

4,3

Trabajo 1

Equipo 19

Estudiantes

**Victoria Valencia Velez
Evamarina Pastor Ochoa
Jose Luis Suarez Ledesma
Jorge Alexander Palomá Villa**

Docente

Julieth Veronica Guarin Escudero

Asignatura

Estadística II



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medellín
30 de marzo de 2023

Índice

1. Glosario	3
2. Pregunta 1	6
2.1. Modelo de regresión ajustado	6
2.2. Significancia de la regresión	7
2.3. Significancia de los parámetros	7
2.4. Interpretación de los parámetros	8
2.5. Coeficiente de determinación múltiple R^2	8
3. Pregunta 2	9
3.1. Planteamiento pruebas de hipótesis y modelo reducido	9
3.2. Estadístico de prueba y conclusión	9
4. Pregunta 3	10
4.1. Prueba de hipótesis lineal general y prueba de hipótesis matricial	10
4.2. Estadístico de prueba	10
5. Pregunta 4	11
5.1. Supuestos del modelo	11
5.1.1. Normalidad de los residuales	11
5.1.2. Varianza constante	12
5.1.3. Media cero	12
5.2. Verificación de las observaciones	13
5.2.1. Datos atípicos	13
5.2.2. Puntos de balanceo	13
5.2.3. Puntos influyentes	15
5.3. Conclusión	16

Índice de figuras

1.	Gráfico cuantil-cuantil y normalidad de residuales	11
2.	Gráfico residuales estudentizados vs valores ajustados	12
3.	Identificación de datos atípicos	13
4.	Identificación de puntos de balanceo	14
5.	Criterio distancias de Cook para puntos influenciales	15
6.	Criterio Dffits para puntos influenciales	16

Índice de cuadros

1.	Tabla de valores coeficientes del modelo	6
2.	Tabla ANOVA para el modelo	7
3.	Resumen de los coeficientes	8
4.	Resumen tabla de todas las regresiones	9
5.	<i>hii</i> de los puntos de balanceo	14
6.	Diagnóstico de los residuos	16

1. Glosario

Modelo de Regresión Lineal Múltiple:

A diferencia de la regresión lineal simple, en este tipo de modelos se emplean más de una variable independiente que contribuyen a la predicción de la variable dependiente Y . El modelo estadístico para este tipo de regresión se puede escribir como una generalización del modelo lineal simple para cierto número de covariables y es posible expresarlo de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i, \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2), 1 \leq i \leq 65$$

esto no es general es n

Se trata de un modelo de primer orden donde los Betas son los parámetros (p) del modelo que indican el cambio en la respuesta media de Y por unidad de incremento en la respectiva variable X ; uno por cada variable predictora considerada más uno por el intercepto ($p = k+1$), los son los valores en la i -ésima observación muestral de las k variables predictoras y un error aleatorio con distribución $N(0, \sigma^2)$ ✓

Para la estimación de los parámetros del modelo se recurre al método de mínimos cuadrados. ✓

Prueba de hipótesis generalizada y matricial:

Es una extensión de la suma de cuadrados extras, que permite hacer pruebas de hipótesis con unas restricciones adicionales, permite saber si el efecto que tiene X_j sobre la variable respuesta, es el mismo efecto que tiene cualquier X_k sobre la variable respuesta. Se plantea a partir una matriz que tiene filas linealmente independientes. La cual al multiplicar por el vector B da un sistema de ecuaciones donde se formulan hipótesis sobre los parámetros del modelo. β

más bien sería es un caso particular

es más que eso

puede hacerse individual y usar prueba t

Prueba de significancia:

Este método consiste en realizar un análisis de varianza similar al empleado en RLS para probar la significancia de la regresión. Se establece un juego de hipótesis H_0 y H_1 que mediante la tabla ANOVA se obtiene el estadístico F_0 .

A una significancia α dada, $F_0 > f_{\alpha, k, n-p}$ se rechaza la hipótesis nula H_0 y se prueba la existencia de una relación de regresión pero, si $F_0 < f_{\alpha, k, n-p}$ no es posible rechazarla.

De forma equivalente, al definir un P-valor para la prueba como $vp = P(f_{k, n-p} > F_0)$, se puede rechazar H_0 si $vp < \alpha$.

Coefficiente de determinación múltiple:

Denotado por R^2 , se define como la suma de los cuadrados residuales entre la suma de los cuadrados totales y mide la proporción de la varianza total de la variable explicada por la regresión.

Al ser una proporción, la cantidad varía entre 0 y 1, pero también puede ser escrita en forma de porcentaje.

Un R^2 igual a cero implica que todos los coeficientes de regresión ajustados son iguales a cero.

Un valor igual a 1, implica que todas las observaciones caen sobre la superficie de regresión ajustada.

Aunque el coeficiente es empleado como una medida de bondad del ajuste de la función de regresión, valores grandes no necesariamente significa que la superficie ajustada es útil. En general, este coeficiente se puede interpretar como el porcentaje de la variabilidad total de Y que es explicada por el modelo ajustado.

