Estadística II - Taller 03 **Semestre: 2021-02**

Profesores: Carlos M. Lopera-Gómez y Raúl Alberto Pérez

Monitor: Simon Pedro Galeano

Este taller se divide en dos secciones, en la primera se trabajará lo relacionado a la validación del modelo. Posterior a esto, se considera un ejercicio en el que se realiza la prueba de falta de ajuste a un modelo.

En primer lugar considere el siguiente conjunto de datos.

Cuadro 1: Presentación de los datos

y	x
36.460743	5.878002
6.075999	-0.624800
47.402151	-3.727324
62.778036	-4.232765
28.971238	-2.318757
6.675766	3.712115

El día de hoy, la misión será realizar los siguientes ejercicios, claro está, haciendo uso de `R`.

1. Genere la base de datos que se muestra previamente usando el siguiente código.

```
gen_dat <- function(n, seed = 7) {
  varianza <- 16
  set.seed(seed)
  x <- runif(n=n, min=-5, max=6)
  media <- 4 - 6 * x + 2 * x^2
  set.seed(seed^2)
  y <- rnorm(n=n, mean=media, sd=sqrt(varianza))
  marco_datos <- data.frame(y=y, x=x)
  return(marco_datos)
}

datos <- gen_dat(75)
```

2. Ajuste el modelo de regresión lineal simple

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2);\ 1 \leq i \leq 75$$

3. Determine que parámetros son significativos y cuales no en el modelo, hágalo de manera rápida aprovechando alguna de las funciones de `R` usadas hasta el momento.

4. Extraiga los residuales del modelo y verifique que estos tengan media igual a 0, dé un argumento de por qué este supuesto siempre se cumple.
5. Determine si los residuales tienen varianza constante, argumente por qué esto es o no es así, además, si nota algún patrón o algo que considere anormal, coméntelo.
6. Evalúe el supuesto de normalidad de los residuales, hágalo usando un histograma, un gráfico cuantil - cuantil y finalmente una prueba de hipótesis.
7. Finalmente verifique si los residuales son o no independientes, hágalo de manera gráfica. Los valores de las variables están ingresados en la base de datos por orden cronológico.
8. Con la base de datos table.b3 del paquete `MPV`, realice la prueba de falta de ajuste, del modelo

$$y_i = \beta_0 + \beta_1 x4_i + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2);\ 1 \leq i \leq 32$$

para ello use la función `rsm` del paquete `rsm`.

Nota: se propone como ejercicio realizar la validación del modelo.

Solución

Ejercicio 1

Corriendo el código que se deja en el enunciado es lo único que se debe hacer para la realización de este ejercicio.

Ejercicio 2

Recuerde usar la función `lm` para el ajuste del modelo.

Ejercicio 3

La respuesta a este ejercicio es que ambos parámetros (β_0 y β_1) son significativos, esto se puede verificar rápidamente con la función `summary`.

Ejercicio 4

Los residuales se pueden extraer de varias maneras, sin embargo se sugiere el uso de la función `residuals`.

Se da un argumento de por qué la media de los residuales siempre es cero.

Recuerde $S(\beta_0,\ \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$.

Derivando parcialmente respecto a β_0 se sigue que.

$$\begin{aligned}
\partial_{\beta_0} S(\beta_0,\ \beta_1) &= \partial_{\beta_0} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \\
&= \sum_{i=1}^{n} \partial_{\beta_0}(y_i - \beta_0 - \beta_1 x_i)^2 \\
&= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)
\end{aligned}$$

Luego, igualando a cero la anterior expresión se tiene que

$$
\begin{aligned}
0 &= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \\
&= \sum_{i=1}^{n} \underbrace{(y_i - \beta_0 - \beta_1 x_i)}_{e_i} \\
&= \sum_{i=1}^{n} e_i \\
&= \sum_{i=1}^{n} \frac{e_i}{n} \quad \text{¡Esta es la media de los residuales!}
\end{aligned}
$$

Dado que para la estimación de los parámetros se requiere que la suma de los residuales sea nula, se tiene de inmediato que su promedio también será nulo.

