Trabajo 1

3,7

Estudiantes

Juan Felipe Aldana Mejia
Maria Fernanda Jaramillo Gomez
Juan Jose Martinez Grisales
Monica Sofia Restrepo Leon

Equipo 39

Docente

Carlos Mario Lopera

Asignatura

Estadística II

Sede Medellín
5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

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

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

Donde:

- **Y**: Riesgo de infección
- **X1**: Duración de la estadía
- **X2**: Rutina de cultivos
- **X3**: Número de camas
- **X4**: Censo promedio diario
- **X5**: 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
β_0	-0.6914
β_1	0.1427
β_2	0.0292
β_3	0.0670
β_4	0.0086
β_5	0.0016

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.6914 + 0.1427 X_{i1} + 0.0292 X_{i2} + 0.067 X_{i3} + 0.0086 X_{i4} + 0.0016 X_{i5};\ 1 \leqslant i \leqslant 64$$

Una vez planteada la ecuación de regresión ajustada, es esencial evaluar la significancia de cada variable predictora. Esto nos permitirá determinar si estas variables tienen un impacto significativo en la probabilidad promedio estimada de adquirir una infección en el hospital. La importancia de cada parámetro se basa en el valor p, que se encuentra en la tabla de parámetros estimados. Luego, comparamos estos valores p con un nivel de significancia conocido como α (alfa). En este caso, hemos establecido un nivel de significanci a $\alpha = 0.05$. Si el valor p de una variable es menor que α, consideramos que esa variable es significativa en el modelo.

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:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,58} \tag{1}$$

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	70.0311	5	14.006227	15.7774	8.69241e-10
Error	51.4889	58	0.887739		

De la tabla Anova, se observa un valor P aproximadamente muy cercano a 0 (8.69241e-10), y es menor al nivel de significancia $\alpha = 0.05$ 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. Esto se puede interpretar como que existe al menos una variable predictora que influye significativamente en la probabilidad promedio estimada de adquirir una infección.

Otra manera de llegar a este resultado es con el estadístico de prueba $F_0 = 15.7774 > f_{0.95,5,58} = 2.3738$

Se llega a la misma conclusión y se rechaza la hipótesis nula.

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	-0.6914	1.4295	-0.4837	0.6304
β_1	0.1427	0.0783	1.8223	0.0736
β_2	0.0292	0.0278	1.0521	0.2971
β_3	0.0670	0.0150	4.4534	0.0000
β_4	0.0086	0.0071	1.2086	0.2317
β_5	0.0016	0.0007	2.3146	0.0242

6 pt

Para determinar la significancia individual de los parámetros se plantean las siguientes hipótesis:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_a : \beta_j \neq 0 \end{cases}$$

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_3 y β_5 son significativos individualmente mientras los demás se mantienen constantes pues sus P-valores son menores a α.

Por otro lado los valores β_0, β_1, β_2 y β_4 no resultan significativos individualmente en presencia de los demás parámetros

Ahora que se han definido los parámetros significativos individuales del modelo de regresión múltiple lo siguiente será realizar su respectiva interpretación

1.4. Interpretación de los parámetros

3 pt

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

$\hat{\beta}_3$: a medida que el número promedio de camas en el hospital (X3) aumenta en una unidad, la probabilidad promedio estimada de adquirir una infección aumenta en 0.067015624 porciento cuando las demás predictoras se mantienen fijas

$\hat{\beta}_5$: a medida que el numero promedio de enfermeras (X5) aumenta en una unidad, La probabilidad promedio estimada de adquirir una infección aumenta en 0.01572896 porciento cuando las demás variables predictoras se mantienen fijas

1.5. Coeficiente de determinación múltiple R^2

1 pt

El siguiente paso es calcular el coeficiente de variación para hacerse una idea de que tan alejados están los puntos Yi de la ecuación de regresión, además de eso veremos si el uso de todas las variables predictoras propuestas para este modelo es realmente necesario para tener un valor alto de R^2 usando una medida de bondad de ajuste

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

Estos valores serán tomados de la tabla Anova por lo que reemplazando se tiene que:

$$R^2 = \frac{70.0311}{(70.0311 + 51.4889)} = 0.5763117614 \tag{3}$$

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.57631$, lo que significa que aproximadamente el 0.57631 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