Modelo reducido:

En ocasiones se requiere analizar la significancia de un subconjunto de variables dentro del conjunto de las variables predictoras, en este caso se define el subconjunto de las variables de análisis y el subconjunto de las otras variables presentes en el modelo inicial, donde la unión de los subconjuntos forma parte del conjunto total.

La ecuación del modelo reducido puede plantearse de la misma forma que el modelo inicial, no obstante, ahora son descartados de la ecuación, los términos correspondientes al subconjunto de análisis.

Debido a la hipótesis \rightarrow MR en forma más general es MF bajo H_0
Una vez planteado el modelo, se puede llevar a cabo la prueba de significancia mediante la construcción del estadístico de prueba F_0 como el cociente de la suma de cuadrados extra de los parámetros de análisis entre los grados de libertad del modelo reducido y el error cuadrático medio del modelo completo.

Se sigue el mismo criterio de rechazo que en la prueba del modelo de regresión inicial, con la diferencia de que un resultado significativo, no implica que el subconjunto sea contribuyente y un resultado negativo, tampoco es suficiente para descartarlo. \rightarrow En general S.

Validación de los supuestos en los errores:

En los modelos de regresión se debe cumplir que los errores distribuyan normal, tengan media cero, varianza constante y sean mutuamente independientes.

Para el supuesto de normalidad pueden emplearse tanto métodos analíticos como el Shapiro-test o gráficos, no obstante, la veracidad de los métodos analíticos dependerá también del tamaño de la muestra. Por otra parte, para el supuesto de varianza constante, se puede emplear un gráfico de estudentizados vs ajustados.

\rightarrow Residuales studentizados

Gráfico cuantil – cuantil y normalidad de los residuales:

La interpretación de este tipo de gráficos permite comparar la distribución de los residuos con la distribución normal teórica, ya que el modelo de regresión lineal supone que los residuos siguen una distribución normal.

En este gráfico tendremos la diagonal de los valores que predice el modelo, los puntos como valores observados y la diferencia entre estos y la diagonal como los errores o residuos. Es posible encontrar en un caso ideal que los puntos se distribuyan sobre la diagonal o en su defecto, que se aparten de esta.

Si se observan patrones que no corresponden a una línea recta, pueden interpretarse como una falta de normalidad, pero también puede deberse a otros factores que estén alterando el modelo, de forma general es posible encontrar asimetrías en la distribución como colas largas y cortas, outliers o puntos alejados de la recta y cambios en la pendiente.

Gráficos residuales estudentizados vs valores ajustados:

En este gráfico idealmente se deber encontrar un patrón de residuos al azar, la presencia de tendencias, dispersiones irregulares y outliers, compromete la validez del modelo.

Verificación de observaciones Datos atípicos:

Resulta importante corroborar la presencia de datos atípicos (alejados de Y), para ello se define una observación atípica como aquella cuyo residual estudentizado presenta una magnitud superior a 3. Se puede acudir a un gráfico de distancias de Cook para esta observación.

al revés!!! $2 \frac{p}{n} < 1$

Puntos de balanceo:

Este tipo de puntos son aquellos que se alejan de las variables predictoras, no causan gran impacto en los coeficientes estimados de la regresión, pero sí el R^2 y los errores estándar de los coeficientes. Pueden ser detectados mediante las distancias hii. Estas cantidades proporcionan una medida estandarizada de la distancia de la i -ésima observación al centro del espacio definido por las predictoras. Para valores de $2p/n > 1$ se asume una observación i como punto de balanceo si se cumple el criterio $h_{ii} > 2p/n$, donde las h_{ii} siempre son menores a 1. Observaciones con h_{ii} grandes posiblemente sean también puntos influyentes.

Puntos influyentes:

Estos puntos poseen valores inusuales tanto en el espacio de las predictoras como en la respuesta, además el impacto sobre los coeficientes es notorio, es decir, son observaciones que al ser obviadas del modelo pueden modificar la pendiente de la regresión, generar tendencias en la varianza y en general, lo afectan drásticamente y por tanto no se pueden descartar.

Para caracterizarlos se puede analizar el modelo mediante el gráfico de distancia de cook y el diagnóstico DFFITS.

También afectan los y_i

Distancia de Cook:

Medida de la distancia cuadrática entre el estimador beta por mínimos cuadrados basado en las n observaciones y el estimador beta obtenido eliminando la i -ésima observación. De este planteamiento, se tiene el vector de parámetros estimados cuando no se

admite la observación i en el ajuste y cuando sí, de tal forma, que una Distancia de cook i -ésima alta corresponde a una influencia sobre el vector de parámetros. Luego, se define y se establece que si la i -ésima observación es influencia.