Ejercicio 5

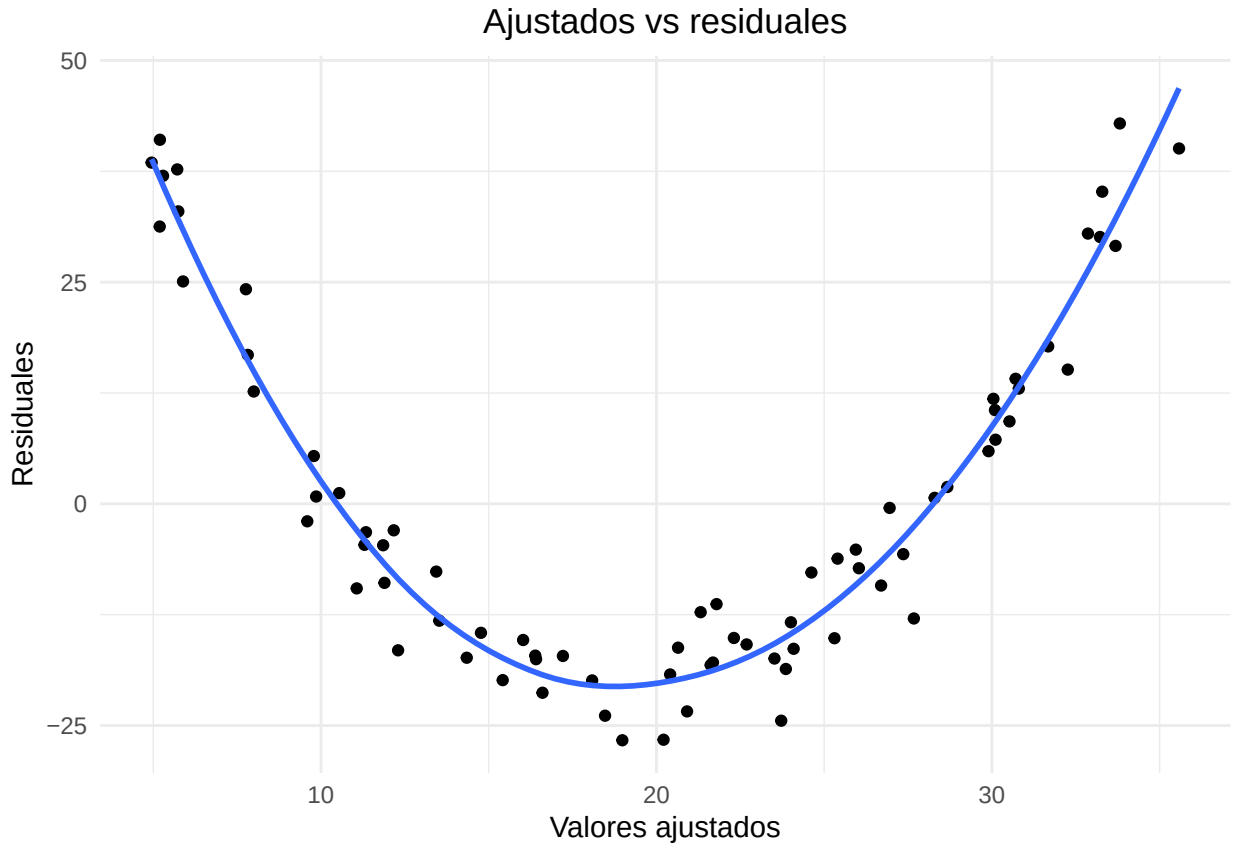

Figura 1: Verificación del supuesto de homocedasticidad

Como se puede ver en la figura anterior, se viola notoriamente el supuesto de varianza constante, además, se tiene una posible no linealidad debido a la forma en "U" de los

residuales cuando son graficados contra los valores ajustados.

