Trabajo 1

4,1

Estudiantes

Yuricik Cañas Quintero
Luzarait Cañas Quintero
Edwin David Noguera Pantoja
Andres Esteban Monsalve Vasquez
Equipo # 53

Docente

Carlos Mario Lopera Gomez

Asignatura

Estadística II

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medellín
Poner fecha de entrega

Índice

1. Pregunta 1 **3**

1.1. Modelo de regresión lineal múltiple . . . 3

1.2. Modelo de regresión ajustado . . . 3

1.3. Significancia de la regresión . . . 3

1.4. Significancia de los parámetros individuales . . . 4

1.5. Interpretación de los parámetros . . . 4

1.6. Calculo e interpretación del coeficiente de determinación múltiple R^2 . . . 4

2. Pregunta 2 **5**

2.1. Planteamiento prueba de hipótesis y modelo reducido . . . 5

2.2. Estadístico de prueba y conclusiones . . . 5

3. Pregunta 3 **5**

3.1. Prueba de hipótesis y prueba de hipótesis matricial . . . 5

3.2. Estadistico de prueba . . . 6

4. Pregunta 4 **6**

4.1. Supuestos del modelo . . . 6

4.1.1. Normalidad de los residuales . . . 6

4.1.2. Media 0 y Varianza constante . . . 7

4.2. Observaciones extremas . . . 8

4.2.1. Datos atípicos . . . 9

4.2.2. Puntos de balanceo . . . 10

4.2.3. Puntos influenciales . . . 10

Índice de figuras

1. Gráfico cuantil-cuantil y normalidad de los residuales 7
2. Gráfico residuales estudentizados vs valores ajustados 8
3. Identificación de datos atípicos 9
4. Identificación de puntos de balanceo 10
5. Criterio distancias de Cook para puntos influenciales 11
6. Criterio Dffits para puntos influenciales 12

Índice de tablas

1. Tabla de valores de los coeficientes estimados 3
2. Tabla anova significancia de la regresion 3
3. Resumen de los coeficientes 4
4. Resumen de todas las regresiones 5
5. Tabla de diagnostico 9

1. Pregunta 1

1.1. Modelo de regresión lineal múltiple

Teniendo en cuenta la base de datos asignada al Equipo53.txt, las covariables del modelo de regresión lineal múltiple son: Duración de la estadía (D), Rutina de cultivos (R), Número de camas (N), Censo promedio diario (C) y Número de enfermeras (E).

El modelo propuesto basado en lo anterior es:

$$X.Bfat_i = \beta_0 + \beta_1 D_i + \beta_2 R_i + \beta_3 N_i + \beta_4 C_i + \beta_5 E_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

donde $1 \leq i \leq 69$

1.2. Modelo de regresión ajustado

Después de ajustar el modelo inicial, se calculan los estimadores de los parámetros, se obtienen todos los coeficientes correspondientes.

Tabla 1: Tabla de valores de los coeficientes estimados

	Valor del parametro
$\hat{\beta}_0$	-0.8201
$\hat{\beta}_1$	0.1848
$\hat{\beta}_2$	0.0304
$\hat{\beta}_3$	0.0648
$\hat{\beta}_4$	0.0074
$\hat{\beta}_5$	0.0009

Por lo tanto, el modelo de regresión ajustado es:

$$\widehat{X.Bfat}_i = -0.8201 + 0.1848 D_i + 0.0304 R_i + 0.0648 N_i + 0.0074 C_i + 9 \times 10^{-4} E_i$$

Donde $1 \leq i \leq 69$

1.3. Significancia de la regresión

En base al modelo ajustado con los respectivos supuestos, y teniendo en cuenta las 5 variables regresoras, se realiza el siguiente análisis de la significancia de la regresión para probar si al menos una de las predictoras es relevante para el ajuste:

Tabla 2: Tabla anova significancia de la regresion

	Sumas de cuadrados	g.l	Cuadrado medio	F_0	Valor-P
Modelo de regresion	61.3830	5	12.27660	14.8498	1.2987e-09
Error	52.0834	63	0.82672		

De la tabla ANOVA, se obtiene un Valor-P tan pequeño (1.2987e-09) que al compararlo con un $\alpha = 0.05$ se puede concluir que, se rechaza la hipótesis nula para la no significancia de los parámetros y por

lo tanto se concluye que la regresión es significativa. Esto implica que al menos uno de los parámetros debe ser relevante para el modelo.

Hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \neq 0 \text{ para } j = 1, 2, 3, 4, 5 \end{cases}$$

1.4. Significancia de los parámetros individuales

En la tabla se muestra el resumen de los coeficientes y su información relevante, la cual permitirá determinar la significancia de cada uno de estos, individualmente:

Tabla 3: Resumen de los coeficientes

	Estimacion β_j	$se(\hat{\beta}_j)$	T_{0j}	Valor-P
β_0	-0.8201	1.4452	-0.5674	0.5724
β_1	0.1848	0.0824	2.2427	0.0284
β_2	0.0304	0.0256	1.1877	0.2394
β_3	0.0648	0.0133	4.8635	0.0000
β_4	0.0074	0.0065	1.1321	0.2619
β_5	0.0009	0.0006	1.5520	0.1257

En base a los resultados de la tabla, y de la columna de los Valores P se puede concluir que al compararlos con un valor de significancia de $\alpha = 0.05$ (Valor-P $< \alpha = 0.05$), los parámetros que individualmente son relevantes para el modelo ajustado son β_1 y β_3, debido a que el Valor-P de estos está por debajo del valor α seleccionado.

1.5. Interpretación de los parámetros

Los resultados y las variables de la sección 1.4 se interpretan como:

- $\hat{\beta}_1$: Este parámetro hace referencia a la duración promedio de la estadía de todos los pacientes en el hospital en días. Teniendo en cuenta lo anterior, podemos explicar que por cada día que aumenta (D), la probabilidad de infección aumenta en un 18.48 porciento, mientras las demás co-variables se mantengan fijas.
- $\hat{\beta}_3$: Este parámetro hace referencia al número promedio de camas en el hospital durante el periodo del estudio. Considerando esto, se puede interpretar que a medida que aumenta el número promedio de camas en el hospital (N), la probabilidad de infección aumenta en un 6.48 porciento, mientras las demás co-variables se mantengan fijas.

1.6. Calculo e interpretación del coeficiente de determinación múltiple R^2

De la información obtenida en la tabla ANOVA en el modelo ajustado, se puede utilizar la fórmula $R^2 = \frac{SSR}{SST} = \frac{SSR}{SSE+SSR}$ para hallar el coeficiente de determinación múltiple de la siguiente manera.

$$\mathbf{R^2} = \frac{SSR}{SSE+SSR} = \frac{61.3830}{52.0834+61.3830} = \mathbf{0.5409}$$

Dado que el coeficiente $\mathbf{R^2} = \mathbf{0.5409}$ se sabe que aproximadamente el 54.09 % de la variabilidad total es explicado por el modelo.

2. Pregunta 2

4 pt

2.1. Planteamiento prueba de hipótesis y modelo reducido

De la regresion se obtiene que los parametros β_1 y β_3 presentan un Valor-P menor a $\alpha = 0.05$, adicionalmente el siguiente parametro con menor Valor-P es β_5; el resto de parametros presentan un Valor-P muy alto. Por tanto se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_3 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 3, 5 \end{cases}$$

El modelo completo es el definido en la sección 1.1, y el modelo reducido es:

$$\text{MR: } \cancel{X.Bfat_i} = \beta_0 + \beta_2 R_i + \beta_4 C_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Se presenta la siguiente tabla del resumen de todas las regresiones, solo con la información necesaria para plantear el estadístico de prueba:

Tabla 4: Resumen de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	52.083	X1 X2 X3 X4 X5
Modelo reducido	95.008	X2 X4

con $X_1 = D_i$, $X_2 = R_i$, $X_3 = N_i$, $X_4 = C_i$ y $X_5 = E_i$.