Diagnóstico DFFITS:

Número de desviaciones estándar que el valor ajustado i -ésimo se mueve si la observación i se omite. Se establece el criterio de punto influyente si $|DFFITS| > 2\sqrt{p/n}$

Nota sobre la verificación de puntos influyentes:

En modelos que presenten datos ~~atípicos~~, resulta importante corroborar que no sean puntos influyentes, pues de esto depende que se puedan omitir del modelo en un proceso de selección de datos. \rightarrow No sólo los influyentes

2. Pregunta 1 18,5

Para estudiar la eficacia del control de infecciones hospitalarias en Estados Unidos, se analiza una base de datos de 65 hospitales. A partir de esta se propone que el riesgo a infectarse se puede modelar como una regresión lineal múltiple donde la variable respuesta Y es el Riesgo de Infección a enfermedades hospitalarias y las variables predictoras son: X_1 Duración de la estadía, X_2 rutina de cultivo, X_3 Número de camas, X_4 censo promedio diario, X_5 Número de enfermeras. El modelo de regresión tiene la siguiente forma:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2); \quad 1 \leq i \leq 65$$

2.1. Modelo de regresión ajustado 3pt

Al realizar el ajuste del modelo con R, se obtienen los siguientes valores de los coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	0.1088
β_1	0.2052
β_2	0.0112
β_3	0.0408
β_4	0.0075
β_5	0.0015

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 0.1088 + 0.2052X_{1i} + 0.0112X_{2i} + 0.0408X_{3i} + 0.0075X_{4i} + 0.0015X_{5i}; \quad 1 \leq i \leq 65$$

2.2. Significancia de la regresión

5 pt

Para evaluar la significancia de la regresión, se plantea la siguiente prueba de hipótesis:

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

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} = \frac{10.022}{0.958} = 10.457 \stackrel{H_0}{\sim} f_{5,59} \quad (1)$$

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	50.1124	5	10.022472	10.4567	3.23816e-07
Error	56.5501	59	0.958476		

Se obtiene un valor P aproximadamente igual a 0, por lo tanto se rechaza la hipótesis nula, se concluye que el modelo de regresión es significativo y que por lo menos un coeficiente es distinto cero, para lo cual se procede a verificar la significancia individual de los parámetros.

2.3. Significancia de los parámetros

6 pt

Para evaluar la significancia de los parámetros, se plantea las siguientes pruebas de hipótesis:

$$\begin{cases} H_0 : \beta_j = 0 \text{ para } j = 0, 1, 2, \dots, 5 \\ H_a : \beta_j \neq 0 \text{ para } j = 0, 1, 2, \dots, 5 \end{cases}$$

A partir del análisis de datos realizado con R, se obtiene la siguiente tabla:

¿Estadístico de prueba $T_{0,j}$?

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	0.1088	1.6462	0.0661	0.9475
β_1	0.2052	0.0783	2.6195	0.0112
β_2	0.0112	0.0293	0.3814	0.7043
β_3	0.0408	0.0136	2.9893	0.0041
β_4	0.0075	0.0081	0.9296	0.3564
β_5	0.0015	0.0007	2.2035	0.0315

Los valores P que se presentan en la tabla de resultados indican que los coeficientes de regresión β_1 , β_3 , y β_5 son estadísticamente significativos a un nivel de confianza del 95 %, es decir, con un nivel de significancia $\alpha = 0.05$. Esto se debe a que los valores P correspondientes a estos coeficientes son menores que α , con valores de 0.0112, 0.0041, y 0.0315, respectivamente. ✓

2.4. Interpretación de los parámetros 1,5 pt

$\hat{\beta}_1$: Por cada día más de duración promedio de estadía cuando el número promedio de camas y el número promedio de enfermeras es constante, la probabilidad promedio estimada de adquirir una infección en el hospital aumenta un 20.52 % ✓

$\hat{\beta}_3$: Por cada unidad adicional en el número promedio de camas en el hospital, manteniendo constante la duración promedio de estadía y el número promedio de enfermeras, la probabilidad promedio estimada de adquirir una infección aumenta en un 4.08 %.

$\hat{\beta}_5$: Por cada unidad adicional en la cantidad promedio de enfermeras en el hospital, manteniendo constante la duración promedio de estadía y el número promedio de camas, la probabilidad promedio estimada de adquirir una infección aumenta en un 0.15 %.

2.5. Coeficiente de determinación múltiple R^2 3 pt

El cálculo del coeficiente de determinación múltiple se realizó de la siguiente manera:

$$R^2 = \frac{SSR}{SSE + SSR} = \frac{50.1124}{56.5501 + 50.1124} = 0.4698 \quad (2) \quad \checkmark$$

El modelo de regresión ajustado muestra un coeficiente de determinación múltiple de $R^2 = 0.4698$, lo que indica que el modelo es capaz de explicar aproximadamente el 46.98 de la variabilidad total observada en el riesgo de infección. En otras palabras, el 53.02 de la variabilidad total en el riesgo de infección no se explica por el modelo y puede ser atribuible a otros factores que no están incluidos en el modelo de regresión. ✓

al error

3. Pregunta 2 4 p +

3.1. Planteamiento pruebas de hipótesis y modelo reducido

Las variables con el valor P más alto en el modelo fueron X_2 , X_4 , X_5 , por lo tanto, para verificar si es posible descartarlas del modelo y con ayuda de la tabla de todas las regresiones posibles se pretende evaluar la siguiente prueba de hipótesis:

$$\begin{cases} H_0 : \beta_2 = \beta_4 = \beta_5 = 0 \\ H_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 2, 4, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo				
Modelo completo	56.550	X1	X2	X3	X4	X5
Modelo reducido	61.830	X1	X3			

