Regresión lineal simple

Alumno: Maldonado Aguilar Angel Julian.

Ejercicio 1

Se supone que el alargamiento de un cable de acero está relacionado linealmente con la intensidad de la fuerza aplicada. Cinco especímenes idénticos de cable dieron los resultados siguientes:

Fuerza (x)	1.0	1.3	2	2.5	3.5
Alargamiento (y)	3	3.5	5	6.5	8

a) Estudia el grado de asociación lineal entre ambas variables.

b) Predice el alargamiento para una fuerza de 2.2.

c) Obtener un intervalo de confianza del 95 % para el valor que se predice de alargamiento para una fuerza de 2.2.

Solución.

a) Grado de asociación lineal entre ambas variables. Para el grado de asociación de ambas variables debemos obtener los valores del *coeficiente de correlación lineal* y el *coeficiente de determinación*.

Para ello primero cargamos los datos, se calcula la media de cada variable y se calculan los coeficientes del modelo.

```
# Se ingresan los datos.
x <- c(1.0, 1.3, 2, 2.5, 3.5)
y <- c(3, 3.5, 5, 6.5,  8)
# Se obtienen las medias de los datos.
barx <- mean(x)
bary <- mean(y)
# Coeficientes del modelo.
b1 <- (sum( (x - barx)*(y - bary) )) / (sum((x - barx)^2))
b0 <- (bary - b1*barx)
```

- Coeficiente de correlación lineal.

```
Sx <- sd(x)
Sy <- sd(y)
r <- Sx * b1 / Sy

round(r, 3)
```

```
## [1] 0.994
```

Con el resultado anterior podemos concluir que como el coeficiente de correlación lineal $r = 0.994$ (un valor bastante cercano a 1), existe una importante relación lineal positiva entre las variables de estudio.

- Coeficiente de determinación.

```
yhat <- b0 + b1*x
R2 <- (sum((yhat - bary)^2)) / (sum((y - bary)^2))

round(R2, 3)
```

```
## [1] 0.988
```

Como el coeficiente de determinación $R^2 = 0.988$, existe un muy buena relación entre las variables de interés.

b) Predicción de alargamiento para una fuerza de 2.2. Como previamente calculamos los valores de β_0 y β_1, podemos usar el modelo de regresion lineal $y = \beta_0 + \beta_1 x$ y asi predecir el valor del alargamiento para cuando la fuerza es igual a 2.2. Esto es:

```
b0 + b1 * 2.2
```

```
## [1] 5.490433
```

Podemos corroborar nuestra predicción usando la funcion de R *predict()*, la cual nos da:

```
reg <- lm(y~x)
pred_force <- data.frame(x = c(2.2))
predict(reg, pred_force)
```

```
##        1
## 5.490433
```

Con ambos método podemos concluir que para una fuerza de 2.2 se predice un alargamiento de 5.49 (redondeado a 2 decimales).

c) Intervalo de confianza del 95 % para el valor de alargamiento que se predijo para una fuerza de 2.2 Para el intervalo de confianza, podemos volver a usar la función *predict()* solo que ahora tambien hay que pasarle que el intervalo sea de tipo *confianza* y el nivel de confianza. Con lo cual obtenemos:

```
conf_level <- 0.95
predict(reg, pred_force, interval = "confidence", level = conf_level)
```

```
##        fit      lwr      upr
## 1 5.490433 5.112965 5.867901
```

Del resultado anterior los valores de $lwr \approx 5.113$ y $upr \approx 5.868$ son los limites inferior y superior respectivamente para el intervarlo de confianza del 95 % para el valor que se predijo de alargamiento (5.49) dada una fuerza de 2.2.

Ejercicio 2

Las bodegas modernas utilizan vehículos guiados computarizados y automatizados para el manejo de materiales. En consecuencia, la disposición física de la bodega debe diseñarse con cuidado a modo de evitar el congestionamiento de los vehículos y optimizar el tiempo de respuesta. En *The journal of Engineering for Industry* (agosto 1993) se estudió el diseño óptimo de una bodega automatizada. La disposición empleada supone que los vehículos no se bloquean entre sí cuando viajan dentro de la bodega, es decir, no hay congestionamiento. La validez de este supuesto se verificó simulando por ordenador las operaciones de la bodega. En cada simulación se varió el número de vehículos y se registró el tiempo de congestionamiento (tiempo total que un vehículo bloquea a otro). Los datos se muestran en la tabla de abajo. Los investigadores están interesados en conocer la relación entre el tiempo de congestionamiento (y) y el número de vehículos (x).

x	1	2	3	4	5	6	7	8	9	10
y	0	0	0.02	0.01	0.01	0.01	0.03	0.03	0.02	0.04

a) Obtén la recta de regresión que expresa el tiempo de congestión en función del número de vehículos.

b) Calcula los coeficientes de correlación y el coeficiente de determinación e interpreta los resultados.

