Trabajo 1

4,4

Estudiantes

Sara Castaño Montoya
Katrin Talana Copete Arroyo
Samuel Rodriguez Ruiz
Eimy Tatiana Sanabria Moreno

Equipo 60

Docente

Javier Armando Lozano Rodriguez

Asignatura

Estadística II

Sede Medellín
5 de octubre de 2023

Índice

1. Pregunta 1 **3**
- 1.1. Modelo de regresión . . . 3
- 1.2. Significancia de la regresión . . . 3
- 1.3. Significancia de los parámetros . . . 4
- 1.4. Interpretación de los parámetros . . . 4
- 1.5. Coeficiente de determinación múltiple R^2 . . . 5

2. Pregunta 2 **5**
- 2.1. Planteamiento pruebas de hipótesis y modelo reducido . . . 5
- 2.2. Estadístico de prueba y conclusión . . . 5

3. Pregunta 3 **6**
- 3.1. Prueba de hipótesis y prueba de hipótesis matricial . . . 6
- 3.2. Estadístico de prueba . . . 7

4. Pregunta 4 **7**
- 4.1. Supuestos del modelo . . . 7
  - 4.1.1. Normalidad de los residuales . . . 7
  - 4.1.2. Varianza constante . . . 9
- 4.2. Verificación de las observaciones . . . 9
  - 4.2.1. Datos atípicos . . . 10
  - 4.2.2. Puntos de balanceo . . . 11
  - 4.2.3. Puntos influenciales . . . 12
- 4.3. Conclusión . . . 14

Índice de figuras

1. Gráfico cuantil-cuantil y normalidad de residuales 8
2. Gráfico residuales estudentizados vs valores ajustados 9
3. Identificación de datos atípicos 10
4. Identificación de puntos de balanceo 11
5. Criterio distancias de Cook para puntos influenciales 12
6. Criterio Dffits para puntos influenciales 13

Índice de cuadros

1. Tabla de valores coeficientes del modelo 3
2. Tabla ANOVA para el modelo 4
3. Resumen de los coeficientes 4
4. Resumen tabla de todas las regresiones 5

1. Pregunta 1

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

Donde,

- 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
β_0	-0.7609
β_1	0.1714
β_2	0.0115
β_3	0.0709
β_4	0.0134
β_5	0.0025

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.7609 + 0.1714 X_{1i} + 0.0115 X_{2i} + 0.0709 X_{3i} + 0.0134 X_{4i} + 0.0025 X_{5i};\ 1 \leqslant i \leqslant 64$$

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:

$$F_0 = \frac{MSR}{MSE} \overset{f_0}{\sim} f_{5,58} \tag{1}$$

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	85.7240	5	17.144793	23.964	5.50379e-13
Error	41.4954	58	0.715438		

De la tabla Anova, se observa un valor P igual a 5.50379e-13, como Vp < 0.05, entonces 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, es decir, el promedio del riesgo de infeccion depende significativamente de al menos una de las variables.

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	-0.7609	1.3973	-0.5445	0.5882
β_1	0.1714	0.0743	2.3059	0.0247
β_2	0.0115	0.0272	0.4247	0.6726
β_3	0.0709	0.0128	5.5307	0.0000
β_4	0.0134	0.0068	1.9598	0.0548
β_5	0.0025	0.0008	3.1595	0.0025

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 individualmente cuando los demás parámetros se mantienen constantes, pues sus P-valores son menores a α, por lo que se podría decir que los parámetros β_2, y β_4 no son significativos individualmente, cuando los otros se mantienen fijos.

1.4. Interpretación de los parámetros

Solo se pueden interpretar los parámetros significativos:

$\hat{\beta}_1$**:** Tenemos $\hat{\beta}_1 = 0.1714$ esto indica que, por cada unidad de aumento en la duración de la estadía, en promedio, el riesgo de infección aumenta en 0.1714 unidades cuando las demás variables predictoras se mantienen fijas.

$\hat{\beta}_3$**:** Tenemos $\hat{\beta}_3 = 0.0709$ esto indica que, por cada unidad de aumento en el número de camas, el promedio del riesgo de infección aumenta 0.0709 unidades cuando las demás variables predictoras se mantienen constantes.

