Estadística II - Taller 03 **Semestre: 2023-01**

Profesores: Francisco Javier Rodríguez Cortés, Julieth Verónica Guarín Escudero

Monitor: Matheo Muñoz Betancur

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.

Tabla 1: Presentación de los datos

y	x
7.775644	3.696441
26.212254	-2.318757
32.596627	-2.963425
12.922859	4.317268
54.455274	-3.727324

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 <- rep(runif(n=floor(n/2)+1, min=-5, max=6),2)[sample(2*floor(n/2)+2,n)]
  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. **Nota:** Asuma que se cumple el supuesto de independencia de los residuales.

4. Extraiga los residuales del modelo y verifique que estos tengan media igual a 0.

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 gráfico cuantil - cuantil y finalmente una prueba de hipótesis.

7. Realice la prueba de falta de ajuste para los datos del modelo

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

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

8. Con la base de datos del archivo decaimiento.xlsx, haga un análisis de si se puede ajustar un modelo, bien sea lineal o intrínsecamente lineal, escriba el modelo y cúales son sus supuestos, reporte los coeficientes estimados e interprételos.

Nota: se propone como ejercicio realizar la validación del modelo

Solución

Ejercicio 1

```
gen_dat <- function(n, seed = 7){
  varianza <- 16
  set.seed(seed)
  x <- rep(runif(n=floor(n/2)+1, min=-5, max=6),2)[sample(2*floor(n/2)+2,n)]
  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) #Generando los datos
```

Ejercicio 2

```
mod <- lm(y ~ x, data = datos)
```

Ejercicio 3

```
summary(mod)
```

```
Call:
lm(formula = y ~ x, data = datos)

Residuals:
   Min     1Q Median     3Q    Max
-30.33 -19.23  -2.35  19.49  41.58

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.3865     2.3899  11.459  < 2e-16 ***
x            -2.5510     0.6747  -3.781 0.000317 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.61 on 73 degrees of freedom
Multiple R-squared:  0.1638,    Adjusted R-squared:  0.1523
F-statistic:  14.3 on 1 and 73 DF,  p-value: 0.0003168
```

Ambos parámetros son individualmente significativos a un nivel de significancia $\alpha = 0.05$.

Ejercicio 4

```
residuales <- residuals(mod) #extrayendo los residuales
mean(residuales) #la media de los residuales da 0
```

```
[1] -4.34236e-16
```

Esto se da producto de utilizar mínimos cuadrados ordinarios, del cual

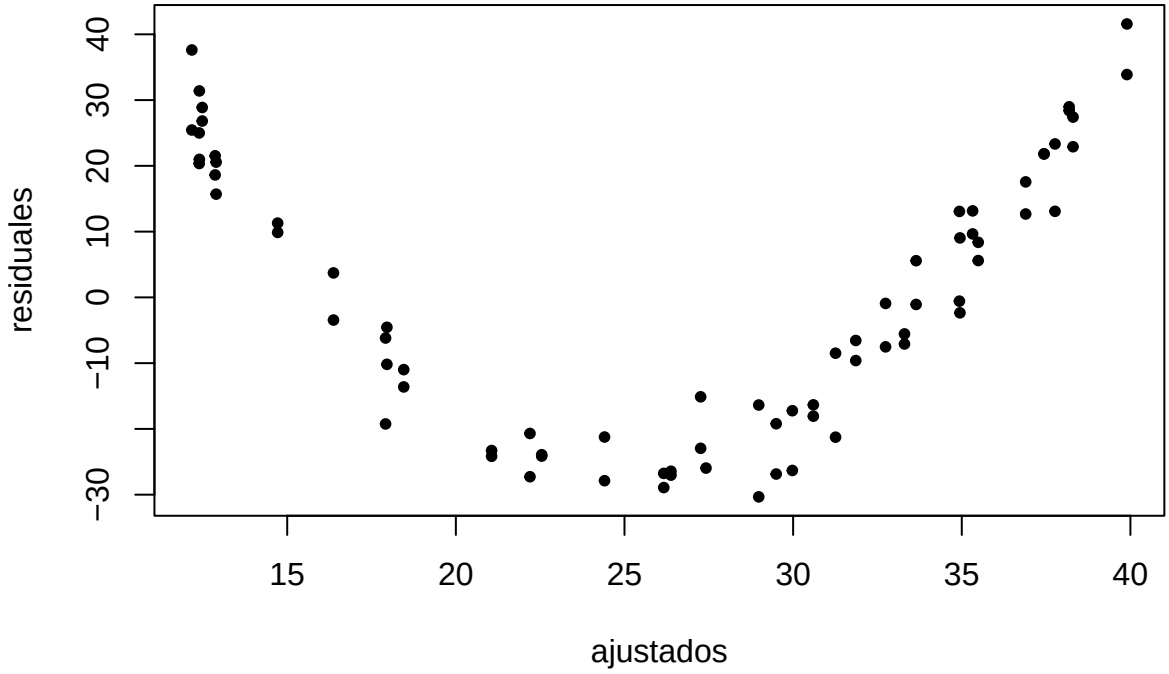
$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \sum_{i=1}^{n} (Y_i - \bar{Y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i) = \sum_{i=1}^{n} [(y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X})] = 0$$