Solución.

a) Recta de regresión que expresa el tiempo de congestión en función del número de vehículos.

```
# Se ingresan los datos.
x <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
y <- c(0, 0, 0.02, 0.01, 0.01, 0.01, 0.03, 0.03, 0.02, 0.04)
# Se obtienen las medias de los datos.
barx <- mean(x)
bary <- mean(y)
# Coeficientes del modelo.
b1 <- (sum( (x - barx)*(y - bary) )) / (sum((x - barx)^2))
b0 <- (bary - b1*barx)

print( c(round(b0, 4), round(b1, 4)) )
```

```
## [1] -0.0033  0.0037
```

Por lo tanto, la recta de regresión es

$$y = -0.0033 + 0.0037x$$

b) Coeficiente de correlación y coeficiente de determinación.

- Coeficiente de correlación lineal.

```
Sx <- sd(x)
Sy <- sd(y)
r <- Sx * b1 / Sy

round(r, 3)
```

```
## [1] 0.837
```

Con el resultado anterior podemos concluir que como el coeficiente de correlación lineal $r = 0.837$ (un valor bastante cercano a 1), existe una importante relación lineal positiva entre las variables de estudio.

- Coeficiente de determinación.

```
yhat <- b0 + b1*x
R2 <- (sum((yhat - bary)^2)) / (sum((y - bary)^2))

round(R2, 3)
```

```
## [1] 0.7
```

Como el coeficiente de determinación $R^2 = 0.7$, un valor cercano a 1, entonces existe una buena relación entre las variables de interés.

Por lo anterior, realizar un modelo de regresión lineal es bastante adecuado para este problema.

Ejercicio 3

Los siguientes datos se refieren al crecimiento de una colonia de bacterias en un medio de cultivo:

x	3	2	9	12	15	17
y	11500	14700	23900	35600	57900	86400

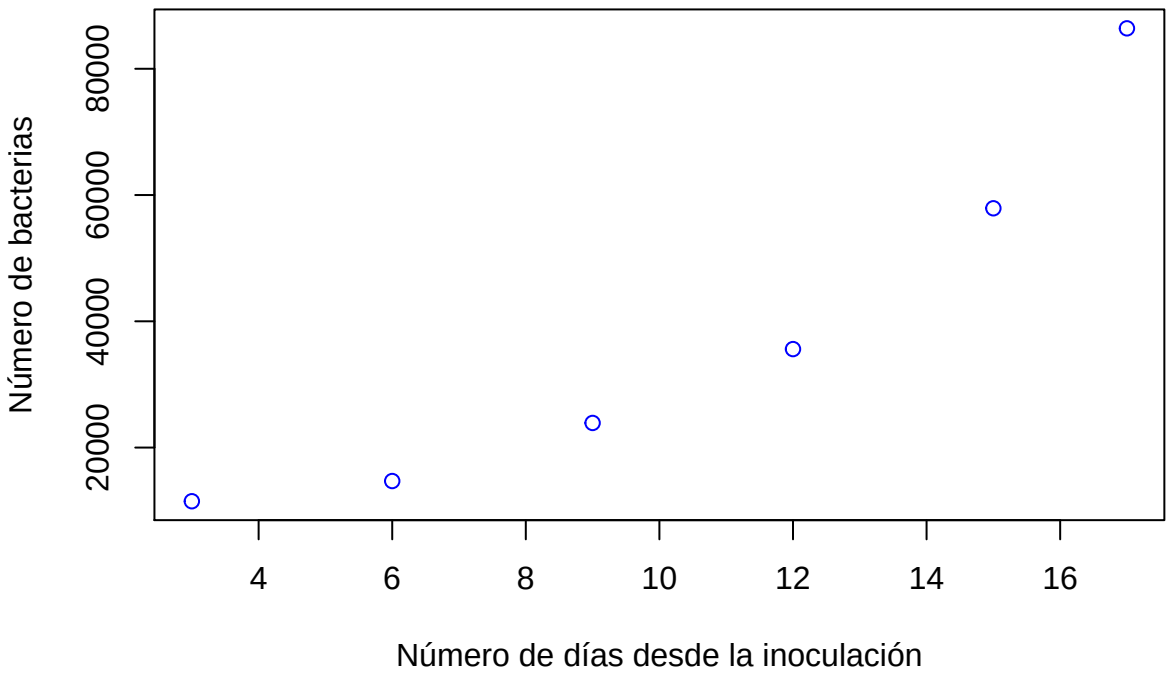
Siendo x el número de días desde la inoculación e y el número de bacterias. Comprobar gráfica y matemáticamente que el tipo de asociación entre ambas variables no es lineal.