Ejercicio 6

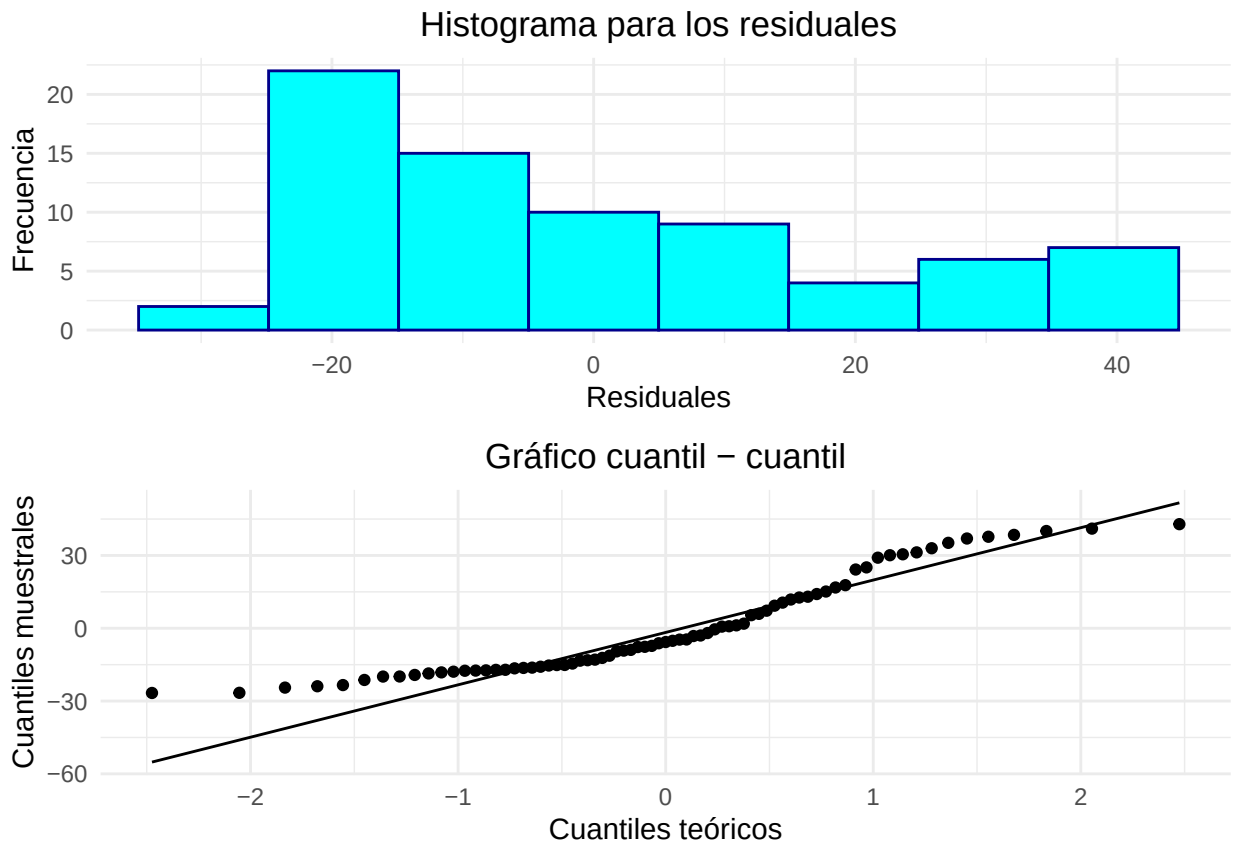

Figura 2: Verificación del supuesto de normalidad

Tanto del histograma como del gráfico cuantil - cuantil se puede empieza a sospechar de la normalidad de los residuales, además, al realizar la prueba de Shapiro-Wilk se obtiene un valor p de 0, (en realidad es del orden de 10^{-5}) lo cual implica que los residuales no provienen de una población normal.

