

# Trabajo 1

4,1

Estudiantes

**Felipe Osorio Sepulveda**  
**Nicolas Stiven Torres Rodriguez**  
**Juan Diego Ramirez Ruiz**  
**Felipe Vasquez Zuluaga**

Equipo

Docente

**Julieth Veronica Guarin Escudero**

Asignatura

**Estadística II**



UNIVERSIDAD  
**NACIONAL**  
DE COLOMBIA

Sede Medellín  
5 de octubre de 2023

# Índice

<b>1. Pregunta 1</b>	<b>3</b>
1.1. Modelo de regresión . . . . .	4
1.2. Significancia de la regresión . . . . .	4
1.3. Significancia de los parámetros . . . . .	5
1.4. Interpretación de los parámetros . . . . .	6
1.5. Coeficiente de determinación múltiple $R^2$ . . . . .	6
<b>2. Pregunta 2</b>	<b>6</b>
2.1. Planteamiento pruebas de hipótesis y modelo reducido . . . . .	6
2.2. Estadístico de prueba y conclusión . . . . .	7
<b>3. Pregunta 3</b>	<b>7</b>
3.1. Prueba de hipótesis y prueba de hipótesis matricial . . . . .	7
3.2. Estadístico de prueba . . . . .	8
<b>4. Pregunta 4</b>	<b>9</b>
4.1. Supuestos del modelo . . . . .	9
4.1.1. Normalidad de los residuales . . . . .	9
4.1.2. Varianza constante . . . . .	10
4.2. Verificación de los valores extremos . . . . .	11
4.2.1. Datos atípicos . . . . .	11
4.2.2. Puntos de balanceo . . . . .	12
4.2.3. Puntos influyentes . . . . .	13
4.3. Conclusión . . . . .	14

## Índice de figuras

1.	Grafica de matriz de dispersion con boxplot y correlaciones . . . . .	3
2.	Gráfico cuantil-cuantil y normalidad de residuales . . . . .	9
3.	Gráfico residuales estudentizados vs valores ajustados . . . . .	10
4.	Identificación de datos atípicos . . . . .	11
5.	Identificación de puntos de balanceo . . . . .	12
6.	Criterio distancias de Cook para puntos influenciales . . . . .	13
7.	Criterio Dffits para puntos influenciales . . . . .	14

## Índice de cuadros

1.	Tabla de valores coeficientes del modelo . . . . .	4
2.	Tabla ANOVA del modelo . . . . .	5
3.	Resumen de los coeficientes . . . . .	5
4.	Resumen tabla de todas las regresiones . . . . .	6

# 1. Pregunta 1

Previamente se realizó una observación del siguiente gráfico para entender la relación y el comportamiento entre cada una de las variables predictoras con la variable respuesta. Adicionalmente se espera que entre las variables regresoras no exista una relación lineal importante para evitar posibles problemas en el modelo.