El modelo reducido para la prueba de significancia del subconjunto es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i}, \quad 1 \leq i \leq 65$$

3.2. Estadístico de prueba y conclusión 2 p +

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_3) - SSE(\beta_0, \dots, \beta_5))/3}{MSE(\beta_0, \dots, \beta_5)} \stackrel{H_0}{\sim} F_{3,59} \\ &= \frac{(61.830 - 56.550)/3}{0.9584} \\ &= 1.8363 \end{aligned} \quad (3)$$

Ahora, comparando el F_0 con $f_{0.95,3,59} = 2.7608$, se puede ver que $1.8363 < 2.7608$ y no se rechaza la hipótesis nula, por lo tanto en conjunto las variables no son significativas y es posible descartar del modelo las variables del subconjunto.

4. Pregunta 3 3 p+

4.1. Prueba de hipótesis lineal general y prueba de hipótesis matricial

A partir de la siguiente pregunta: ¿el valor de la variable β_2 es equivalente 4 veces el valor de la variable β_3 y el valor de β_4 es equivalente al valor de β_5 ? Se plantea la siguiente prueba de hipótesis:

$$\begin{cases} H_0 : \beta_2 = 4\beta_3; \beta_4 = \beta_5 \\ H_1 : \beta_2 \neq 4\beta_3; \beta_4 \neq \beta_5 \end{cases}$$

¿? β_2 es un parámetro $\beta_2 \neq 4\beta_3 \vee \beta_4 \neq \beta_5$ es alguna o ambas, no ambas al mismo tiempo

Se puede notar que la prueba de hipótesis constituye un sistema de ecuaciones 2×6 , donde las variables con coeficientes diferentes de cero son β_2, β_3 y β_4, β_5 , de manera que se puede reescribir de forma matricial, así:

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

✓ 1,5 p+

Con \mathbf{L} dada por

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & 1 & -4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

✓

El modelo reducido está dado por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 (X_{2i} + 4X_{3i}) + \beta_4 (X_{4i} + X_{5i}) + \varepsilon_i, \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2); 1 \leq i \leq 65$$

X 0 p+

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 (X_{23i}^*) + \beta_4 (X_{45i}^*) + \varepsilon_i, \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2); 1 \leq i \leq 65$$

X

Con $X_{23i}^* = X_{2i} + 4X_{3i}$ y $X_{45i}^* = X_{4i} + X_{5i}$

X $Y_i = \beta_0 + 4\beta_3 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_4 X_{5i} + \varepsilon_i$
 $= \beta_0 + \beta_3 (4X_{2i} + X_{3i}) + \beta_4 (X_{4i} + X_{5i}) + \varepsilon_i$

4.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \stackrel{H_0}{\sim} f_{2,59}$$

✓ 1,5 p+

Dado que no se solicita calcularlo, se deja denotado recordando que el criterio de rechazo de la hipótesis nula es cuando el estadístico de prueba $F_0 > f_{2,59}$ con $f_{2,59} = 3.1531$ o cuando el valor P es menor que la significancia.

✓

Reemplazar lo que conocen

5. Pregunta 4 16,5

5.1. Supuestos del modelo

Los supuestos del modelo a evaluar son los siguientes:

- Los errores distribuyen normal
- Los errores tienen media cero
- Los errores tienen varianza constante σ^2

Se asume que los errores son independientes entre sí y estos



5.1.1. Normalidad de los residuales

4p+

Para la validación del supuesto de normalidad se evalúa la siguiente prueba de hipótesis ~~shapiro-wilk~~, junto con el gráfico cuantil-cuantil:

esto no es Shapiro wilk, es una prueba, no la formulación de hipótesis en sí

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


Gráfico cuantil-cuantil

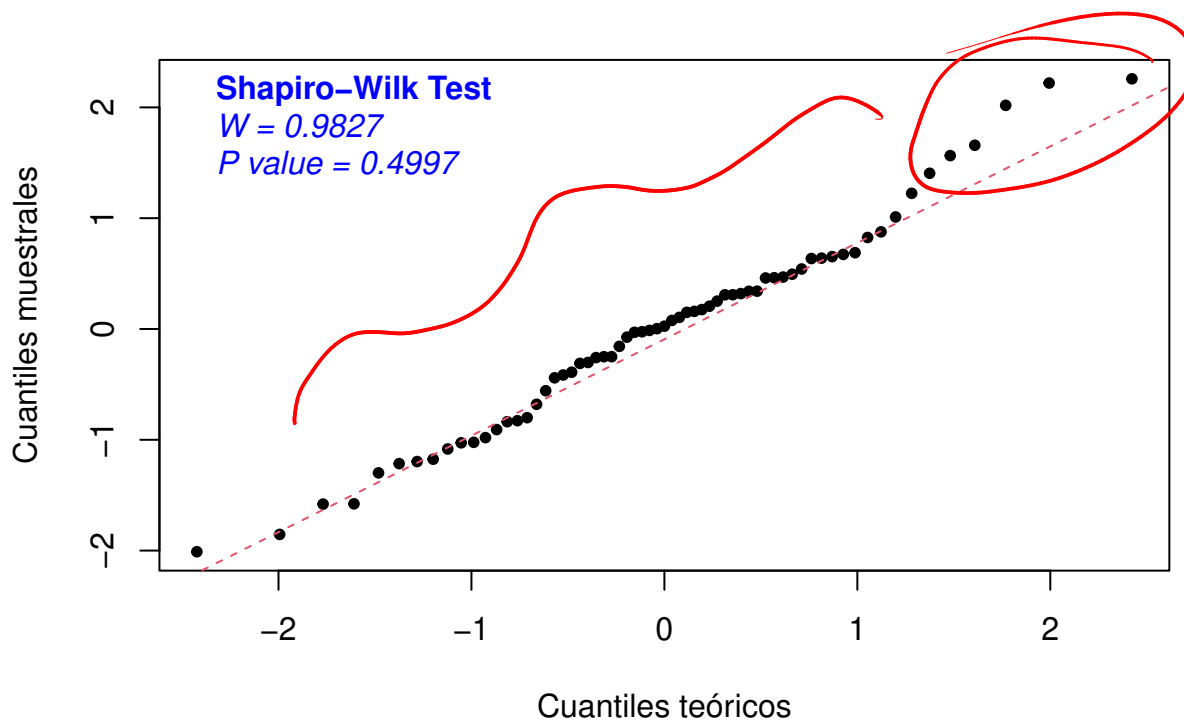


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

las que lo afectan
son las de
balanceo

A pesar de obtener un valor P de 0.4997 mayor a 0.05, por criterio gráfico, no se cumple el supuesto de normalidad, observe que se tienen patrones irregulares y la cola derecha pesada, se presume que esto pueda ocurrir debido a la presencia de observaciones influyentes. Este al ser un criterio más fidedigno, se rechaza hipótesis nula y se concluye que los errores no distribuyen normal. ✓

5.1.2. Varianza constante

Les faltó terminar de concluir con VP, pero se explicaron en glosario así que no les baja

Para validar el supuesto de varianza constante se procede a graficar los residuales estudentizados con respecto a los valores ajustados:

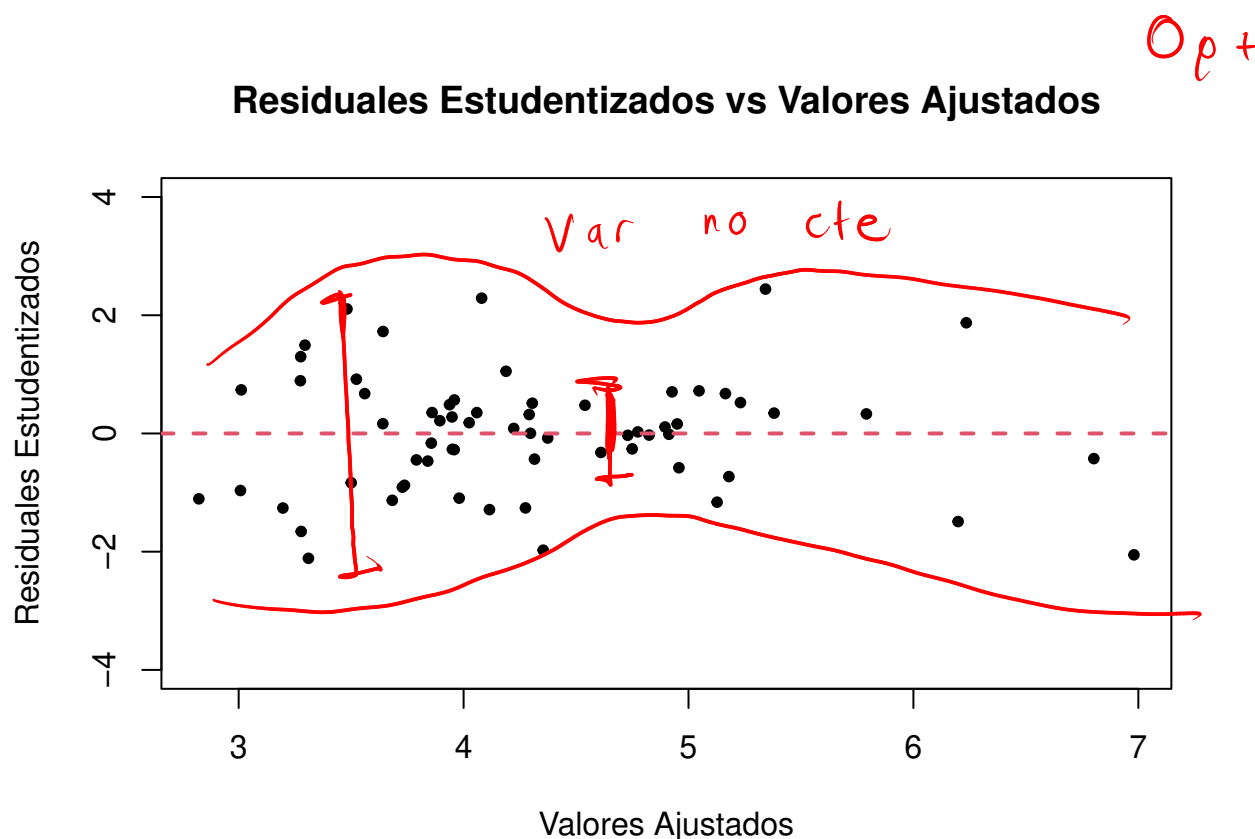


Figura 2: Gráfico residuales estudentizados vs valores ajustados

El gráfico no se logra observar algún patrón de decrecimiento, crecimiento, cuadrático, etc. Por lo tanto se asume que el modelo cumple con el supuesto de varianza constante. ✗