A continuación, se presenta el cálculo del R^2_{adj}:

$$R^2_{\text{adj}} = 1 - \frac{(n-1)MSE}{SST} \tag{4}$$

Y recordemos que MSE= SSE/ (g.l SSE) reemplazando se obtiene que:

$$R^2_{\text{adj}} = 1 - \frac{(64-1)(51.4889/58)}{121.52} = 0.539766475 \tag{5}$$

$R^2_{\text{adj}} = 0.539766475 < R^2 = 0.5763117614$ esto puede ser explicado como que en el modelo de regresión existen variables que no aportan significativamente o son variables redundantes por lo tanto se pueden depurar más adelante

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más alto en el modelo fueron X_1, X_2, X_4. Para poder verificar la significancia del subconjunto (Duracion de la estadia(X1),Numero de camas promedio (X3) Y numero promedio de enfermeras (X5) en este caso) se requiere realizar la tabla de todas las combinaciones de regresiones posibles “myallregtable” enfocándonos en el conjunto mencionado anteriormente bajo la siguiente prueba de hipótesis:

$$\begin{cases} \text{H}_0 : \beta_1 = \beta_2 = \beta_4 = 0 \\ \text{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 2, 4 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	51.489	X1 X2 X3 X4 X5
Modelo reducido	69.852	X2 X3

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

$$Y_i = \beta_0 + \beta_3 X_{i3} + \beta_5 X_{i5} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,58} \\ &= \frac{[(83.242 - 51.4889)/3]}{0.887739} \\ &= 11.92283618 \end{aligned} \tag{6}$$

Ahora, comparando el F_0 con $f_{0.95,3,58} = 2.7636$, se puede ver que $F_0 > f_{0.95,1,45}$ y por tanto qué se rechaza H_0 a favor de H_a, concluyendo así que la probabilidad promedio estimada de adquirir una infección en el hospital depende significativamente por lo menos de una de las variables asociadas al subconjunto (Duracion de la estadia(X1),Numero de camas promedio (X3) Y numero promedio de enfermeras (X5))

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si ¿La influencia de las variables duración promedio de la estadía (X1), la rutina de cultivos(X2) y numero de camas promedios durante el estudio (X3) es similar respecto al censo promedio diario (X4)? por consiguiente se plantea la siguiente prueba de hipótesis:

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

reescribiendo matricialmente:

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

Con **L** dada por

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

3 pt

El modelo reducido está dado por:

$$Y_i = \beta 0 + \beta 4(X1 + X2 + X3 + X4) + \beta 5 X5$$

Hay que fijarse en que la matriz L tiene 3 filas que son Linealmente independientes (no se puede escribir ninguna fila como un múltiplo escalar de las otras) por ende r = 3.Donde $X_{1,2,3,4} = X1 + X2 + X3 + X4$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

2 pt

$$F_0 = \frac{(SSE(RM) - SSE(FM))/3}{MSE(FM)} \overset{H_0}{\sim} f_{0.95,3,58} \tag{7}$$

$$F_0 = \frac{([SSE(RM) - 51.4889]/3)}{0.887739} \tag{8}$$

4. Pregunta 4

1 pt

Por ultimo se requiere verificar los supuestos de los errores (independencia, varianza constante,normalidad y media igual 0, cabe recalcar que por la manera como se tomaron los datos vamos a asumir el supuesto de independencia y el supuesto de media 0 siempre se cumple para los errores del modelo) para eso utilizaremos la prueba de Shapiro-Wilk donde se tomaran el valor p para determinar si los errores se distribuyen normales y luego se realizara el análisis grafico para determinar si a primera vista se pueden detectar puntos influénciales que afecten el supuesto de normalidad, es decir los errores se distribuirán normales si no se detectan dichos puntos y el valor p es mayor a 0.05

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:

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

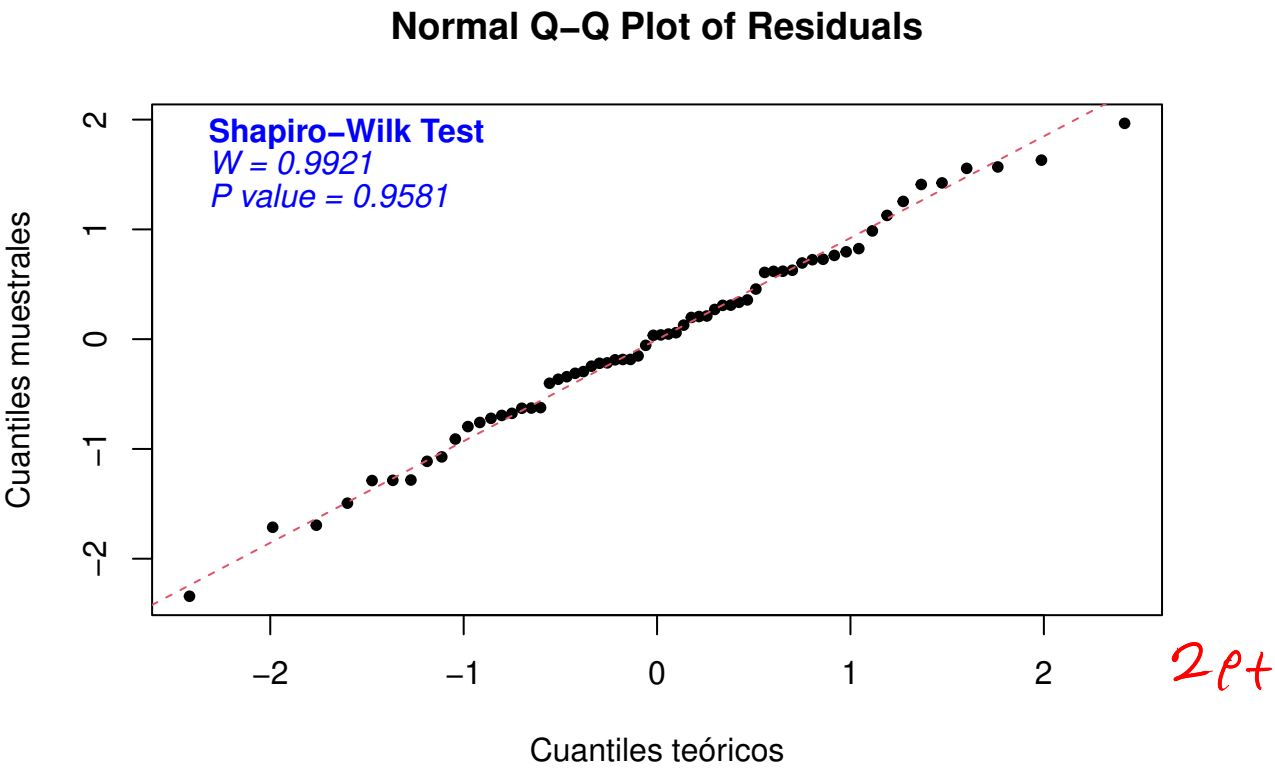

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Bajo H_0: errores distribuidos normales el valor p es cercano a 1 en particular tenemos el valor p de 0.9581 de esta prueba de hipótesis que es mayor a 0,05 por ende no se tendría la suficiente evidencia muestral para rechazar H_0 , por lo tanto se acepta el supuesto de normalidad teniendo en cuenta que considerando únicamente este análisis gráfico no se pueden percibir puntos que influyan significativamente para rechazar el supuesto de normalidad sin embargo debemos fijarnos también en la posible existencia de valores influénciales por medio de las pruebas numéricas