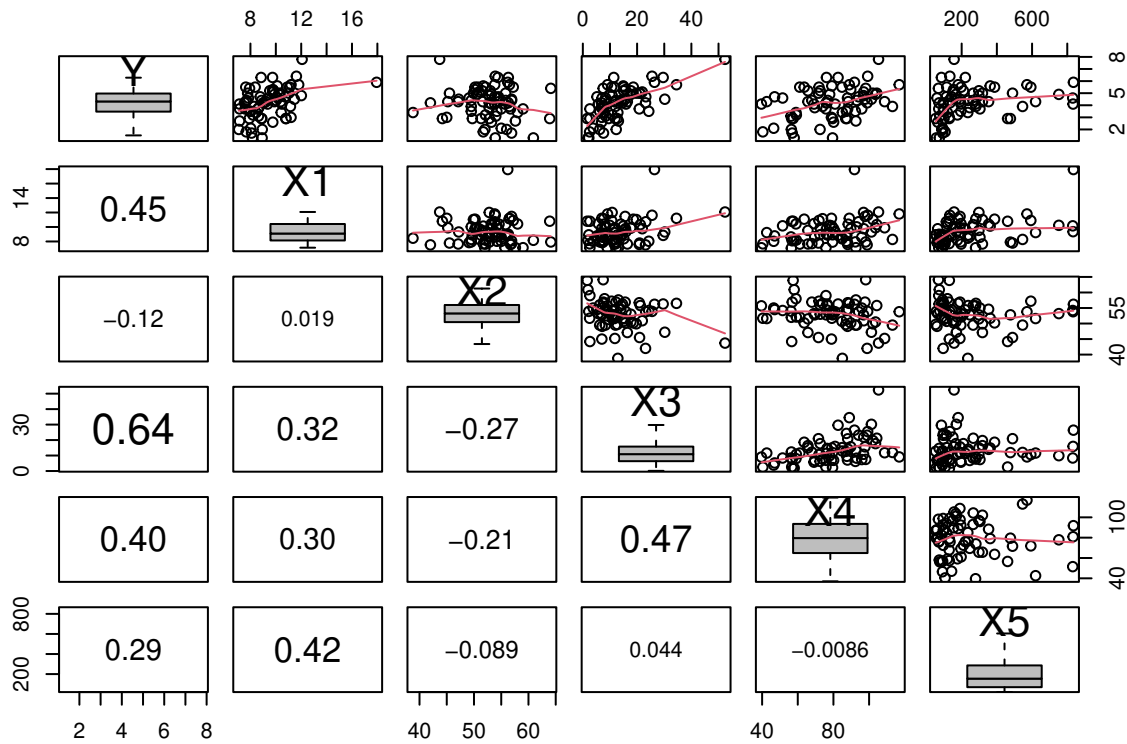


Figura 1: Grafica de matriz de dispersion con boxplot y correlaciones

Con base en el análisis de la matriz de gráficos de dispersión se puede plantear un modelo de regresión lineal Múltiple.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

Bajo los supuestos de:

$$\varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2); \quad i = 1, \dots, 69$$

En forma matricial:

$$\underline{Y} = \underline{\beta} \underline{X} + \underline{\varepsilon}$$

$$\underline{\varepsilon} \stackrel{\text{iid}}{\sim} N_n(0, \sigma^2 I)$$

Las variables utilizadas para explicar el riesgo de infección (Y) serán:

No la  
7 analizaron ni  
siguiera XD

- $X_1$ : 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

Posteriormente se realiza una estimación de cada uno de los parámetros con la cual se ajusta el modelo inicial dando como resultado:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} + \hat{\beta}_4 X_{i4} + \hat{\beta}_5 X_{i5}$$

Cuadro 1: Tabla de valores coeficientes del modelo

Valor estimado del parametro	
$\beta_0$	-0.0204
$\beta_1$	0.1113
$\beta_2$	0.0199
$\beta_3$	0.0827
$\beta_4$	0.0083
$\beta_5$	0.0014

Por lo tanto, la ecuación ajustada es:

$$\hat{Y}_i = -0.0204 + 0.1113X_{1i} + 0.0199X_{2i} + 0.0827X_{3i} + 0.0083X_{4i} + 0.0014X_{5i}$$

## 1.2. Significancia de la regresión

Se quiere verificar si el conjunto de variables predictoras explica significativamente el riesgo de infección, para ello se plantea el siguiente juego de hipótesis:

$$\left\{ H_0 : \beta_j; \forall j = 1, 2, 3, 4, 5 \right\} \quad VS \quad H_1 : \beta_j \neq 0; \text{para algún } j = 1, 2, 3, 4, 5$$

Cuyo estadístico de prueba es:

$$F_0 = \frac{SSR/k}{SSE/(n-p)} = \frac{MSR}{MSE} \quad \sim ? \quad (1)$$

Acontuniación, la tabla ANOVA:

Cuadro 2: Tabla ANOVA del modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	$F_0$	P-valor
Regresion	60.3366	5	12.067327	13.7113	4.67748e-09
Error	55.4463	63	0.880099		

De la tabla ANOVA; observamos que el valor p para esta prueba de hipótesis es aproximadamente 0, en consecuencia, podemos concluir que este conjunto de variables efectivamente si explica significativamente el riesgo de infección(Y).  $\rightarrow$  conclusión un poco meh...

### 1.3. Significancia de los parámetros

Se quiere verificar que parámetros de forma individual son significativos en presencia de los demás, para esto se plantean las siguientes hipótesis:

$$H_0 : \beta_j = 0 \quad H_1 : \beta_j \neq 0 \text{ para } j = 1, 2, 3, 4, 5$$

Con su respectivo estadístico de prueba:

$$T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \quad (2)$$

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
$\beta_0$	-0.0204	1.5907	-0.0128	0.9898
$\beta_1$	0.1113	0.0810	1.3743	0.1742
$\beta_2$	0.0199	0.0265	0.7505	0.4558
$\beta_3$	0.0827	0.0156	5.3090	0.0000
$\beta_4$	0.0083	0.0073	1.1467	0.2559
$\beta_5$	0.0014	0.0006	2.1573	0.0348

Por medio de la tabla de resumen de los coeficientes podemos concluir que:

- $\beta_0, \beta_1, \beta_2, \beta_4$  son individualmente no significativos en presencia de los demás parámetros.
- $\beta_3, \beta_5$  son individualmente significativos en presencia de los demás parámetros.