Basado en los residuales, los errores tienen media 0:

$$\sum_{i=1}^{n} \frac{e_i}{n} = 0$$

Ejercicio 5

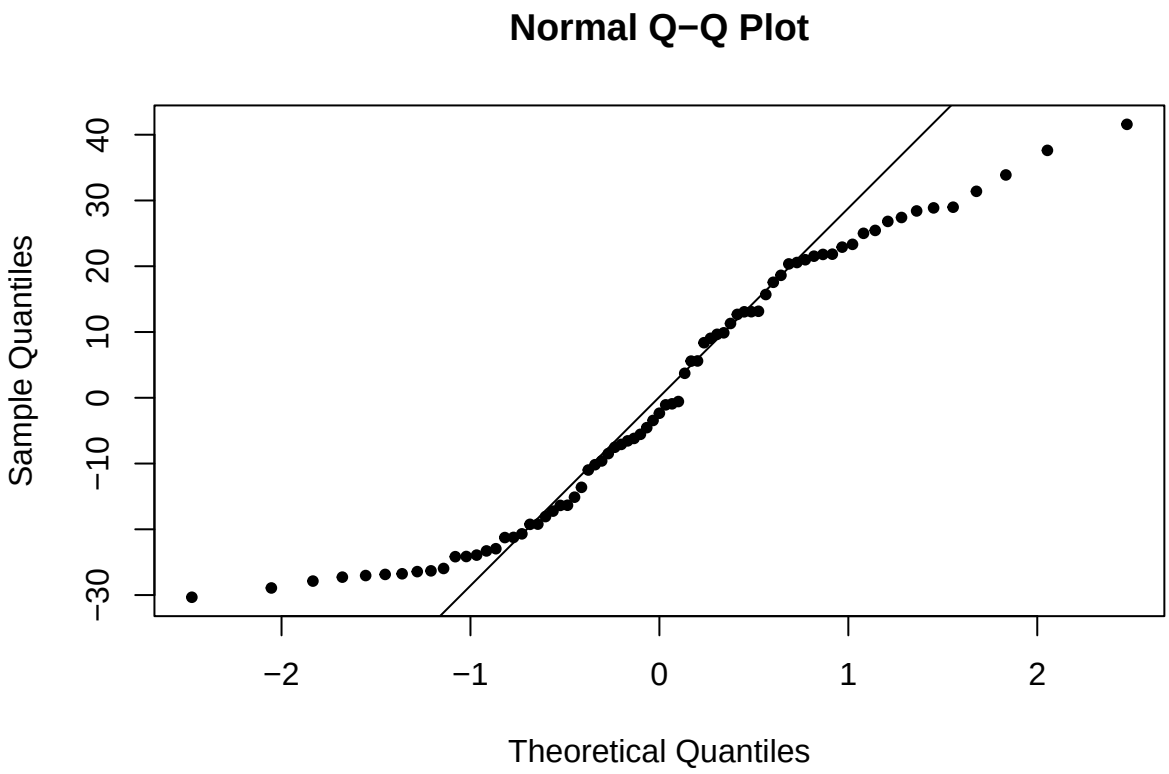
```
ajustados <- fitted(mod) #valores ajustados
plot(ajustados, residuales, pch=20)
```



Es clara la forma en U de los residuales vs valores ajustados, lo que da indicio de falta de ajuste, sin embargo no hay evidencia suficiente en contra de varianza constante, por lo que para este caso se considera que se cumple el supuesto.