Solución.

Comprobación gráfica. Al realizar una nube de puntos de las dos variables tenemos lo siguiente.

```
x <- c(3, 6, 9, 12, 15, 17)
y <- c(11500, 14700, 23900, 35600, 57900, 86400)

plot(x, y, xlab = "Número de días desde la inoculación", ylab = "Número de bacterias", col = "blue")
```



De la anterior gráfica se observa que los puntos no parecen ser aleatorios y que sigan una linea recta, sino que parecen seguir un patrón exponencial por lo que el tipo de asociación entre ambas variables no es lineal.

Comprobación matemática. Para comprobar matematicamente que el tipo de asociación entre las variables no es lineal, lo primero que podemos analizar son los coeficientes tanto de correlación como de determinación. Posteriormente tambien podriamos analizar los 3 supuestos que se deberian cumplir.

Como primero analizaremos los coeficientes, antes calculamos las medias de los datos y los coeficientes del modelo ya que seran de utilidad.

```
# Se obtienen las medias de los datos.
barx <- mean(x)
bary <- mean(y)
# Coeficientes del modelo.
b1 <- (sum( (x - barx)*(y - bary) )) / (sum((x - barx)^2))
b0 <- (bary - b1*barx)
```

Cálculo de los coeficientes:

- Coeficiente de correlación lineal.

```
Sx <- sd(x)
Sy <- sd(y)
r <- Sx * b1 / Sy

round(r, 3)
```

```
## [1] 0.937
```

- Coeficiente de determinación.

```
yhat <- b0 + b1*x
R2 <- (sum((yhat - bary)^2)) / (sum((y - bary)^2))

round(R2, 3)
```

```
## [1] 0.877
```

Como tanto el coeficiente de correlación ($r = 0.937$) asi como el coeficiente de determinación ($R^2 = 0.877$) dieron resultados cercanos a *1*, se puede concluir que existe una importante relacion entre las variables de estudio. Por lo que ahora pasaremos a analizar los supuestos.

- Supuesto de normalidad.

Hipotesis.

$$H_0 : \text{Los datos son normales}$$

$$H_1 : \text{Los datos no son normales}$$

Cálculo del *p-valor* usando la prueba de Shapiro-Wilks

```
# Modelo de regresión lineal en R.
reg <- lm(y~x)
# Residuales.
e <- reg$residuals
# Prueba de Shapiro-Wilks
shapiro.test(e)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  e
## W = 0.91114, p-value = 0.444
```

Como $p - value = 0.444$ no tenemos evidencia suficiente para rechazar H_0 por lo que se asume que los residuales son normales.

- Supuesto de varianza constante.

Hipótesis:

$$H_0 : \text{La varianza es constante}$$

$$H_1 : \text{La varianza no es constante}$$

Cálculo del *p-valor* usando la prueba de Breusch-Pagan.

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.0.5
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.0.5
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(y~x)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  y ~ x
## BP = 0.90037, df = 1, p-value = 0.3427
```

Como $p - value = 0.3427$ no tenemos evidencia suficiente para rechazar H_0 por lo que se asume que la varianza de los errores es constante.

- Supuesto de independencia.

Hipotesis.

$$H_0: \text{ Existe independencia}$$

$$H_1: \text{ No existe independencia}$$

Cálculo del *p-valor* usando la prueba de Durbin-Watson.