## 1.4. Interpretación de los parámetros

Solo se interpretarán  $\beta_3$  y  $\beta_5$  debido a que son susceptibles a interpretación, dado que son significativos individualmente.

- $\hat{\beta}_3 = 0.0827$ : Indica que por cada unidad que aumente el número promedio de camas en el hospital durante el periodo del estudio, la probabilidad promedio estimada de adquirir infección en el hospital aumentara en 0.0827
- $\hat{\beta}_5 = 0.0014$ : Indica que por cada unidad que aumente el Número promedio de enfermeras, equivalentes a tiempo completo durante el periodo del estudio, la probabilidad promedio estimada de adquirir infección en el hospital aumentara en 0.0014

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

$$R^2 = \frac{60.3366}{55.4463 + 60.3366} = 0.5211$$

Esto indica que el 52.11 % de la variabilidad total de la variable respuesta, es explicada por el modelo propuesto en el presente trabajo. Para tomar decisiones más acertadas sobre el ajuste del modelo, se prefiere usar el  $R^2$  ajustado, debido a que penaliza el número de variables incluidas en el modelo.

$$R^2_{adj} = 1 - \frac{(68)MSE}{SST} = 0.4831$$

Como el  $R^2_{adj}$  es menor que el  $R^2$  indica que puede existir variables que realmente no aportan significativamente al modelo. En otras palabras, se puede depurar el modelo, eliminando las variables que no aportan significativamente.

## 2. Pregunta 2

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

Las variables cuyos parámetros tuvieron el valor P más pequeño son:  $X_1, X_3, X_5$

Cuadro 4: Resumen tabla de todas las regresiones

	$SSE$	Variables en el modelo				
Modelo completo	55.446	X1	X2	X3	X4	X5
Modelo reducido	96.822		X2	X4		

Queremos verificar la significancia del subconjunto anteriormente mencionado, por lo que nos apoyaremos en la tabla de todas las regresiones posibles, para afirmar o rechazar su relevancia en el modelo por tanto se propone el siguiente juego de hipótesis. es:

$$H_0 : \beta_1 = \beta_3 = \beta_5 = 0 \quad \text{V.S.} \quad H_1 : \beta_1 \neq 0 \text{ ó } \beta_3 \neq 0 \text{ ó } \beta_5 \neq 0$$

va antes de tabla

El modelo reducido bajo la hipótesis nula es:

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon; \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2); i = 1, \dots, 69$$

huh?

## 2.2. Estadístico de prueba y conclusión

Para realizar la prueba de significancia del subconjunto debemos utilizar el estadístico de prueba:

$$F_0 = \frac{MS_{extra}}{MSE} = \frac{SS_{extra}/3}{MSE} \stackrel{H_0}{\sim} f_{3,63}$$

$$F_0 = \frac{[SSE(\beta_0, \beta_2, \beta_4) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)]/3}{MSE} = \frac{[96.822 - 55.446]/3}{0.8801}$$

$$F_0 = 15.671$$

1 pt



Como  $F_0 = 15.671 > f_{0.05, 3, 63} = 2.7505$ , denotamos que el subconjunto como unidad es significativo, sin embargo, no podemos aseverar que cada elemento individualmente es significativo; por tanto, el subconjunto debe continuar en el modelo. En consecuencia, podemos inferir que la probabilidad promedio de adquirir la infección, ciertamente está influenciada por la duración promedio de la estadía de los pacientes, el número promedio de camas en el hospital o el número promedio de enfermeras.

2 pt



## 3. Pregunta 3

4,5 pt

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

- • ¿Existe evidencia estadística para concluir que el aumento en una unidad en la duración promedio de estadía en el hospital (X1) tiene un mismo efecto en la probabilidad promedio estimada de adquirir una infección en el hospital en comparación con el aumento en una unidad en el número promedio de pacientes en el hospital por día (X4)?
- • ¿Al disminuir en una unidad el número promedio de camas en el hospital durante el periodo del estudio (X3), la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje) decrece en la misma medida que si mermara en una unidad el número promedio de enfermeras en el hospital (X5)?



$$\begin{cases} H_0 : \beta_1 - \beta_4 = 0 \\ \quad \beta_3 - \beta_5 = 0 \\ H_1 : \beta_1 - \beta_4 \neq 0 \\ \quad \beta_3 - \beta_5 \neq 0 \end{cases} \quad \text{ó !!}$$

En forma matricial se puede escribir como:

$$\begin{cases} H_0 : \mathbf{L}\underline{\beta} = \underline{0} \\ H_1 : \mathbf{L}\underline{\beta} \neq \underline{0} \end{cases}$$