4.1.2. Varianza constante

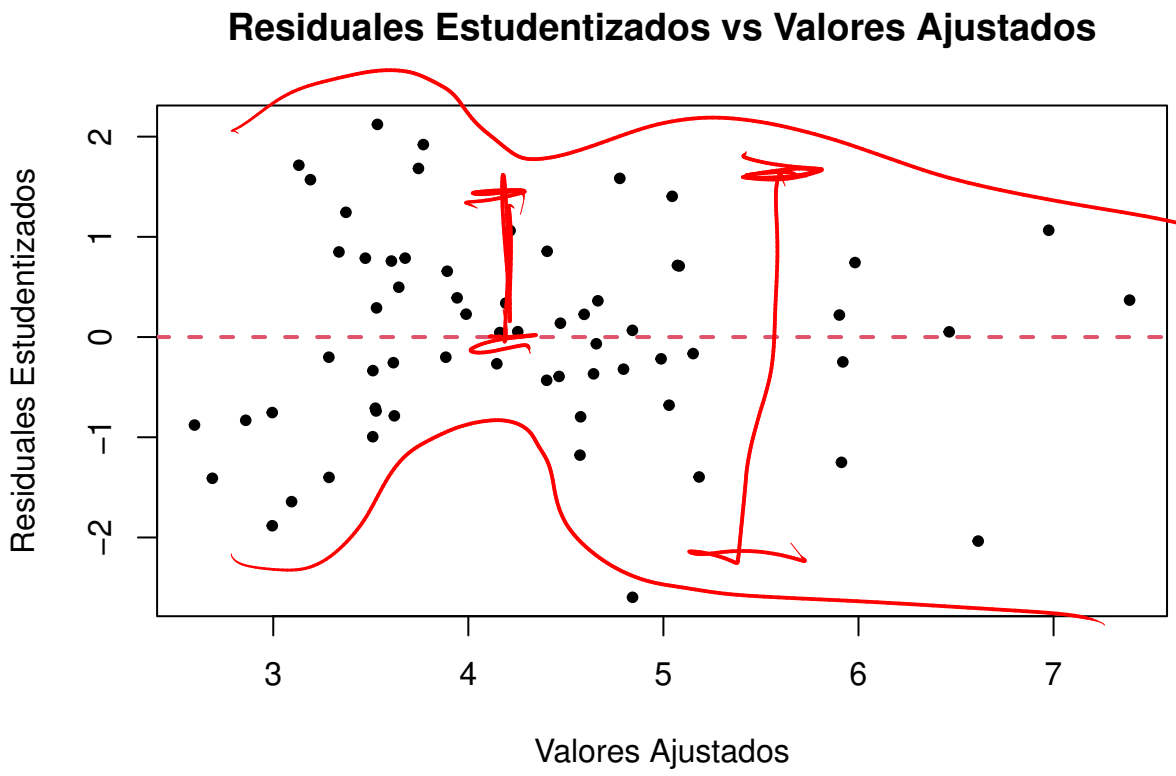

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En este grafico si se dibuja una línea por encima y por debajo de los puntos podemos decir que la concentración de los datos no es concreta (los puntos están muy dispersos entre sí) y esto se puede interpretar como que el supuesto de la varianza constante no se cumple, sin embargo, no se puede determinar si es porque la varianza del modelo en efecto no es constante o existen observaciones atípicas que provocan que este supuesto no se cumpla

4.2. Verificación de las observaciones

Tengan cuidado acá, modifiquen los límites de las gráficas para que tenga sentido con lo que observan en la tabla diagnóstica. También, consideren que en aquellos puntos extremos que identifiquen deben explicar el qué causan los mismos en el modelo.