2.2. Estadístico de prueba y conclusiones

2 pt

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_4) - SSE(\beta_0, \beta_1, \cdots, \beta_5)/3}{MSE(\beta_0, \beta_1, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,63} \\ &= \frac{(95.008 - 52.083)/3}{0.82672} = 17.30731 \end{aligned}$$

Ahora, comparando a un nivel de significancia $\alpha = 0.05$ F_0 con $f_{0.05,3,63} = 2.7505411$

Debido a que el valor F_0 es mayor que al $f_{0.05,3,63}$ entonces se rechaza H_0 y se concluye que el riesgo de infección depende de al menos una de la variables asociadas al modelo reducido planteado en la sección 2.1.

2 pt

3. Pregunta 3

5 pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

¿Las consecuencias que tiene la duración de la estadía (D) y el número de camas (N) sobre la probabilidad de infección son las mismas? De la misma manera, ¿Las consecuencias que tiene el número de enfermeras (E) y la rutina de cultivos (R) sobre la probabilidad de infección es igual?

Para analizar las preguntas que se plantean, se debe formular la siguiente prueba de hipótesis donde se analiza si el efecto de la duración de la estadía en la probabilidad de infección es igual al efecto del número de camas sobre la probabilidad de infección. Igualmente, se busca analizar si el efecto del número de enfermeras es el mismo que el efecto de la rutina de cultivos sobre la probabilidad de infección.

Se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_3,\ \beta_2 = \beta_5 \\ \mathrm{H}_1 : \text{Alguna de las desigualdades no se cumple} \end{cases}$$

Considerando que la hipótesis nula se puede ver como un sistema de dos ecuaciones se puede reescribir de la siguiente manera:

$$\begin{cases} \mathrm{H}_0 : \beta_1 - \beta_3 = 0,\ \beta_2 - \beta_5 = 0 \end{cases}$$

Las hipótesis propuestas también se pueden escribir en forma matricial como se muestra a continuación:

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

En donde $\mathbf{L}$ está dada por las siguientes combinaciones lineales:

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

El modelo reducido en este caso es:

$$\widehat{Riesgo}_i = \beta_0 + \beta_1(D+N)_i + \beta_2(R+E)_i + \beta_4 C_i$$

3.2. Estadistico de prueba

Estadístico de Prueba:

La expresión para calcular el estadístico de prueba F_0 es presentado a continuación en donde MR se refiere al modelo reducido y MF al modelo completo:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,63} = \frac{(SSE(MR) - 52.083)/2}{0.82672} \overset{H_0}{\sim} f_{2,63}$$

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Para la validación de este supuesto, se plantea la siguiente prueba de hipótesis, Shapiro-Wilk Test, acompañado del siguiente gráfico:

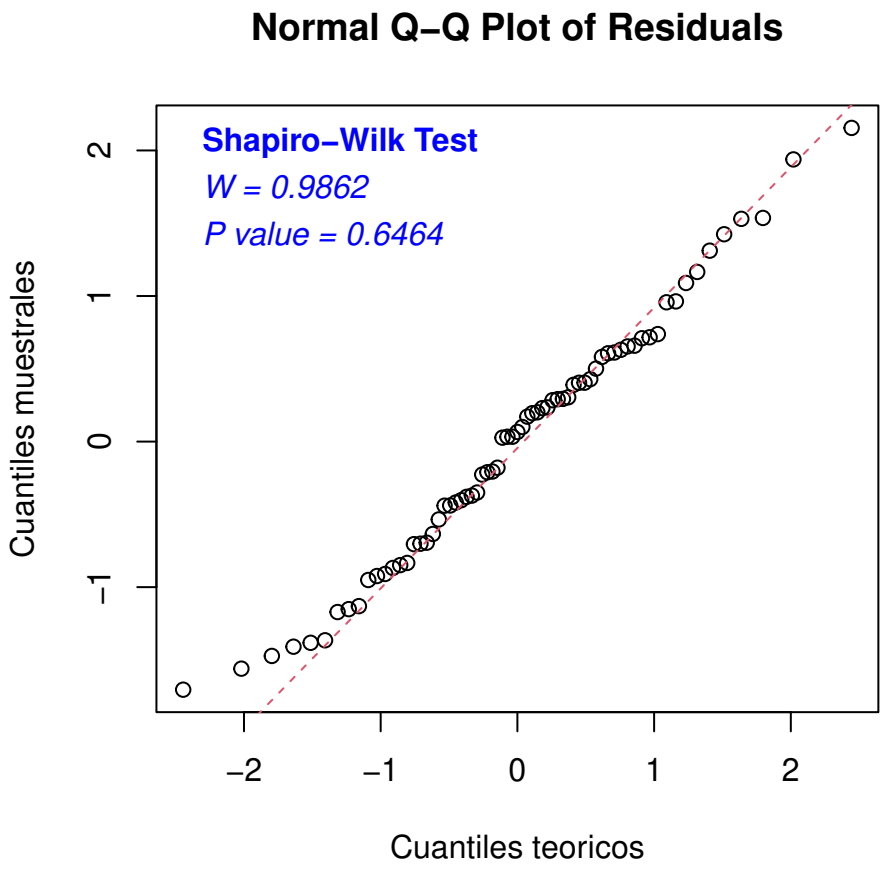

Figura 1: Gráfico cuantil-cuantil y normalidad de los residuales

Planteamos la prueba de hipótesis para probar el supuesto de noramalidad de los errores:

$$\begin{cases} \mathrm{H}_0 : \varepsilon \sim \mathcal{N}(0, \sigma^2) \\ \mathrm{H}_1 : \varepsilon \nsim \mathcal{N}(0, \sigma^2) \end{cases}$$

Se observa que el patrón de los residuales en la zona media sigue la línea roja que representa el ajuste de la distribución de los residuales a una distribución normal, en el extremo superior se aprecia un leve desajuste el cual no representa un problema. En el extremo inferior se puede ver el mayor desajuste de los residuales con la línea roja (no es un desajuste muy grave), aunque además de un buen ajuste en el gráfico, teniendo en cuenta que el valor Valor-P (>0.05) es grande, se debe aceptar H_0. Entonces, se concluye que los residuales del modelo tienen una distribución normal.

4.1.2. Media 0 y Varianza constante

Para verificar el supuesto de varianza constante, realizaremos el análisis de la gráfica de residuales estudentizados vs. valores ajustados del modelo:

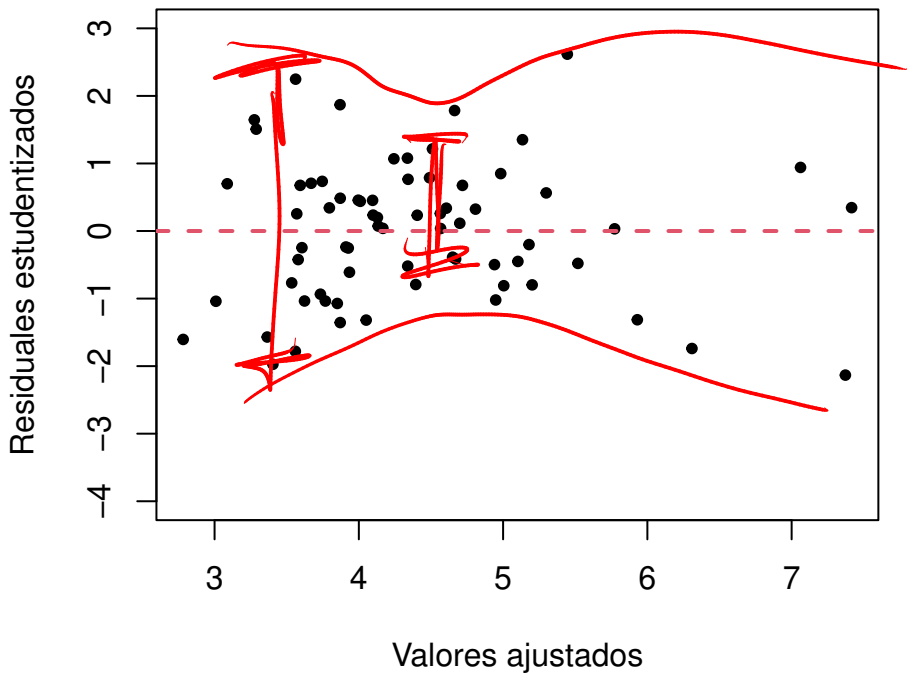

Figura 2: Gráfico residuales estudentizados vs valores ajustados

La línea roja indica que los residuales tienen media 0, asimismo se requiere probar que:

$$H_0 : V[\varepsilon_i] = \sigma^2 \quad \text{vs.} \quad V[\varepsilon_i] \neq \sigma^2$$