$\hat{\beta}_5$**:** Tenemos $\hat{\beta}_5 = 0.0025$ esto indica que, por cada unidad de aumento en el numero de enfermeras, el promedio del riesgo de infección aumenta en 0.0025 unidades cuando las demás variables predictoras se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.6738$, lo que significa que aproximadamente el 67.38 % de la variabilidad total observada en el riesgo de infección es explicada por el modelo de regresión propuesto en el presente informe.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con el P-valor más alto en el modelo fueron X_2, X_4, 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	41.495	X1 X2 X3 X4 X5
Modelo reducido	87.851	X2 X4

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

$$Y_i = \beta_0 + \beta_2 X_{2i} + \beta_4 X_{4i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$
\begin{aligned}
F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_4) - SSE(\beta_0, \ \cdots, \ \beta_5))/3}{MSE(\beta_0, \ \cdots, \ \beta_5)} \overset{H_0}{\sim} f_{3,58} \\
&= \frac{15.452}{0.715438} \\
&= 21.5979
\end{aligned} \tag{2}
$$

Ahora, comparando el F_0 con $f_{0.95,3,58} = 2.7636$, se puede ver que $F_0 > f_{0.95,3,58}$, es decir, pertenece a la región de rechazo, por tanto se rechaza la hipótesis nula donde $\beta_1 = \beta_3 = \beta_5$=0, aceptando la hipotesis alternativa donde alguno de estos betas es diferente de 0. Se concluye que en conjunto estas tres variables son significativas para el modelo y no deben ser descartadas.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

¿Existe una relación estadística entre X_1, X_3 y X_4 y X_5? ¿Es posible compararlas?

Siendo:

- X_1: Duración de la estadía
- X_3: Número de camas
- X_4: Censo promedio diario
- X_5: Número de enfermeras

$$
\begin{cases}
\mathrm{H}_0 : \beta_1 = \beta_3, \beta_4 = \beta_5 \\
\mathrm{H}_1 : \beta_1 \neq \beta_3, \beta_4 \neq \beta_5
\end{cases}
$$

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

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

El modelo reducido está dado por:

$$
Y_i = \beta_0 + \beta_1 X_{1i}^* + \beta_2 X_{2i} + \beta_4 X_{4i}^* + \varepsilon_i, \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); \ 1 \leqslant i \leqslant 64
$$

Donde,

- $X^*_{1i} = X_{1i} + X_{3i}$
- $X^*_{4i} = X_{4i} + X_{5i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,58} \tag{3}$$

Vamos a reemplazar en la formula.

$$F_0 = \frac{(SSE(MR) - 41.4954)/2}{0.7154} \overset{H_0}{\sim} f_{2,58} \tag{4}$$

Se observa que al no tener todos los valores de la expresión, no se puede llegar a un resultado numerico para realizar un analisis de este, por tal motivo se dejan expresados los valores desconocidos y se reemplazan los valores que conocemos.

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Para la validación de este supuesto, haremos uso de la prueba shapiro-wilk, acompañada de un grafico cuantil-cuantil:

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

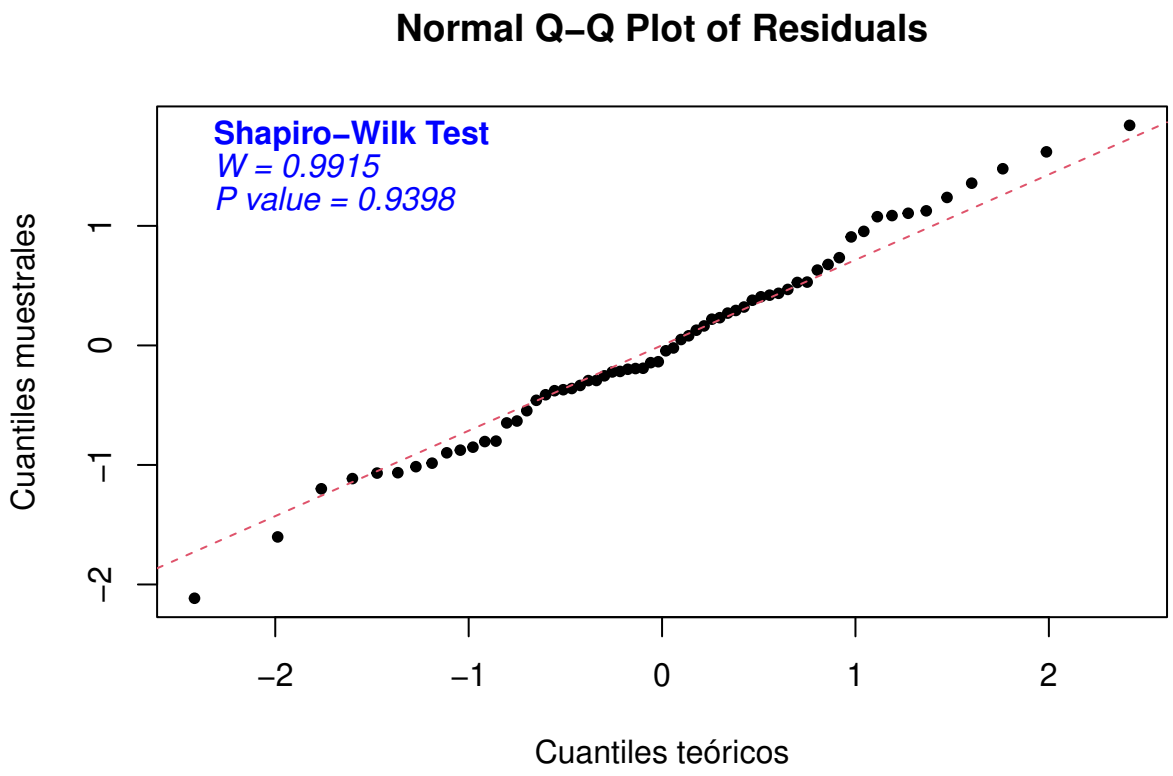

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.9398 y teniendo en cuenta que el nivel de significancia $\alpha = 0.05$, el P-valor es mayor y por lo tanto, no se rechaza la hipotesis nula, es decir que los datos distribuyen normal μ y varianza σ^2, sin embargo la gráfica de comparación de cuantiles permite ver patrones irregulares en las colas,debido a que se presentan datos alejados de la linea de regresión,por lo que se evidencia en la grafica determinamos que los supuestos no se distribuyen normal.