5.1.3. Media cero

it extra por hacer esto y haberlo hecho con los crudos

Para evaluar dicho supuesto se calcula el promedio de los residuales crudos con R y se obtiene un valor de -5×10^{-17} , lo cual confirma el supuesto de media cero. ✓

5.2. Verificación de las observaciones

5.2.1. Datos atípicos

2,5 p +

Para la identificación de datos atípicos se realiza el gráfico de los residuales estudentizados de cada observación, así:

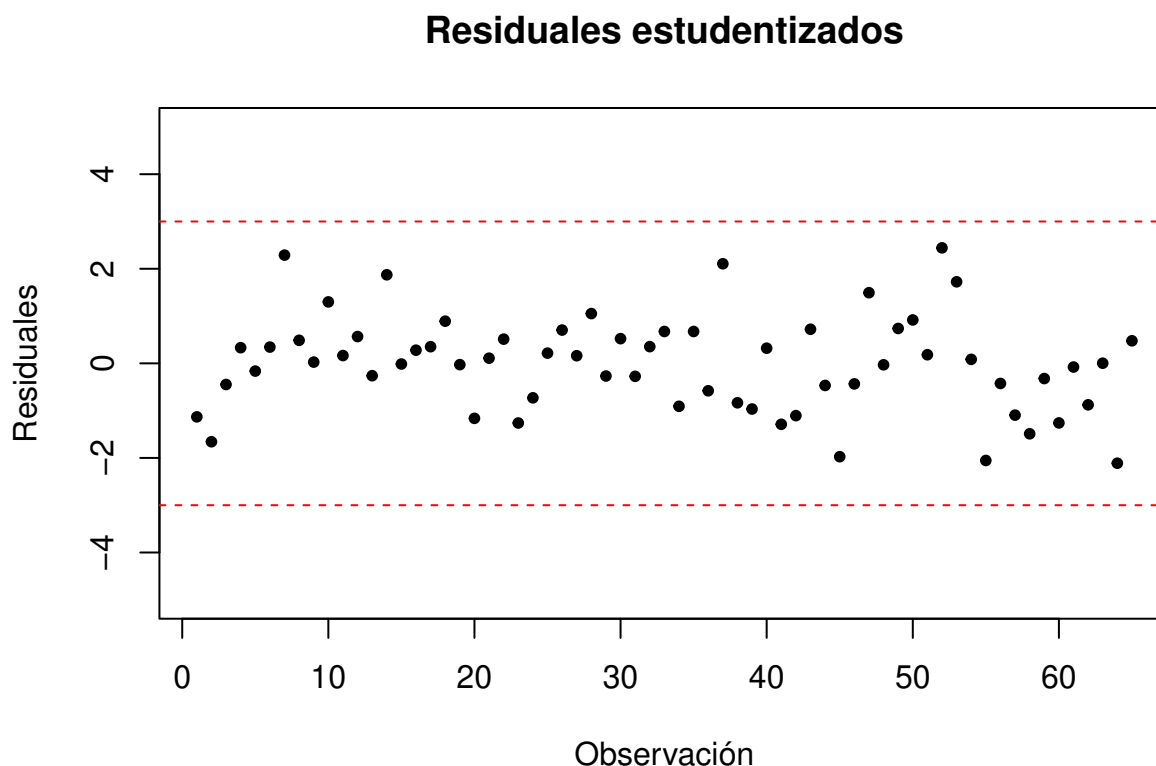


Figura 3: Identificación de datos atípicos

Como se puede observar en la gráfica de distancias de Cook, no hay datos atípicos en la muestra, pues ningún residual estudentizado sobrepasa el criterio de $|r_{estud}| > 3$. Es decir que la distribución de los residuales estudentizados del modelo oscilan en un intervalo de -3 a 3 . ✓

5.2.2. Puntos de balanceo

3 p +

Para identificar los puntos de balanceo, se hace una gráfica de las distancias h_{ii} , las cuales son las distancias de las observaciones al centro del espacio definido por las predictoras. ✓

