, afectar la ecuación de regresión ajustada.

4.3. Conclusión

3pt

Una vez realizadas las respectivas validaciones de supuestos y la verificación de las observaciones, se determina que el modelo es válido debido a que se cumple el supuesto de normalidad de los errores y de varianza constante. Igualmente, se identificaron puntos de balanceo en las observaciones 7, 14, 30, 37, 57, 58 y 74, que aunque podrían afectar la normalidad, no logran tener un efecto negativo en esta. Además, se encontraron puntos de influencia en las observaciones 13, 14, 29, 30, 32 y 57, lo que sugiere que el modelo puede ser sensible a ciertas observaciones, pero que no se ve afectada la validez de las inferencias y estimaciones basadas en el mismo.