Ejercicio 7

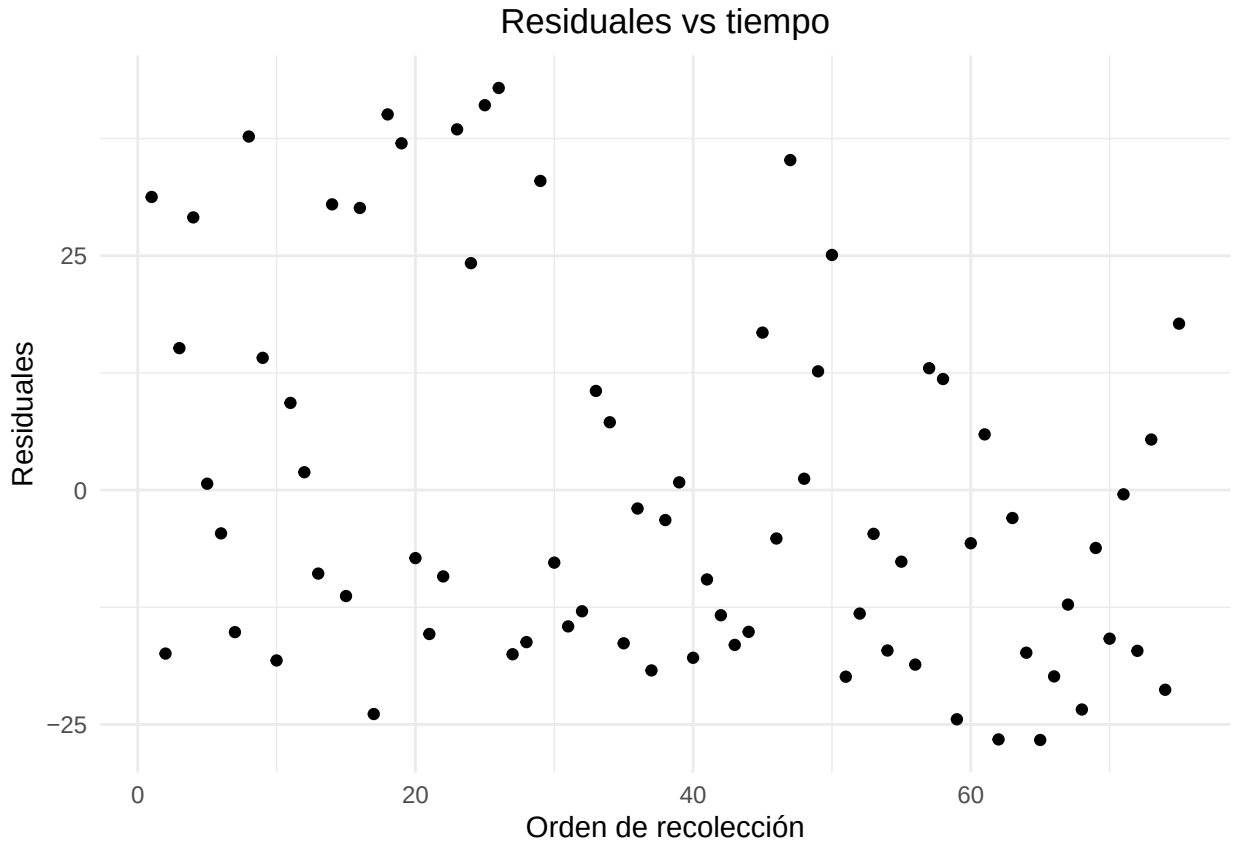

Figura 3: Verificación del supuesto de independencia

En el gráfico anterior no se nota patron alguno, de hecho se nota la “nube de moscas”, por tanto (al menos desde lo que permite ver la gráfico) se cumple el supuesto en cuestión.

Ejercicio 8

Este fue el resultado obtenido luego de ajustar el modelo y realizar la prueba de ajuste.

Cuadro 2: Resumen prueba falta de ajuste

	g.l.	Suma de cuadrados	Cuadrado medio	F_0	Valor p
x4	1	157.3249	157.3249	4.3693	0.0452
Residuales	30	1080.2192	36.0073		
Falta de ajuste	5	141.5328	28.3066	0.7539	0.5912
Error puro	25	938.6864	37.5475		