Gráfica de hii para las observaciones

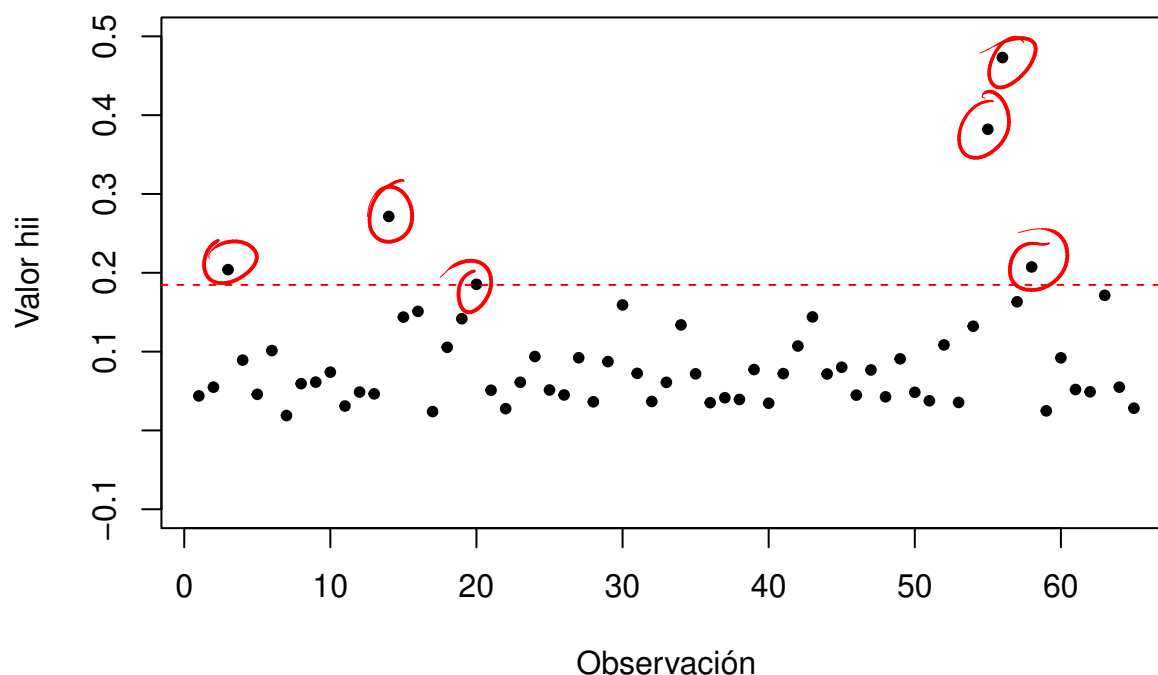


Figura 4: Identificación de puntos de balanceo

De la gráfica se pueden identificar 6 puntos de balanceo, donde el criterio hii es $hii > 2p/n$ que para el modelo es 0.1846 donde las 3, 14, 20, 55, 56, 58 tienen un valor mayor, donde 14, 55 y 58 se encuentran más alejados, como se observa en la siguiente tabla:

Cuadro 5: hii de los puntos de balanceo

	hii
3	0.2040
14	0.2715
20	0.1854
55	0.3821
56	0.4731
58	0.2074

Les bajaría por no decir lo que afectan los puntos de balanceo pero lo hicieron en glosario.

5.2.3. Puntos influenciales

Para identificar las observaciones influenciables, se emplea el uso de una gráfica de distancias de Cook (movimiento del vector de coeficientes) y una de observaciones vs Df-fits (movimiento del vector de la variable respuesta).