4.2.1. Datos atípicos

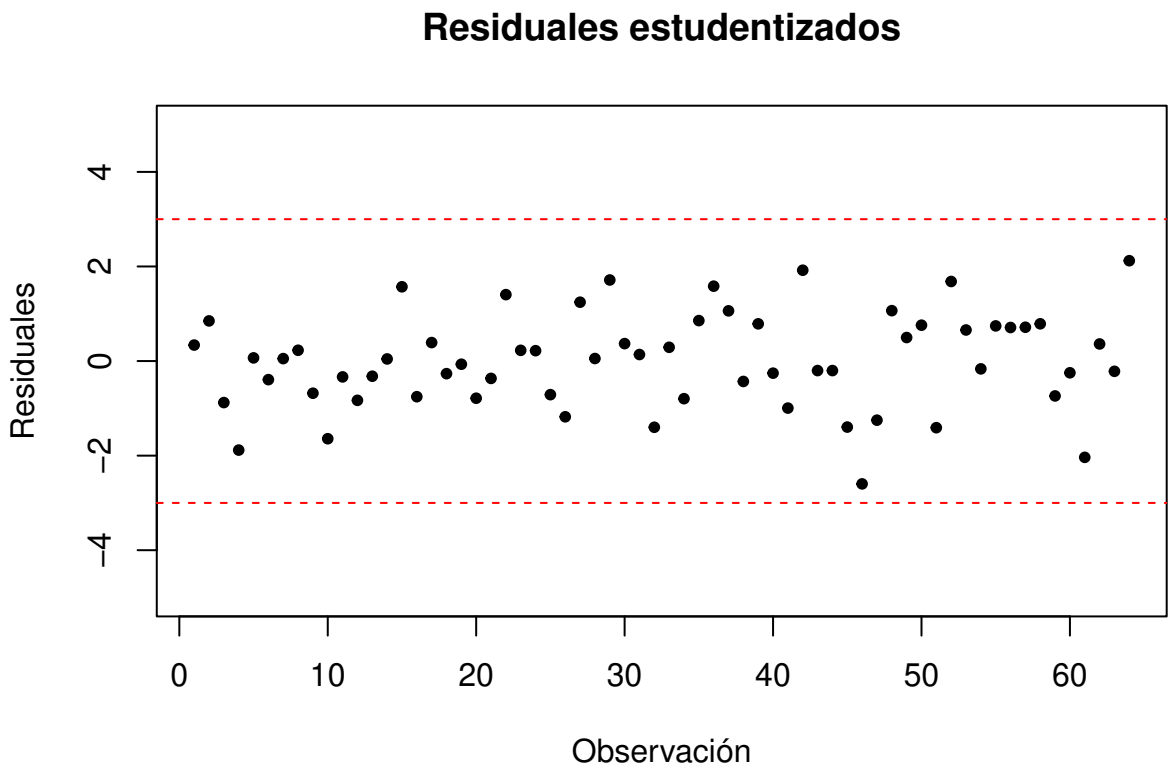

Figura 3: Identificación de datos atípicos

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 $|r_{estud}| > 3$.