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

4.1.2. Varianza constante

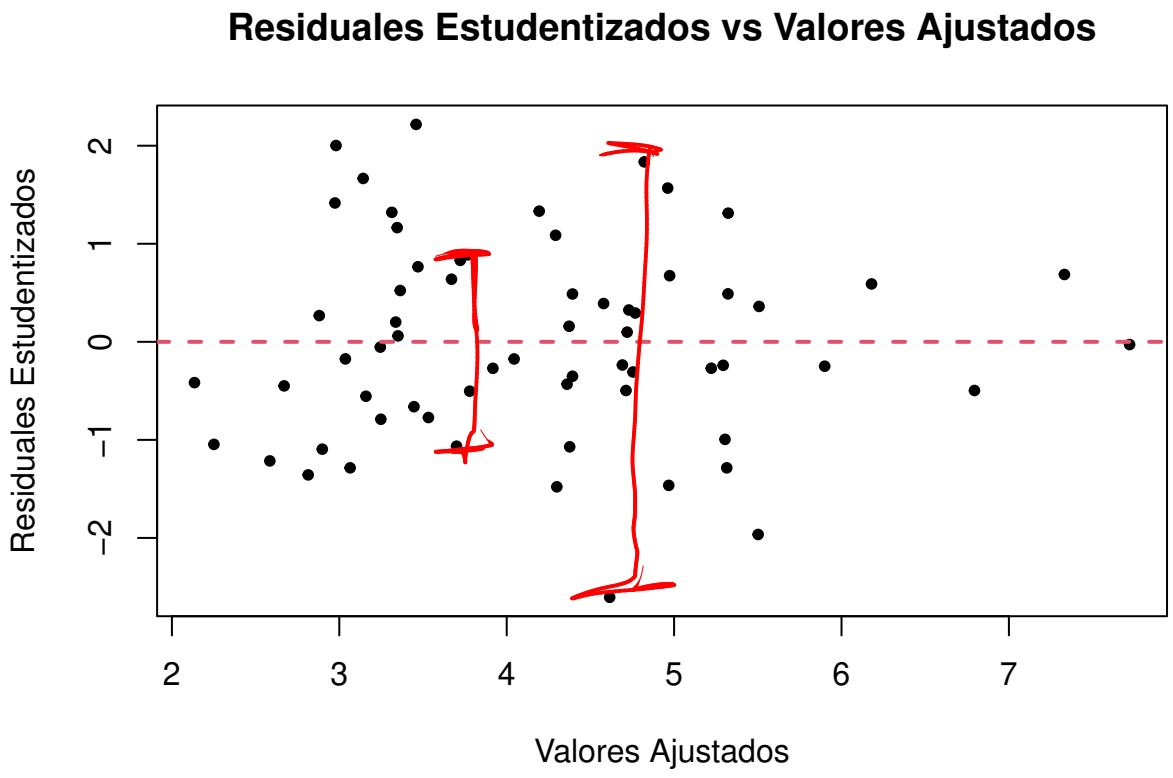

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que no hay patrones en los que la varianza aumente o decrezca, aceptamos este supuesto como cierto ya que cumple con todos los parametros para determinar que es varianza constante. Además es posible observar media 0.

4.2. Verificación de las observaciones

Se realizaran las verificaciones de las observaciones teniendo en cuenta las tablas y graficas generadas en los siguientes puntos, analizando las tendencias de los datos y concluyendo si el modelo es apto o no es apto.