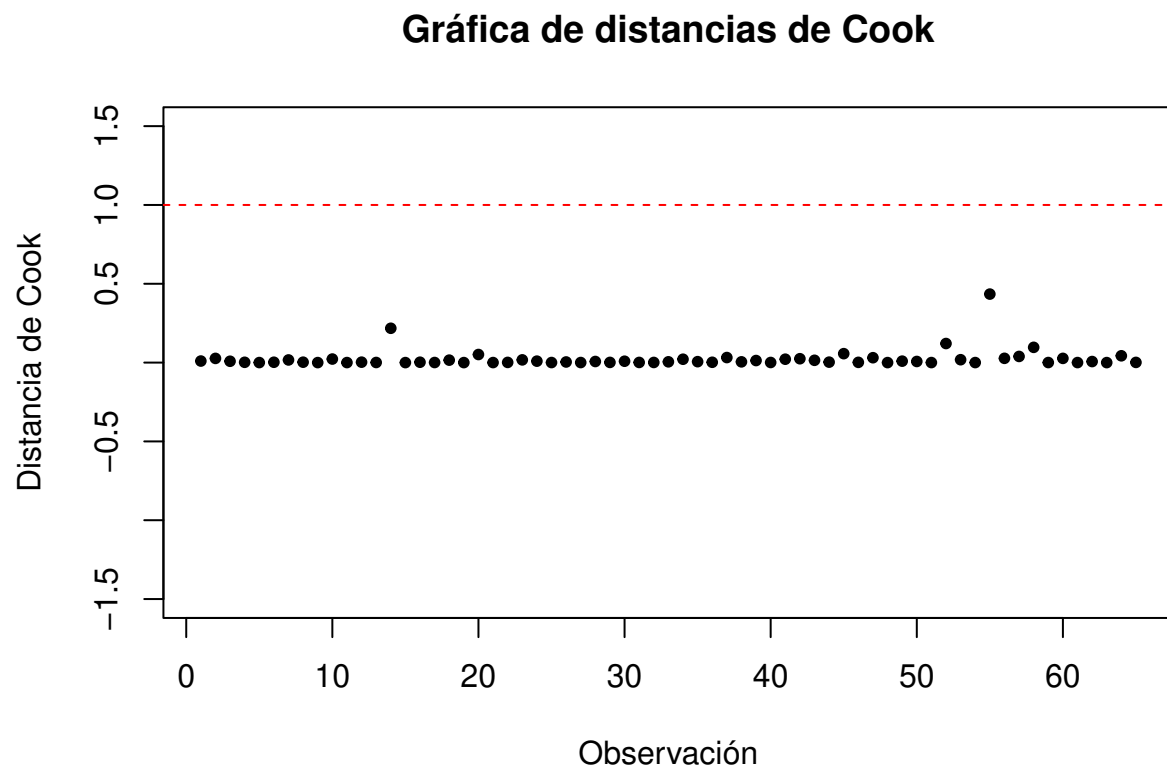


Figura 5: Criterio distancias de Cook para puntos influenciales

De acuerdo a la distancia de cook que es $D_i > 1$, no se encontraron puntos influenciales. Luego al aplicar el criterio Dffits:

✓

2 pt

Gráfica de observaciones vs Dffits

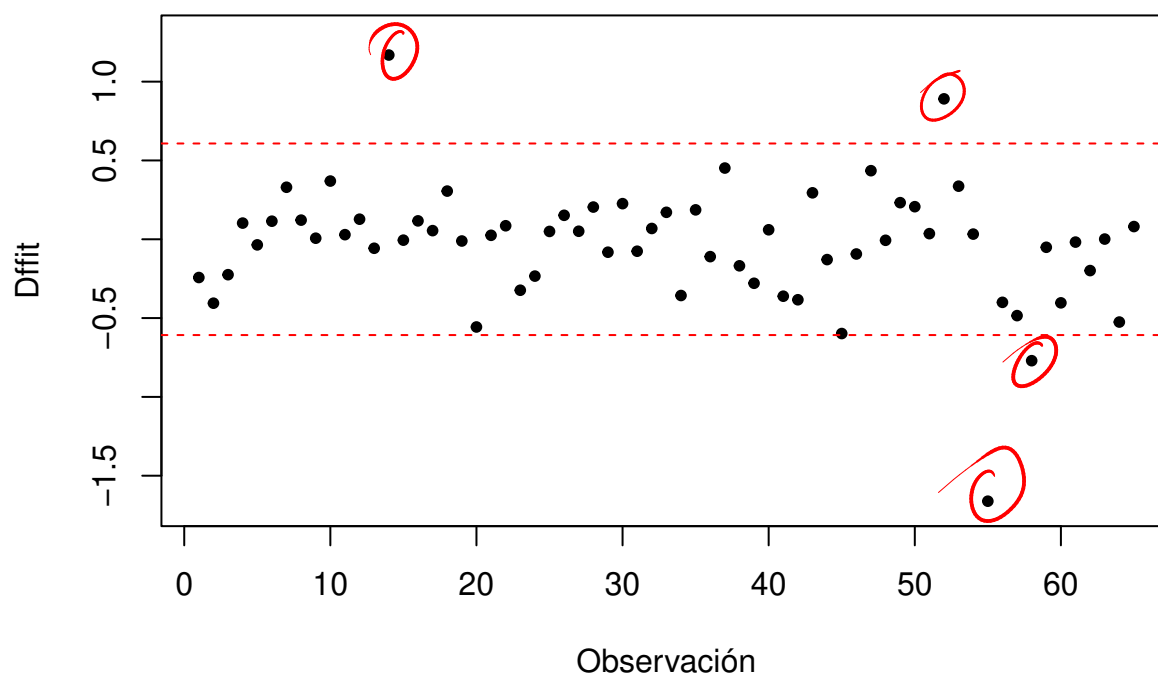


Figura 6: Criterio Dffits para puntos influenciales

Se obtienen cuatro ~~posibles~~ ^{7 lo son} puntos influenciales, 14, 52, 55 y 58, los cuales cumplen con el criterio $|DFFIST| > 2\sqrt{p/n}$ con un valor de 0.6076, como se observa en la tabla:


Cuadro 6: Diagnóstico de los residuos


	res.stud	Cooks.D	hii.value	Dffits
14	1.8726	0.2178	0.2715	1.1687
52	2.4423	0.1209	0.1085	0.8908
55	-2.0537	0.4346	0.3821	-1.6616
58	-1.4899	0.0968	0.2074	-0.7703

5.3. Conclusión

3 pt

El modelo de regresión lineal múltiple planteado no es ~~completamente~~ válido pues no cumple con el supuesto de normalidad de los errores, a pesar de que la prueba de hipótesis Shapiro-Wilk indicara lo contrario, dicha prueba no es adecuada pues el tamaño de la muestra

es mayor a 50 datos y el gráfico cuantil-cuantil es  más fuerte, pues cuenta con suficiente evidencia gráfica para rechazar la hipótesis nula.

Las observaciones 14, 52, 55 y 58 coinciden como posibles puntos de balanceo e influyentes, es decir, que afectan tanto a las variables predictoras, como a la variable respuesta y no se pueden descartar del modelo  ¿por qué no?