Con  $\mathbf{L}$  dada por

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

El modelo reducido(MR) bajo  $H_0$  es:

$$Y_i = \beta_0 + \beta_1 X_{i1,4}^* + \beta_2 X_{i2}^* + \beta_3 X_{i3,5} + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2); i = 1, \dots, 69$$

Donde

- $X_{1,4} = X_1 + X_4$
- $X_{3,5} = X_3 + X_5$

### 3.2. Estadístico de prueba

Recordemos que el Modelo Completo (MF):

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

El estadístico de prueba  $F_0$  esta definido por

$$F_0 = \frac{MSH}{MSE(MF)} \stackrel{H_0}{\sim} f_{2,63}$$

$$F_0 = \frac{MSH}{MSE(MF)} = \frac{SSH/2}{MSE(MF)} = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} = \frac{(SSE(MR) - 55.446)/2}{0,8801}$$

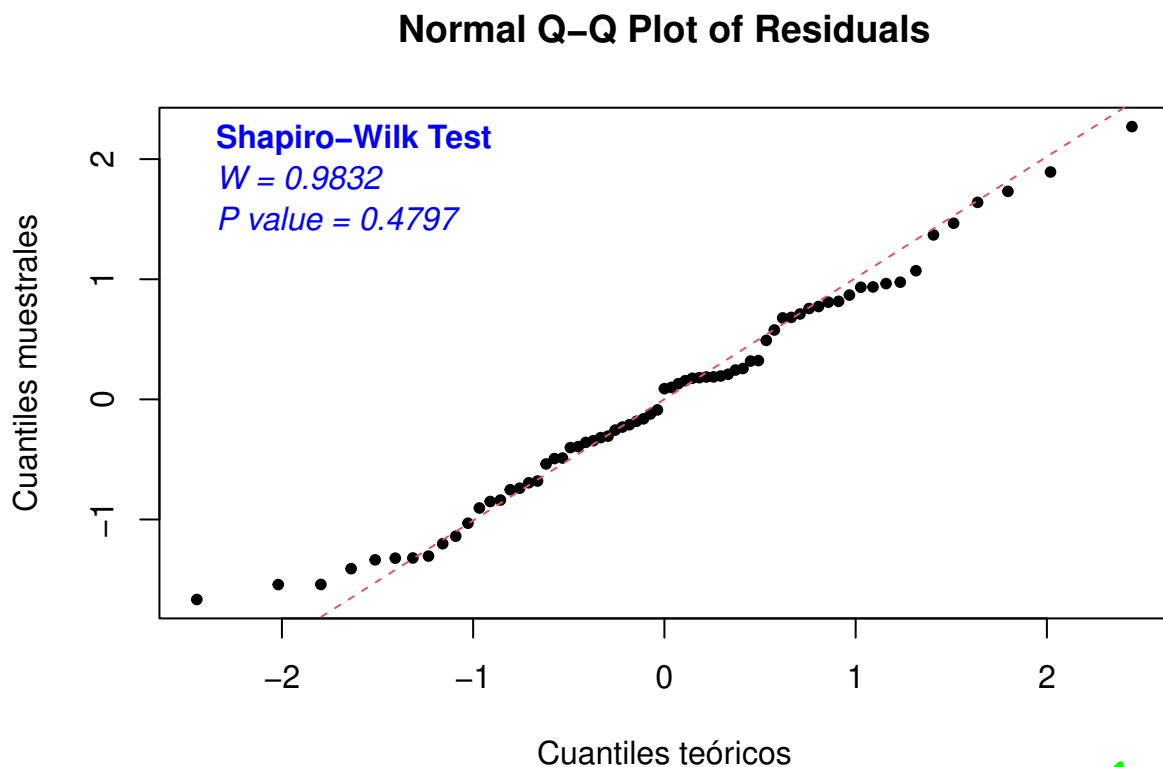
## 4. Pregunta 4 1a p+

### 4.1. Supuestos del modelo

#### 4.1.1. Normalidad de los residuales

Para la validación de este supuesto, se planteará el siguiente juego de hipótesis que se verificara por medio de la prueba shapiro-wilk, acompañada de un gráfico cuantil-cuantil:

$$\begin{cases} H_0 : \varepsilon_i \sim \text{Normal} \\ H_1 : \varepsilon_i \not\sim \text{Normal} \end{cases}$$



4pt

Figura 2: Gráfico cuantil-cuantil y normalidad de residuales

Graficamente se puede visualizar que los residuales del modelo se acoplan bastante bien a la linea roja mientras mas nos acercamos al centro, y cuando se aleja del medio, podemos notar que los residuales se van alejando de la linea, aun asi, se puede ver que la distancia ente los residuales y la linea no es muy grande, esta diferencia puede ser ocasionado por puntos de balanceo. Asi, se concluye que el supuesto de normalidad se cumple. Ademas por medio de la prueba teorica Shapiro-Wilk test, podemos soportar esta conclusion debido a que el valor  $p = 0.4797$  es mayor a  $\alpha = 0.05$ .