4.2.1. Datos atípicos

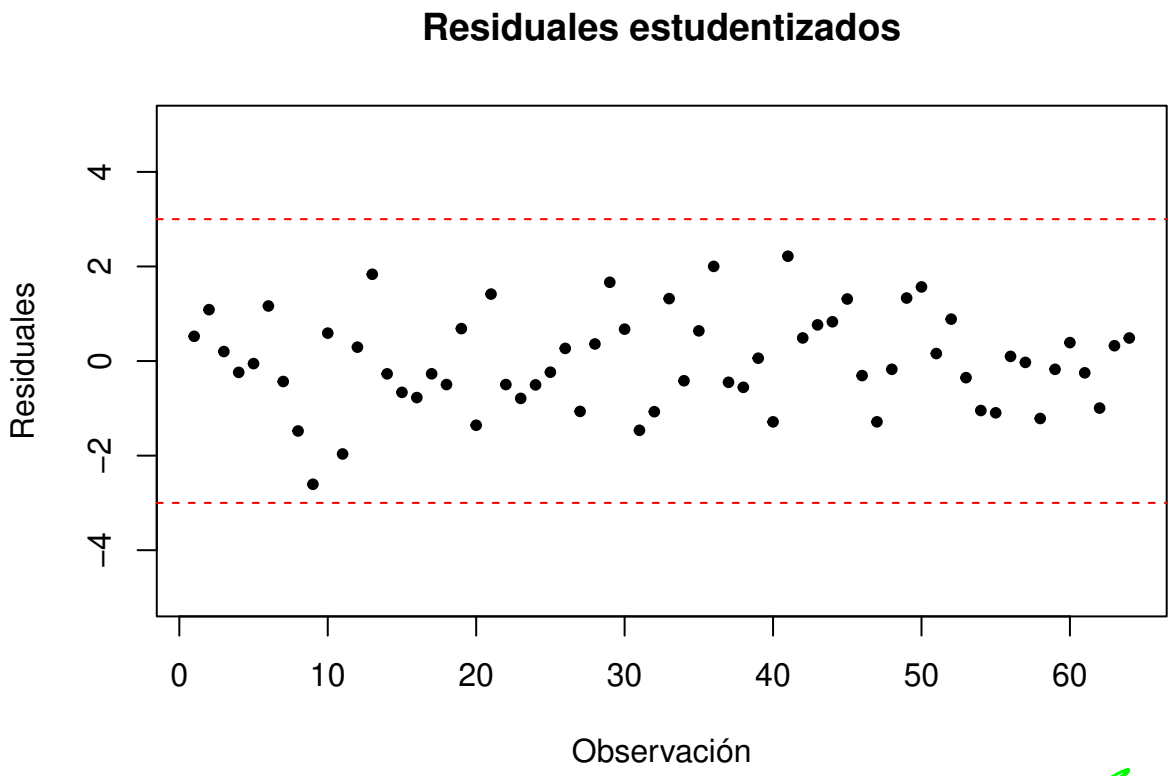

Figura 3: Identificación de datos atípicos

Como se puede observar en la gráfica anterior, no hay datos atípicos pues ningún residual estudentizado sobrepasa el criterio de $|r_{estud}| > 3$ como limite superior, ni el criterio de $|r_{estud}| < -3$ como limite inferior.