Ejercicio 6

```
qqnorm(residuales, pch=20)
qqline(residuales)
```



No se observa una distribución de normalidad en la gráfica de comparación de cuantiles, tanto por el patrón que esta presenta como por tener colas más pesadas, se plantea entonces la siguiente prueba de hipótesis:

$$\begin{cases} H_0 : \text{Los } \varepsilon_i \text{ distribuyen normal} \\ H_a : \text{Los } \varepsilon_i \text{ no distribuyen normal} \end{cases}$$

La siguiente es la prueba de normalidad shapiro wilk para la anterior prueba de hipótesis:

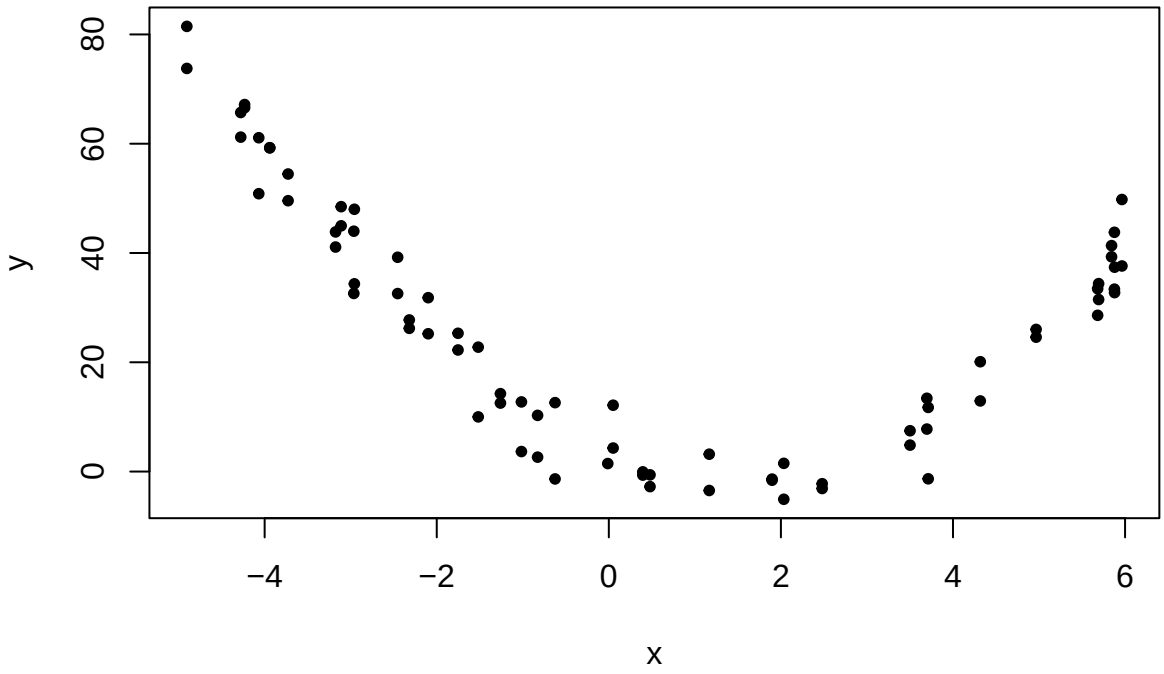
```
shapiro.test(residuales)
```

```
    Shapiro-Wilk normality test

data:  residuales
W = 0.93625, p-value = 0.000944
```

Como Val-P$<\alpha$ entonces se rechaza la hipótesis nula en la que los datos provienen de una población normal, por lo que se acepta la hipótesis en la que no lo hacen.

Ejercicio 7

```
#install.packages("rsm") instalando el paquete necesario
#graficando los datos
with(datos, plot(x, y, pch=20))
```



Es posible hacer prueba de falta de ajuste pues existen valores repetidos para datos en X. Así, la prueba de hipótesis es:

$$\begin{cases} H_0 : \mathrm{E}[y|x_i] = \beta_0 + \beta_1 x_i \\ H_a : \mathrm{E}[y|x_i] \neq \beta_0 + \beta_1 x_i \end{cases}$$

En **R**:

```
library(rsm)
mod.falta.ajuste <- rsm(y ~ FO(x), data = datos)
summary(mod.falta.ajuste)
```