Se observa que puede simular un rectángulo con nube de puntos en la zona media de la gráfica lo cual indica que el supuesto se cumple, es decir que los residuales tienen varianza constante σ^2, aun así, en la zona derecha del grafico se observan valores extremos un poco alejados aunque estos valores podrían estar dentro del rectángulo simulado, es decir, el supuesto se cumple, pero existen valores extremos que lo afectan levemente.

4.2. Observaciones extremas

Para identificar los valores atípicos, de balanceo e influencias del modelo se debe utilizar la tabla de diagnóstico.

Las observaciones atípicas (outliers) son caracterizadas porque su valor en la respuesta Y está separada al resto de las observaciones, implicando afectaciones a los resultados del ajuste del modelo de regresión.

Tabla 5: Tabla de diagnostico

	res.stud	Cooks. D	$H_{ii.value}$	Dffits
6	0.3461	0.0049	0.1953	0.1705
10	2.6178	0.1216	0.1042	0.8930
15	0.9412	0.0511	0.2569	0.5534
17	0.0326	0.0000	0.1838	0.0155
23	-1.7390	0.1075	0.1804	-0.8160
38	1.8695	0.1049	0.1577	0.8090
46	-0.7678	0.0215	0.1788	-0.3583
47	-2.1331	0.4610	0.3911	-1.7094

4.2.1. Datos atípicos

Una observación se considera atípica cuando su residual estudentizado r_i es tal que $|r_{estud}| > 3$. Cuando se analiza la columna res.stud de la tabla 5, se puede notar que ninguna observación tiene un valor de r_i mayor a 3 por lo que se concluye que ninguno de los datos del modelo es atípico. Esta conclusión se puede confirmar en la siguiente gráfica (Residuales vs. Observaciones) puesto que ningún punto del modelo está por fuera de la región delimitada por las líneas rojas.