```
dwtest(y~x)
```

```
##
##  Durbin-Watson test
##
## data:  y ~ x
## DW = 1.1282, p-value = 0.008701
## alternative hypothesis: true autocorrelation is greater than 0
```

Como $p - value \approx 0.0087$ es un valor muy pequeño, hay evidencia suficiente para rechazar H_0, por lo que se concluye que no existe independencia en las observaciones.

Por lo tanto, se puede afirmar que el tipo de asociación entre ambas variables no es lineal ya que con un supuesto que no se cumpla (en este caso el de independencia) cualquier resultado encontrado en el modelo de regresión lineal no tiene validez.

Ejercicio 4

Se ha realizado un estudio para investigar el efecto de un determinado proceso térmico en la dureza de una determinada pieza. Once piezas se seleccionaron para el estudio. Antes del tratamiento se realizaron pruebas de dureza para determinar la dureza de cada pieza. Después, las piezas fueron sometidas a un proceso térmico de templado con el fin de mejorar su dureza. Al final del proceso, se realizaron nuevamente pruebas de dureza y se obtuvo una segunda lectura. Se recogieron los siguientes datos (Kg. de presión):

Dureza previa (x)	182	232	191	200	148	249	276	213	241	480
Dureza post.(y)	198	210	194	220	138	220	219	161	210	313

a) Realiza un gráfico de dispersión para analizar la relación.

b) ¿Se puede afirmar que el proceso de templado mejora la dureza de las piezas?

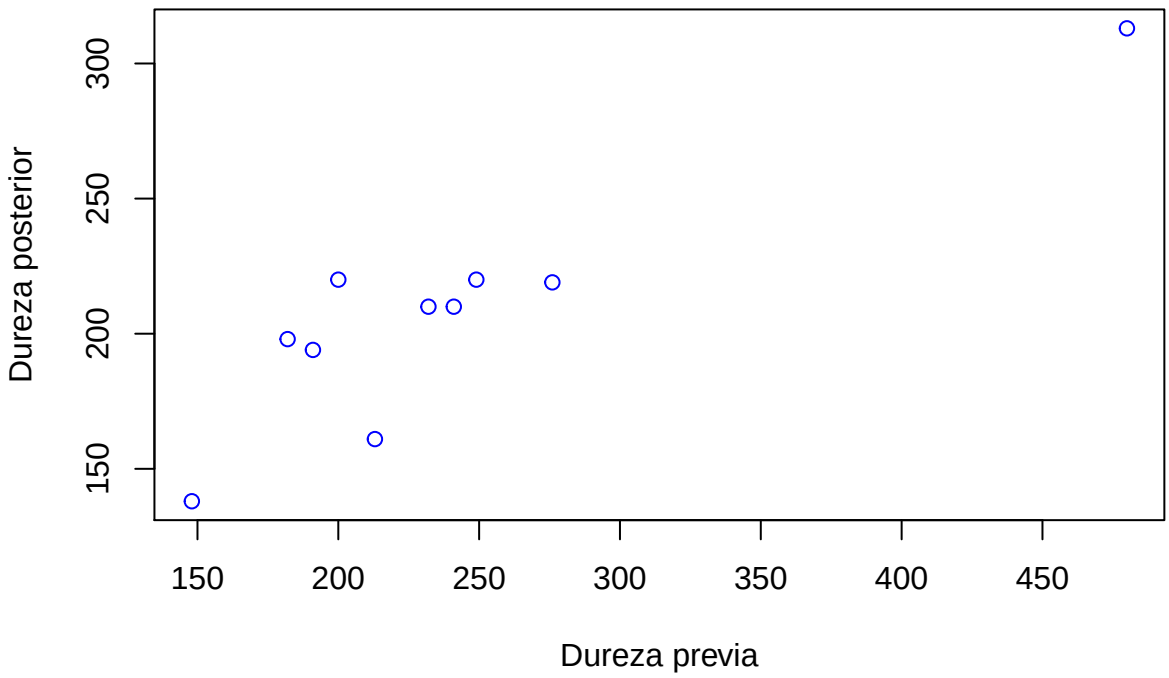
c) Decide si un modelo lineal es adecuado para explicar la dureza posterior en función de la dureza previa. En caso afirmativo obtenlo y predice la dureza tras el proceso de templado de una pieza con un dureza previa de 215.

Solución.

a) Gráfico de dispersión.