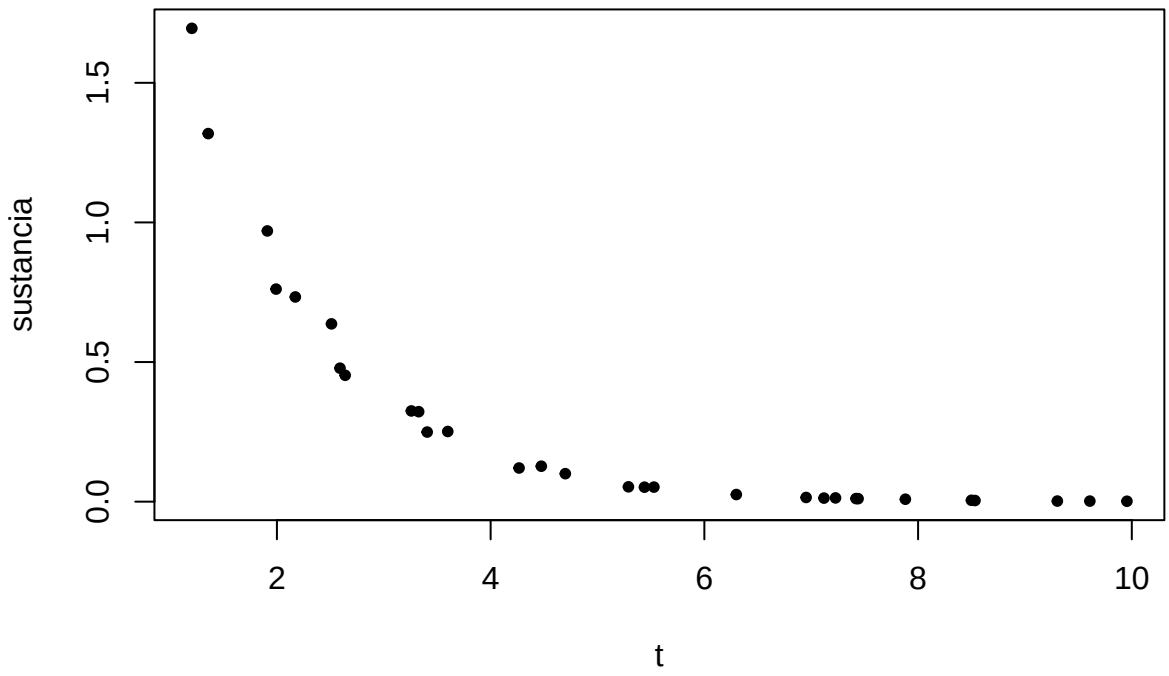
```
Call:
rsm(formula = y ~ FO(x), data = datos)
```

```
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 27.38652    2.38986 11.4595 < 2.2e-16 ***
x           -2.55105    0.67469 -3.7811 0.0003168 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Multiple R-squared:  0.1638,    Adjusted R-squared:  0.1523
F-statistic:  14.3 on 1 and 73 DF,  p-value: 0.0003168

Analysis of Variance Table

Response: y
            Df  Sum Sq Mean Sq F value    Pr(>F)
FO(x)        1  6070.7  6070.7  14.297 0.0003168
Residuals   73 30997.7   424.6
Lack of fit 36 30058.8   835.0  32.905 < 2.2e-16
Pure error  37   938.9    25.4

Direction of steepest ascent (at radius 1):
 x
-1

Corresponding increment in original units:
 x
-1
```

Como el valor P es demasiado pequeño, entonces a cualquier nivel de significancia α se rechaza la hipótesis nula, por lo que el modelo lineal tiene falta de ajuste (hecho que se evidencia gráficamente), también se observa que a un nivel de significancia $\alpha = 0.05$, el modelo es significativo, sin embargo la variabilidad de la respuesta explicada por la regresión es de solamente 16.38 %.

Ejercicio 8

```r
datos.nuevos <- readxl::read_xlsx("decaimiento.xlsx")
with(datos.nuevos, plot(t, sustancia, pch=20))
```



Es claro que el modelo no es lineal, sin embargo tiene un comportamiento exponencial multiplicativo, por lo que se considera $Y = \beta_0 e^{\beta_1 X}\varepsilon$ y se ajusta $Y^* = \beta_o^* + \beta_1 X + \varepsilon^*$ con $Y^* = \ln(Y)$. En **R**:

```
mod.transform <- lm(log(sustancia) ~ t, data = datos.nuevos)
summary(mod.transform)
```

```
Call:
lm(formula = log(sustancia) ~ t, data = datos.nuevos)

Residuals:
     Min       1Q   Median       3Q      Max
-0.12184 -0.08420  0.00987  0.06587  0.15086

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.425612   0.034957   40.78   <2e-16 ***
t           -0.801611   0.006005 -133.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.0863 on 28 degrees of freedom
Multiple R-squared:  0.9984,    Adjusted R-squared:  0.9984
F-statistic: 1.782e+04 on 1 and 28 DF,  p-value: < 2.2e-16
```

Los parámetros son significativos y se leen igual, pero se concluye en función del logaritmo natural de la respuesta. El R^2 de un modelo transformado no es comparable con el R^2 del modelo sin transformar en el caso que se transforme la respesta Y, como en este ejercicio. Se deben analizar los residuales del modelo transformado, pues $\varepsilon_i^* \overset{\text{iid}}{\sim} \text{N ormal}$.