4.2.2. Puntos de balanceo

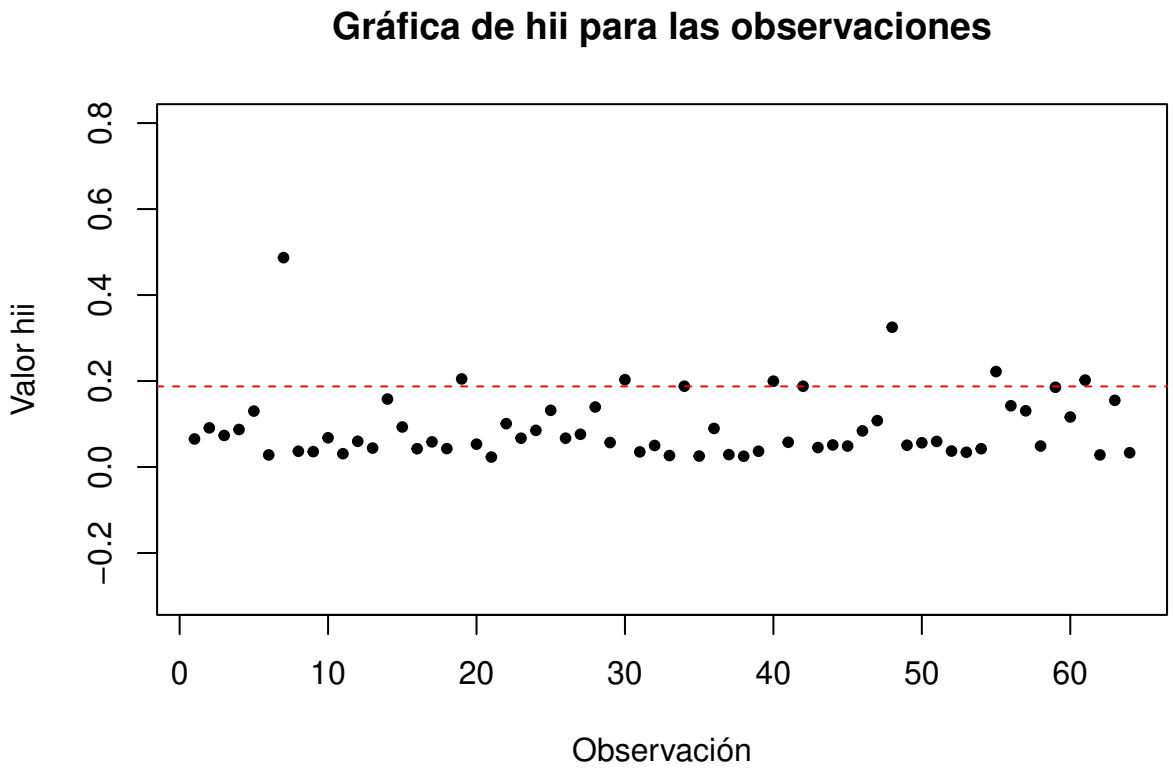

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 7    0.0521  0.0004    0.4870  0.0503
## 19  -0.0669  0.0002    0.2049 -0.0337
## 30   0.3687  0.0058    0.2031  0.1847
## 34  -0.7958  0.0244    0.1880 -0.3817
## 40  -0.2560  0.0027    0.1997 -0.1268
## 42   1.9208  0.1424    0.1880  0.9469
## 48   1.0657  0.0912    0.3251  0.7405
## 55   0.7434  0.0263    0.2220  0.3955
## 61  -2.0361  0.1751    0.2022 -1.0546
```

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 10 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.

4.2.3. Puntos influenciales

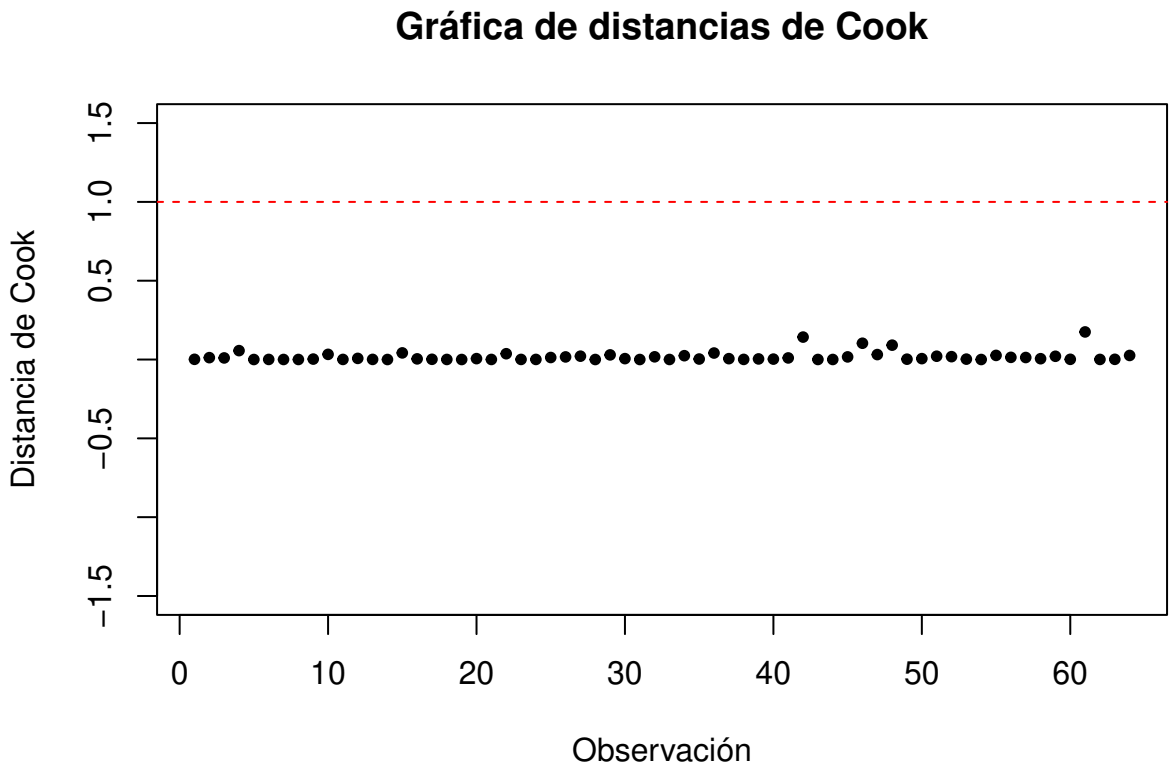

Figura 5: Criterio distancias de Cook para puntos influenciales

Como ningún valor supera el valor de 1 podemos decir que no existen puntos influenciales por parte del criterio de Cook, ahora solo falta probar si existen puntos influenciales por el criterio dffits

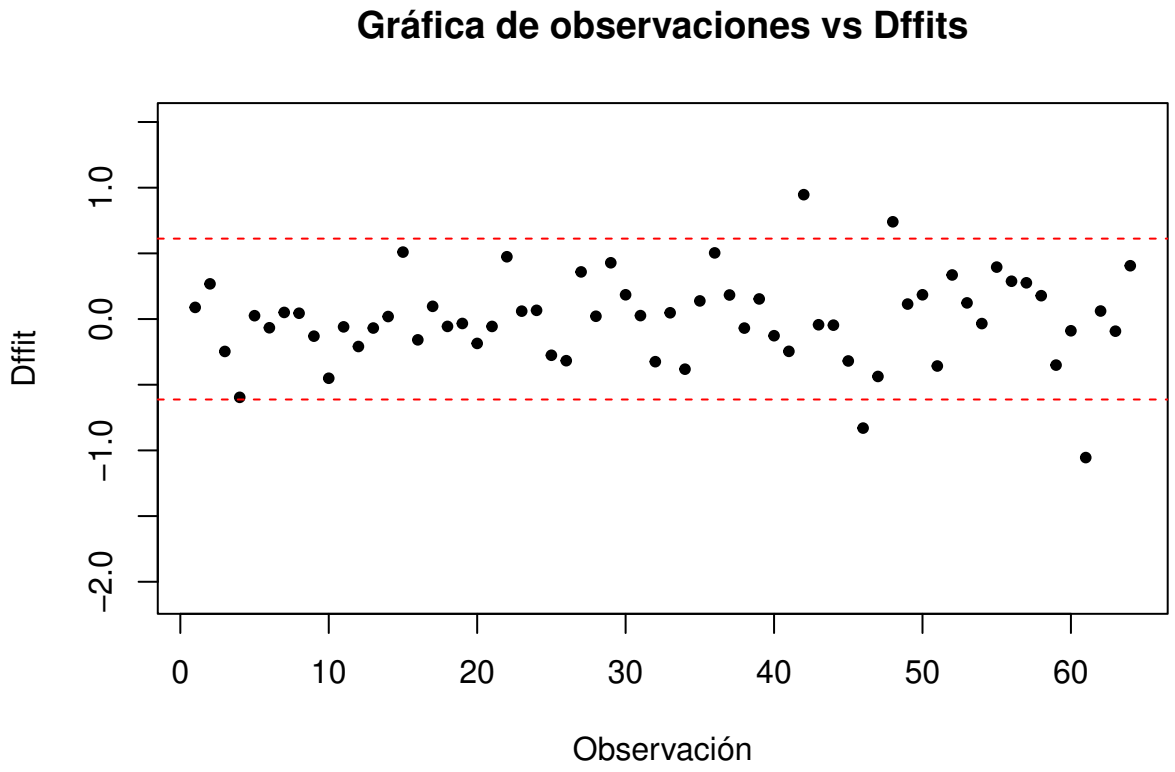

Figura 6: Criterio Dffits para puntos influenciales