```
x <- c(182, 232, 191, 200, 148, 249, 276, 213, 241, 480)
y <- c(198, 210, 194, 220, 138, 220, 219, 161, 210, 313)

plot(x, y, xlab = "Dureza previa", ylab = "Dureza posterior", col = "blue")
```



Como se puede observar en el grafico anterior, los datos si parecen estar distribuidos aleatoriamente ademas de que parecen tener una relación lineal positiva.

b) ¿Se puede afirmar que el proceso de templado mejora la dureza de las piezas? Para llegar a una conclusión podemos hacer una prueba de hipótesis de la diferencia de 2 medias.

1. **Planteamiento de las hipótesis:**

Donde μ_1 representa la media de la dureza previa y μ_2 la media de la dureza posterior. Por lo tanto las hipotesis a plantear son:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ (El proceso de templado no mejora la dureza de las piezas)}$$
$$H_1 : \mu_1 - \mu_2 < 0 \text{ (El proceso de templado mejora la dureza de las piezas)}$$

2. **Cálculo del estadístico de prueba:**

Como las muestras son pequeñas, se desconoce la varianza poblacional y podemos asumir que las varianzas poblacionales son iguales ya que las muestras pertenecen al mismo conjunto, por lo tanto nuestro estadistico de prueba es:

$$t_{n_1+n_2-2} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{k}}$$
$$k = n_1 + n_2 - 2$$

Por lo que el valor del estadistico de prueba es (redondeado a 3 decimales):

```
x1 <- mean(x)
x2 <- mean(y)
s1 <- sd(x)
s2 <- sd(y)
n1 <- 10
n2 <- 10

k <- n1 + n2 - 2
sp <- sqrt(( (n1-1)*s1**2 + (n2-1)*s2**2 ) / k)
t <- ((x1 - x2) - (0)) / (sp * sqrt(1/n1 + 1/n2))

round(t, 3)
```

```
## [1] 1.016
```

3. **Cálculo del p-valor**

Como la prueba es unilateral, tenemos que calcular la siguiente probabilidad.

$$P(t_{18} < 1.016)$$

```
p_value <- pt(t, n1+n2-2)
round(p_value, 3)
```

```
## [1] 0.838
```

4. **Toma de decisión**

Como $p-value \approx 0.838$ no hay evidencia suficiente para afirmar que el proceso de templado mejore la dureza de las piezas.

c) ¿Un modelo lineal es adecuado para explicar la dureza posterior en función de la dureza previa?. En caso afirmativo obtenlo y predice la dureza tras el proceso de templado de una pieza con un dureza previa de 215. Para saber si un modelo lineal es adecuado, analizaremos el valor del coeficiente de determinación (R^2), el cual es:

```
# Se ingresan los datos.
barx <- mean(x)
bary <- mean(y)
# Se obtienen las medias de los datos.
b1 <- (sum( (x - barx)*(y - bary) )) / (sum((x - barx)^2))
b0 <- (bary - b1*barx)
# Coeficientes del modelo.
yhat <- b0 + b1*x
R2 <- (sum((yhat - bary)^2)) / (sum((y - bary)^2))

round(R2, 3)
```

```
## [1] 0.82
```

Como el coeficiente de determinación $R2 = 0.82$ (valor muy cercano a 1) existe un buena relación entre las variables de interés, por lo que se concluye que **un modelo lineal es adecuado** para explicar la dureza posterior en función de la dureza previa.

Como ya sabemos que un modelo lineal es adecuado para este problema a continuacion se calculara la predicción de la dureza tras el proceso de templado de una pieza con un dureza previa de 215. Para ello nos apoyaremos de la recta de regresión $y = 99.2243 + 0.4522x$, la cual al pasarle el valor de la varible independiente obtenemos:

```
pred <- b0 + b1 * 215
round(pred, 2)
```

```
## [1] 196.45
```

Podemos corroborar nuestra predicción usando la funcion de R *predict()*, la cual nos da:

```
reg <- lm(y~x)
pred_hardness <- data.frame(x = c(215))
predict(reg, pred_hardness)
```