## 4.1.2. Varianza constante

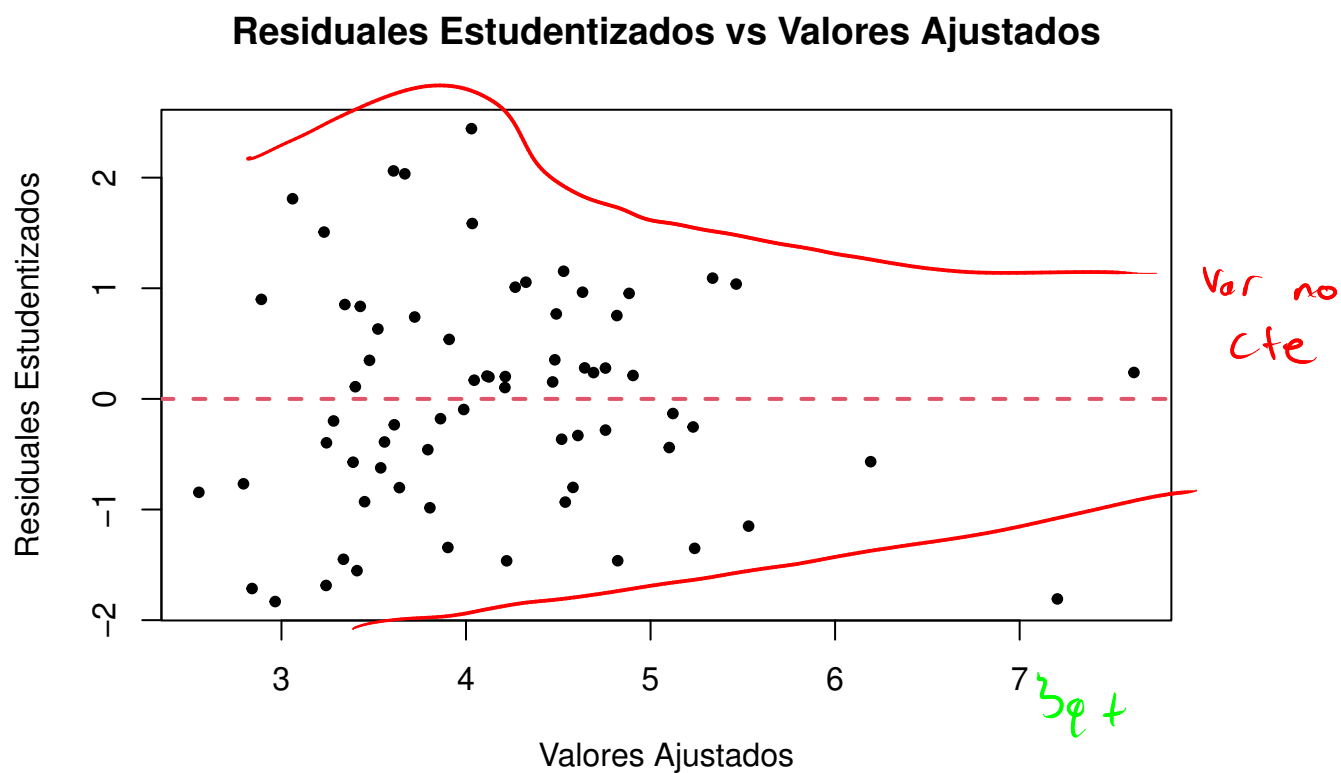


Figura 3: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que existe un patron marcado en la parte superior y uno no tan marcado en la parte inferior, en los cuales la varianza decrece, por tanto, el supuesto de varianza constante no se ~~puede~~ confirmar. no se cumple.