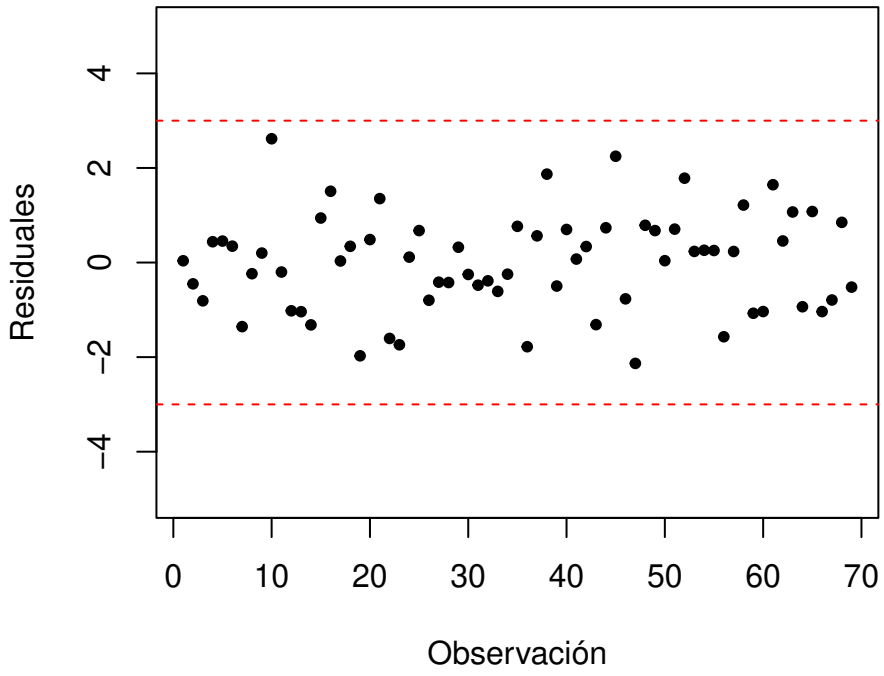

Figura 3: Identificación de datos atípicos

4.2.2. Puntos de balanceo

Los puntos de balanceo de un modelo implican que hay observaciones en las predictoras muy alejadas de las demás observaciones de la muestra. Aquí, se asume que la observación i es un punto de balanceo si $h_{ii} > 2\frac{p}{n}$. En este caso, se tiene como criterio que $h_{ii} > 2\frac{p}{n} = 2(\frac{6}{69}) = 0.1739$.

De la tabla de Diagnóstico, se pueden comparar los valores de la columna hii.value y se obtiene que los puntos 6, 15, 17, 23, 46, 47 son de balanceo.

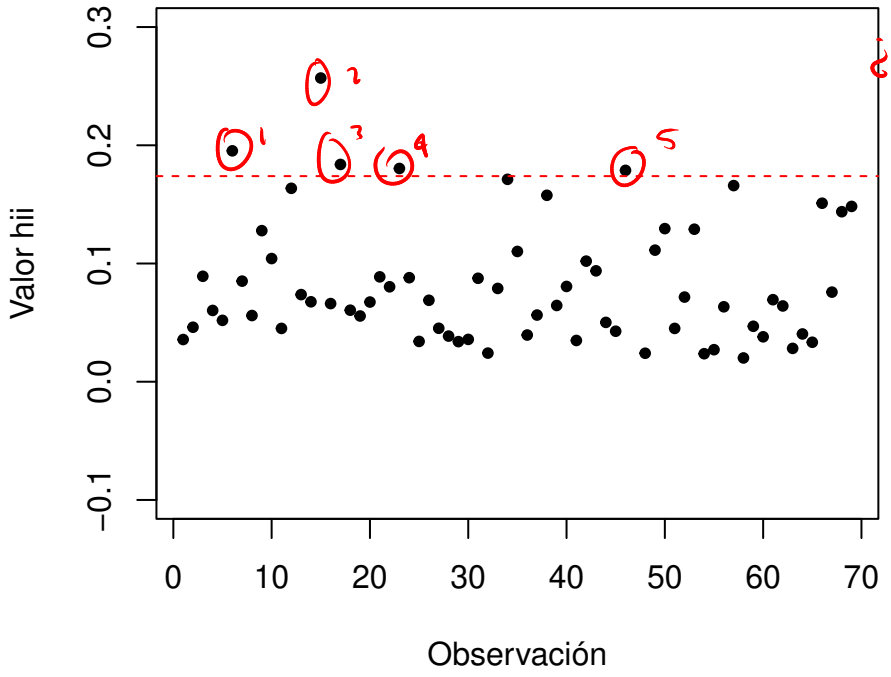

Figura 4: Identificación de puntos de balanceo

4.2.3. Puntos influenciales

Las observaciones influenciales tienen un impacto notable sobre los coeficientes de la regresión ajustada, son aquellos que halan el modelo y causan cambios importantes en la ecuación.

Para determinar cuáles observaciones son influenciables se usan los siguientes dos criterios:

1. La distancia de Cook.

Si el valor D_i es muy alto (mayor a 1), se puede decir que la observación i tiene influencia sobre el vector de parámetros estimados. De la tabla de diagnóstico, se concluye que ninguna observación del modelo es influencial, esto también se logra ver desde la siguiente gráfica.