```
##        1
## 196.4518
```

Con ambos método podemos concluir que para una dureza previa de 215 se predice una dureza posterior de aproximadamente 196.45.

Ejercicio 5

La hidrólisis de un cierto éster tiene lugar en medio ácido según un proceso cinético de primer orden. Partiendo de una concentración inicial (por 10^3 (M)) desconocida del éster, se han medido las concentraciones del mismo a diferentes tiempos (en minutos) obteniéndose los resultados siguientes:

Tiempo (x)	3	5	10	15	20	30	40	50	60	75	90
Concentración (y)	25.5	23.4	18.2	14.2	11	6.7	4.1	2.5	1.5	0.7	0.3

a) Realiza una nube de puntos de las dos variables. La teoría cinética de este tipo de reacciones nos indica que la evolución de la concentración del éster en función del tiempo se rige por $C_t = C_0 e^{-kt}$, donde C_0 es la concentración inicial. ¿Qué transformación de los datos nos lleva a un modelo lineal?. Realiza esta transformación y obtén la concentración inicial C_0 y la velocidad k de desaparición del éster.

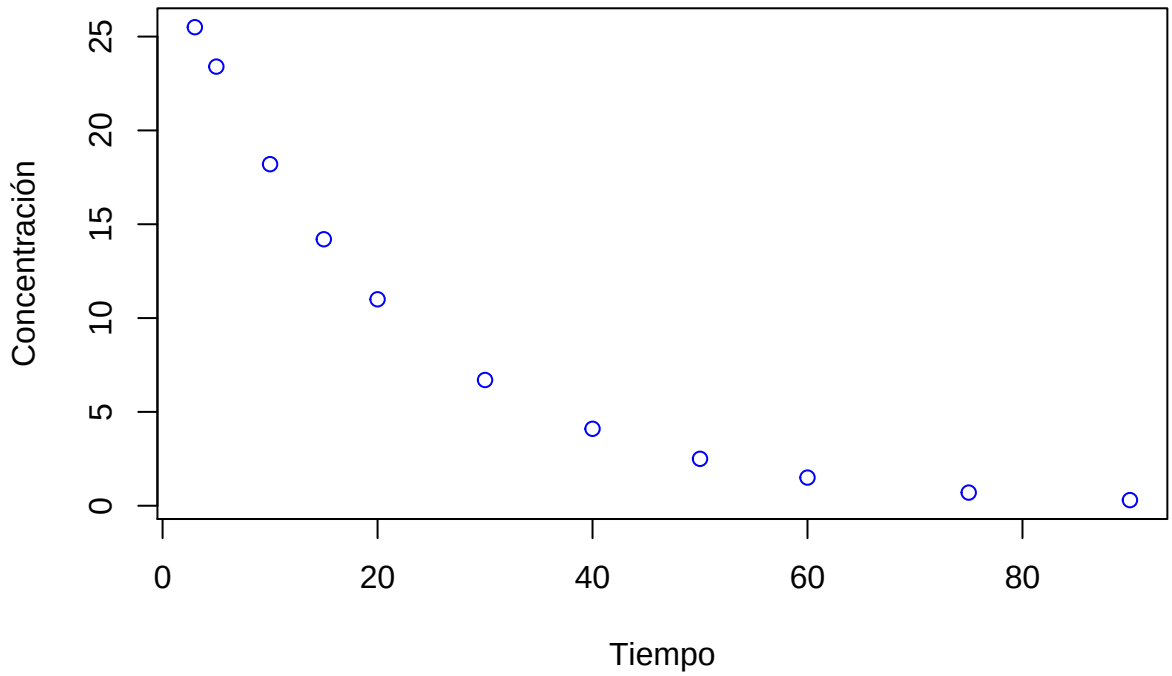
b) Suponemos ahora que nos comunican que la concentración inicial del éster es $C_0 = 3.10^{-2}$(M). ¿Cómo incorporar esta información a nuestro análisis anterior?. Obtén el nuevo valor de k.

Solución.

a) Nube de puntos de las 2 variables.