## 4.2. Verificación de los valores extremos

### 4.2.1. Datos atípicos

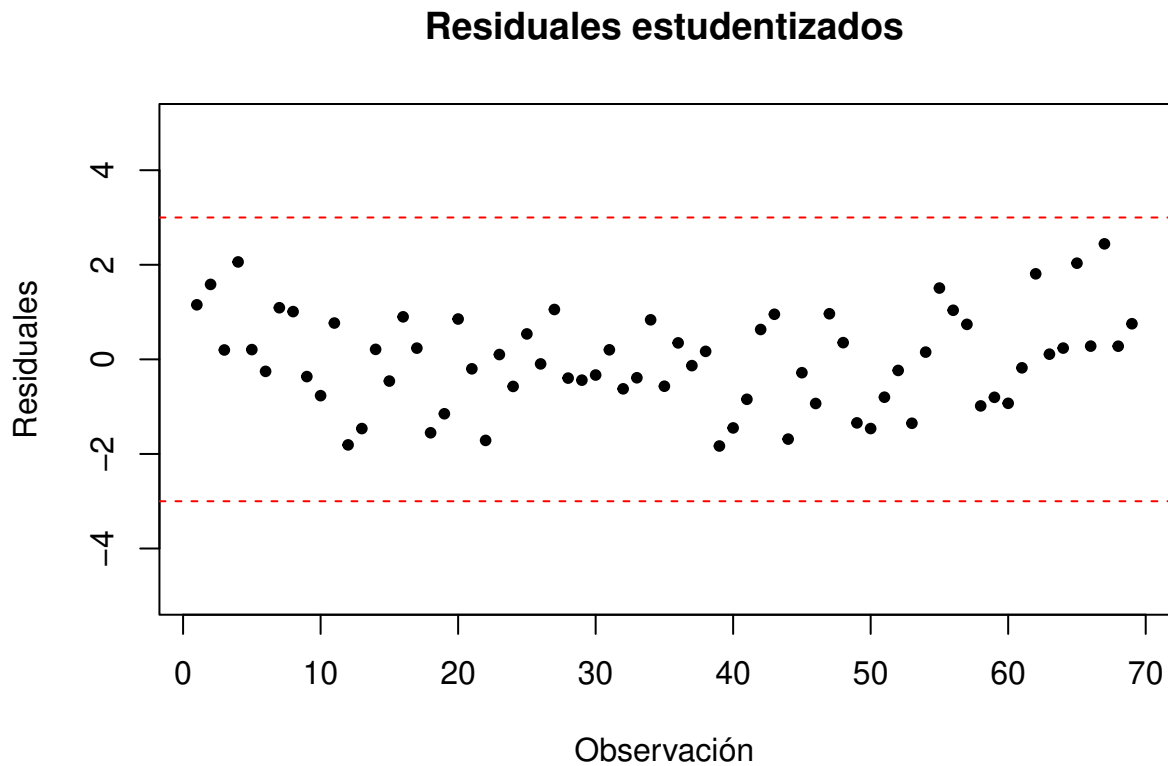


Figura 4: Identificación de datos atípicos

De la figura 3, se puede corroborar que ninguna observación de la muestra es atípica, puesto a que ningún residual estudentizado sobrepasa los límites delimitados; No se cumple el criterio de identificación de observaciones atípicas:  $|r_i| > 3$ .