```
##     res.stud Cooks.D hii.value  Dffits
## 42   1.9208  0.1424    0.1880  0.9469
## 46  -2.5972  0.1033    0.0841 -0.8300
## 48   1.0657  0.0912    0.3251  0.7405
## 61  -2.0361  0.1751    0.2022 -1.0546
```

Como se puede ver,las observaciones 42, 46, 48 y 61 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, es un punto influencial.

4.2.4. Conclusiones de las observaciones extremas

Por lo tanto, se pueden realizar las siguientes conclusiones:

- No se presentan observaciones atípicas en el modelo
- 7, 19, 30, 34, 40, 42, 48, 55, 61 son observaciones que son punto de balanceo
- 42, 46, 48, 61 se consideran observaciones influenciales

4.3. Conclusión

Como conclusión se tiene que, en definitiva, el modelo no cumple con todos los supuestos necesarios en sus errores para permitirse realizar intervalos de confianza, y predicciones de

valores futuros. Luego se tiene que con un modelo un poco mas sencillo (modelo de regresión con únicamente variables no redundantes) justificandolo con que la medida de R^2_{adj} difiere del R^2. (no se tienen las herramientas necesarias para corregir estas medidas), la presencia de observaciones extremas puede influir en el cumplimiento de los supuestos (positiva o negativamente) por ende se sugiere elegir una metodología más puntual para determinar la validez de los supuestos (una solución podría ser una regresión robusta para confirmar a ciencia cierta el comportamiento del modelo de regresión lineal propuesto)