4.2.2. Puntos de balanceo

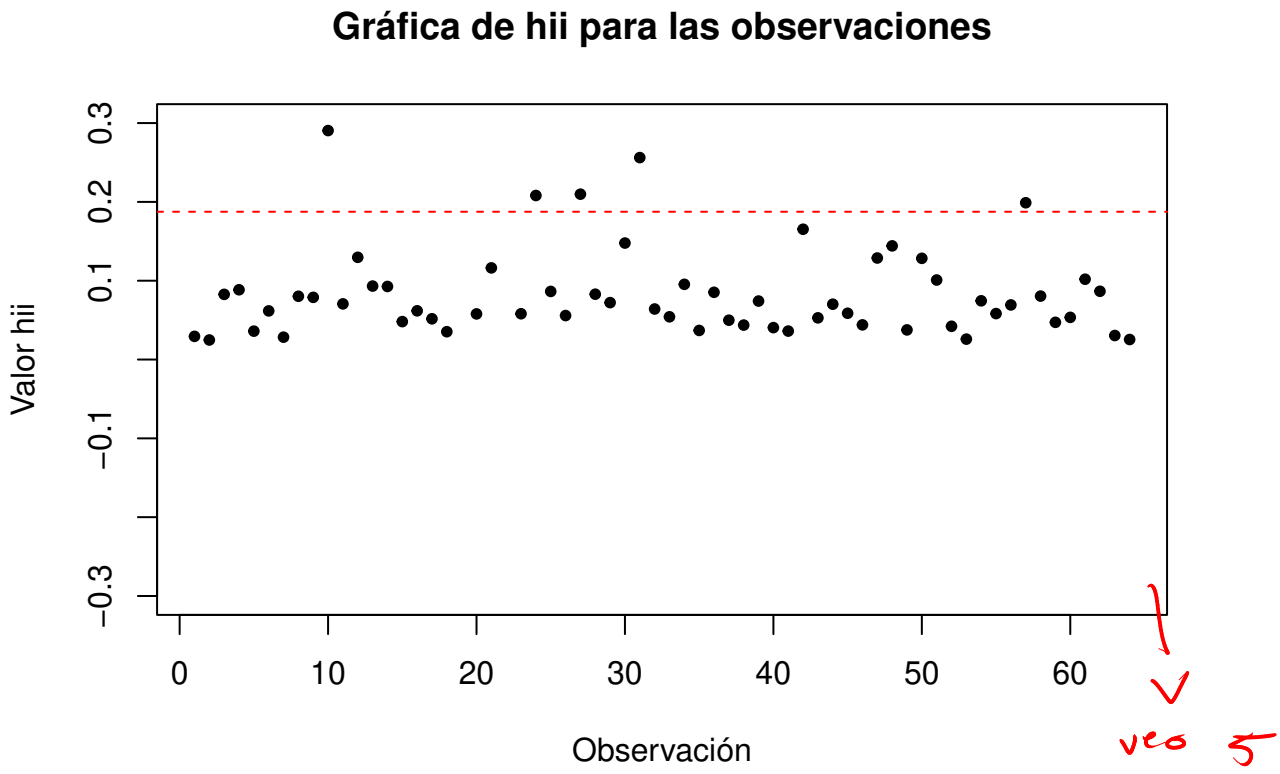

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 10   0.5902  0.0238    0.2905  0.3755
## 19   0.6865  0.0422    0.3494  0.5008
## 22  -0.4968  0.0426    0.5091 -0.5026
## 24  -0.5034  0.0111    0.2081 -0.2564
## 27  -1.0647  0.0502    0.2098 -0.5493
## 31  -1.4642  0.1231    0.2562 -0.8681
## 57  -0.0280  0.0000    0.1989 -0.0138
```

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 5 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.

El valor de $h_{ii} = 2\frac{p}{n}$ es:

$h_{ii} = 2\frac{5}{64}$

$h_{ii} = 2\frac{5}{32}$

$h_{ii} = 0.15625$

Fijandonos que valores especificos sobrepasan el valor de h_{ii},tenemos que:

Observación 19 con hii.value =0.3494

Observación 22 con hii.value =0.5091

Observación 24 con hii.value =0.2081

Observación 27 con hii.value =0.2098

Observación 31 con hii.value =0.2562

4.2.3. Puntos influenciales

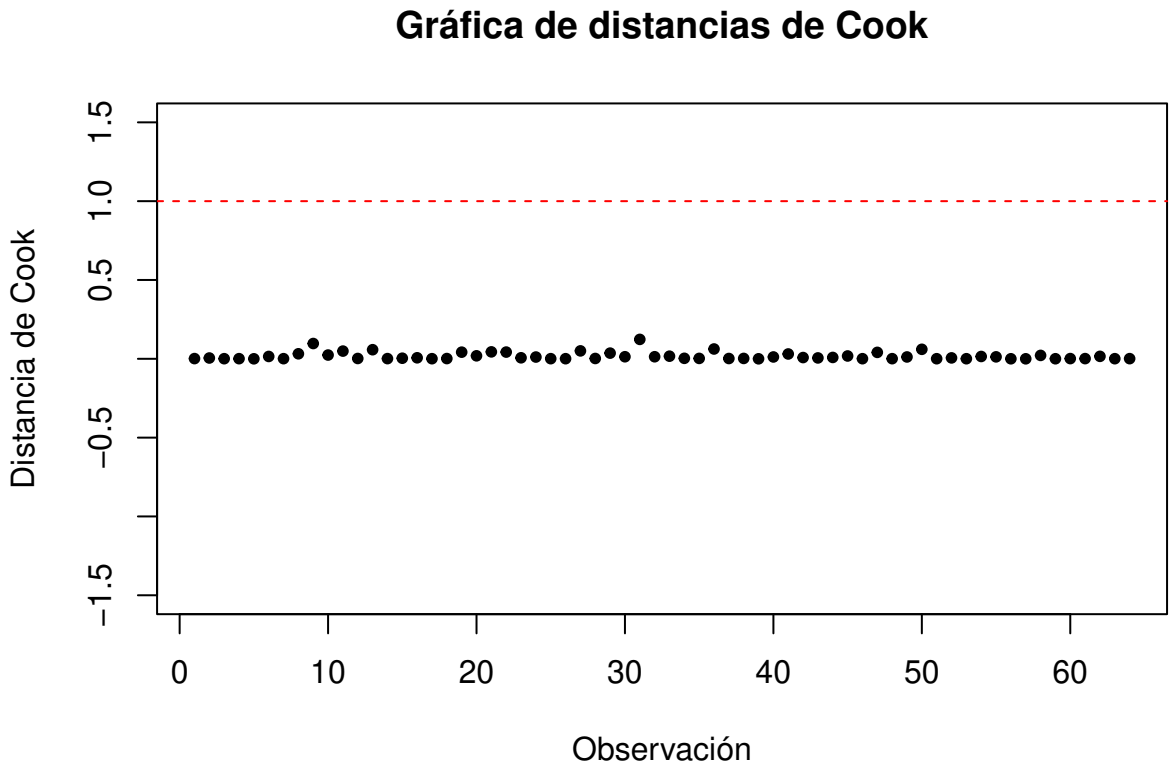

Figura 5: Criterio distancias de Cook para puntos influenciales

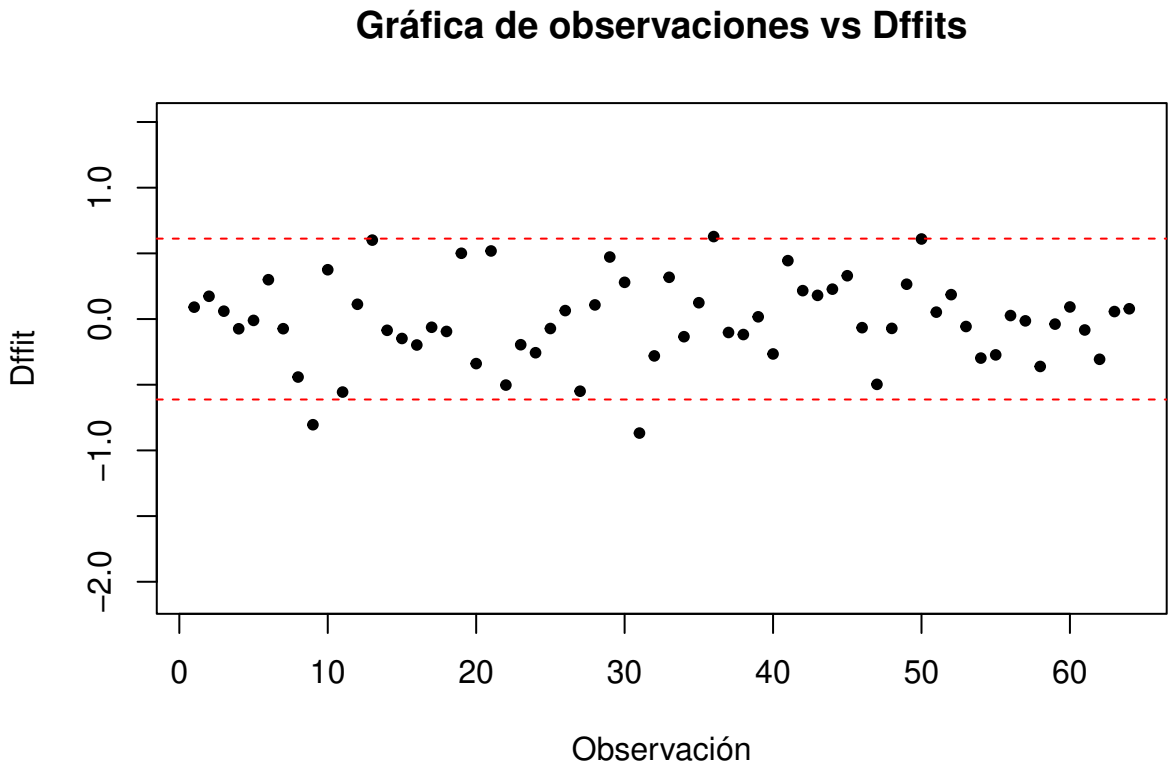

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 9   -2.6057  0.0969    0.0789 -0.8044
## 31  -1.4642  0.1231    0.2562 -0.8681
## 36   2.0019  0.0622    0.0853  0.6279
```

Al realizar el analisis si realizamos el analisis bajo el criterio de Dffits,el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, notamos que ninguno de los puntos cumple con esta caracteristica.Por otro lado, siguiendo el criterio de distancias de Cook, ningún dato de la columna de Cook es mayor a 1 y se dice que bajo este criterio un punto es de influencia si $D_i > 1$ tampoco se cumple que algun punto sea de influencia.

$|D_{ffit}| > 2\sqrt{\frac{p}{n}}$ $|D_{ffit}| = 0.5590$

Observación 9 con Dffits = -0.8044

Observación 31 con Dffits = -0.8681

Observación 36 con Dffits = 0.6279

Causan...? 3pt

$D_i > 1$

Observación 9 con CookD =0.0969

Observación 31 con CookD =0.1231

Observación 36 con CookD =0.0622

4.3. Conclusión

3 pt

Concluimos que el modelo es apto, cumple con los requerimientos basicos para determinar que la regresion lineal es valida,cumple con la linealidad, la homocedasticidad en los patrones de varianza, es decir que su varianza en constance, con una distribucion normal y una cantidad de 5 puntos de balanceo