✓ 3 pt

## 4.2.2. Puntos de balanceo

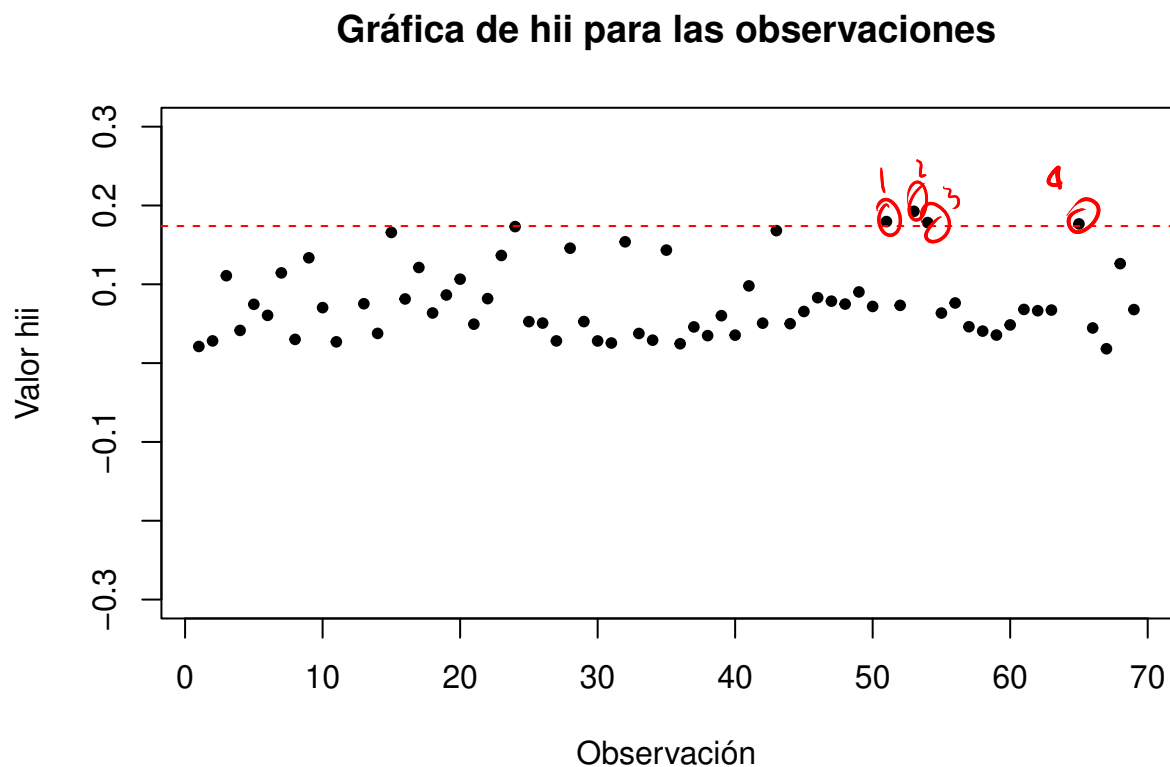


Figura 5: Identificación de puntos de balanceo

##	res.stud	Cooks.D	hii.value	Dffits
## 12	-1.8082	0.3766	0.4086	-1.5315
## 51	-0.8000	0.0234	0.1796	-0.3733
## 53	-1.3511	0.0726	0.1926	-0.6644
## 54	0.1533	0.0009	0.1783	0.0709
## 64	0.2388	0.0051	0.3500	0.1739
## 65	2.0343	0.1480	0.1767	0.9673

Presentan 6 pero sólo se ven 4 en la gráfica.

Al observar la gráfica de observaciones vs valores  $h_{ii}$ , se puede notar algunos puntos que sobrepasan la línea roja, esta representa el valor límite para decidir si es o no un punto de balanceo, este valor es:  $h_{ii} = \frac{2p}{n} = 0.1739$ .

En esta muestra existen 6 puntos de balanceo que corresponden a las observaciones 12, 51, 53, 54, 64, 65.

2pt

#### 4.2.3. Puntos influenciales

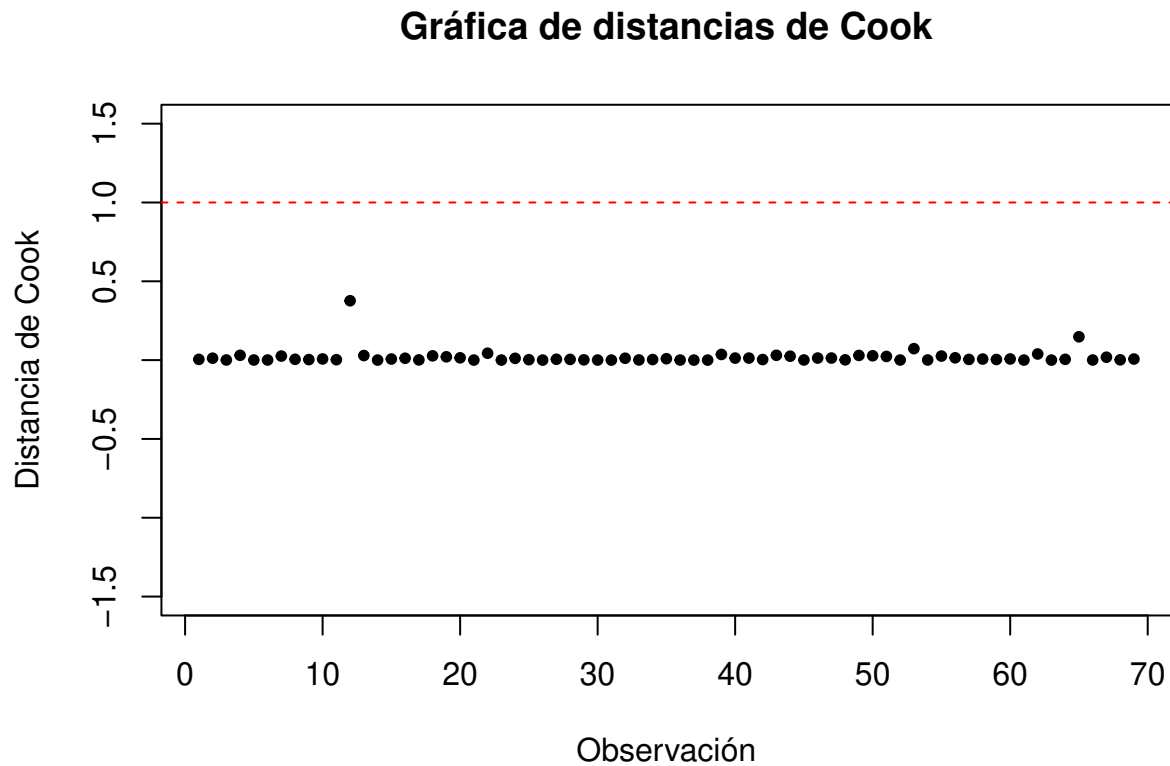


Figura 6: Criterio distancias de Cook para puntos influenciales

Visualmente se infiere que no existe ningún punto influyente debido a que ninguna distancia de Cook sobrepasa el límite de determinación de un punto influyente:  $D_i > 1$ .