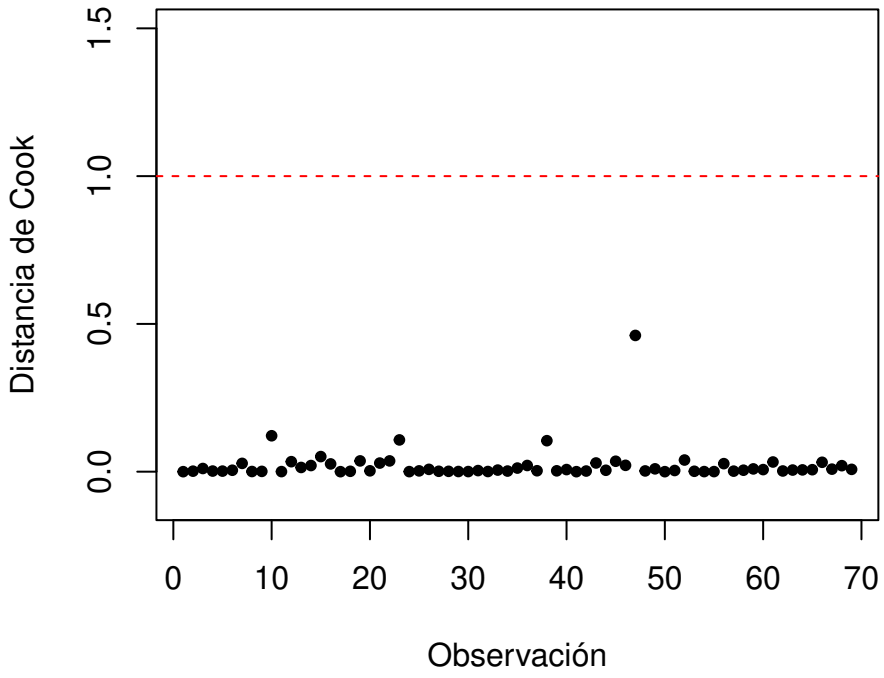

Figura 5: Criterio distancias de Cook para puntos influenciales

2. Diagrama DFFITS

Se refiere al número de desviaciones estándar que el valor ajustado y_i se mueve si la observación i es omitida.

Una observación es influencial si $|DFFITS_I| > 2\sqrt{\frac{p}{n}}$. En este caso, donde $p = 6$y $n = 69$ se tiene que $|DFFITS_I| > 2\sqrt{\frac{6}{69}} = 0.5897$. Comparando las observaciones de la columna DFFITS, se concluye que las observaciones 10, 38, 23, 47 son influenciales.

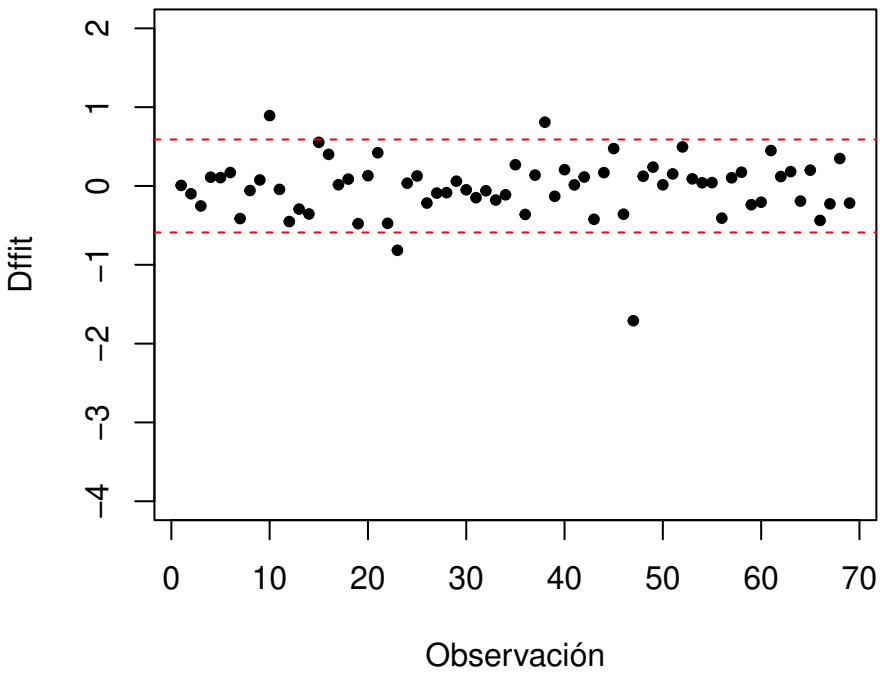

Figura 6: Criterio Dffits para puntos influenciales