### Gráfica de observaciones vs Dffits

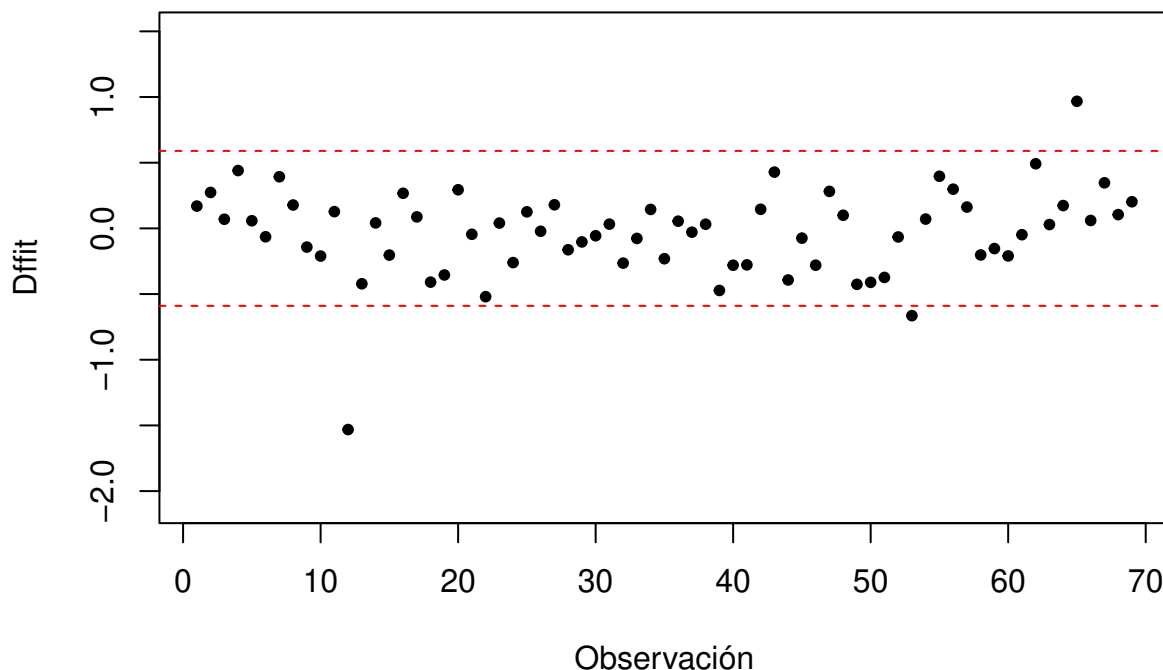


Figura 7: Criterio Dffits para puntos influenciales

##	res.stud	Cooks.D	hii.value	Dffits
## 12	-1.8082	0.3766	0.4086	-1.5315
## 53	-1.3511	0.0726	0.1926	-0.6644
## 65	2.0343	0.1480	0.1767	0.9673

4 p +


Para que un punto sea influyente por el criterio de Diagnostico DFFITS se debe cumplir que  $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$ , por medio de la grafica podemos notar que existen 3 puntos influenciales, los cuales son las observaciones 12, 53, 65.

### 4.3. Conclusión

3 p +

Como se observó anteriormente en el modelo se cumple el supuesto de normalidad, sin embargo, el supuesto de varianza constante no fue validado, por tanto el modelo no es valido para realizar predicciones y estimaciones sobre posibles valores de interes. Se debe tener en cuenta que ambas pruebas (Normalidad y varianza constante) fueron realizadas con posibles valores extremos que contuviera el modelo.

Con el fin de conocer cuales son estos valores extremos y cual es su efecto sobre el modelo, se hizo un analisis de observaciones extremas, donde se obtuvo:

- No hay observaciones atípicas en la muestra.
  - Las observaciones 12, 51, 53, 54, 64, 65 son puntos de balanceo que afectan el  $R^2$  y los errores estandar de cada parametro.
  - Existen 3 observaciones influenciales las cuales halan el modelo en su dirección, modifican notoriamente sus coeficientes, y en consecuencia la regresión se ve afectada; Estas observaciones son: 12, 53, 65.
- 

En efecto se obtuvieron valores extremos que deben ser investigados antes de usar el modelo, y de esta forma evaluar de nuevo su validez como estimador de la variable respuesta.