```r
x <- c(3, 5, 10, 15, 20, 30, 40, 50, 60, 75, 90)
y <- c(25.5, 23.4, 18.2, 14.2, 11, 6.7, 4.1, 2.5, 1.5, 0.7, 0.3)

plot(x, y, xlab = "Tiempo", ylab = "Concentración", col = "blue")
```



Como se puede observar en el grafico anterior la relación entre las variables es de tipo exponencial negativa.

Para convertir la funcion exponencial a lineal, debemos aplicar para este caso logaritmo base e (logaritmo natural) a ambas partes de la ecuación, es decir:

$$C_t = C_0 e^{-kt} \tag{1}$$
$$ln(C_t) = ln(C_0 e^{-kt}) \tag{2}$$
$$ln(C_t) = ln(C_0) + ln(e^{-kt}) \tag{3}$$
$$ln(C_t) = ln(C_0) - kt \tag{4}$$

De tal forma que ahora

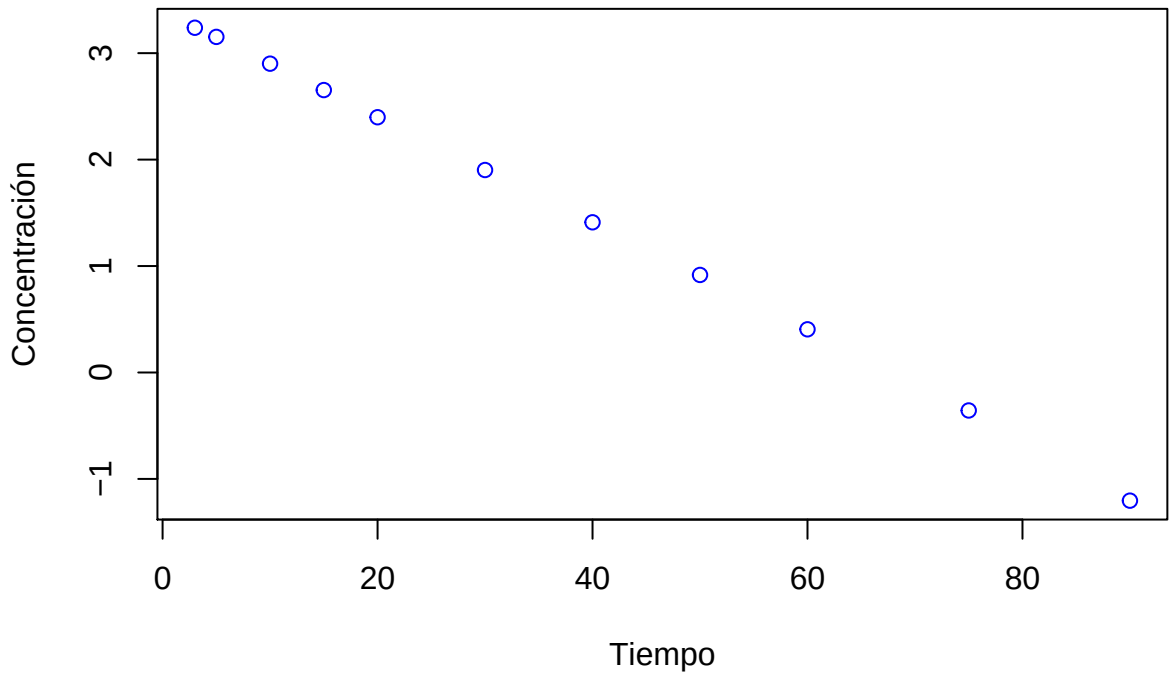
$$y = ln(C_t)$$
$$\beta_0 = ln(C_0)$$
$$\beta_1 = -k$$

Si volvemos a graficar la nube de puntos, donde ahora los valores de

$$y = ln(C_t)$$

tenemos lo siguiente.

```
logy <- log(y)
plot(x, logy, xlab = "Tiempo", ylab = "Concentración", col = "blue")
```



Donde de la gráfica anterior podemos ver que hemos llegado a un modelo lineal con una asociación lineal negativa.

Ahora para poder obtener los valores de la concentración inicial C_0 y la velocidad k de desaparición del éster, podemos despejar sus valores de $\beta_0 = ln(C_0)$ y $\beta_1 = -k$. De esta forma tenemos que:

$$C_0 = e^{\beta_0}$$
$$k = -\beta_1$$

Ya que tenemos sus formulas, ahora nos hace falta calcular los valores de los coeficientes del modelo, los cuales podemos obtener de la siguiente manera

```
reg <- lm(logy~x)
reg
```

```
##
## Call:
## lm(formula = logy ~ x)
##
## Coefficients:
## (Intercept)            x
##      3.4144      -0.0506
```

Del resultado anterior tenemos que $\beta_0 = 3.4144$ y $\beta_1 = -0.0506$

Por lo tanto

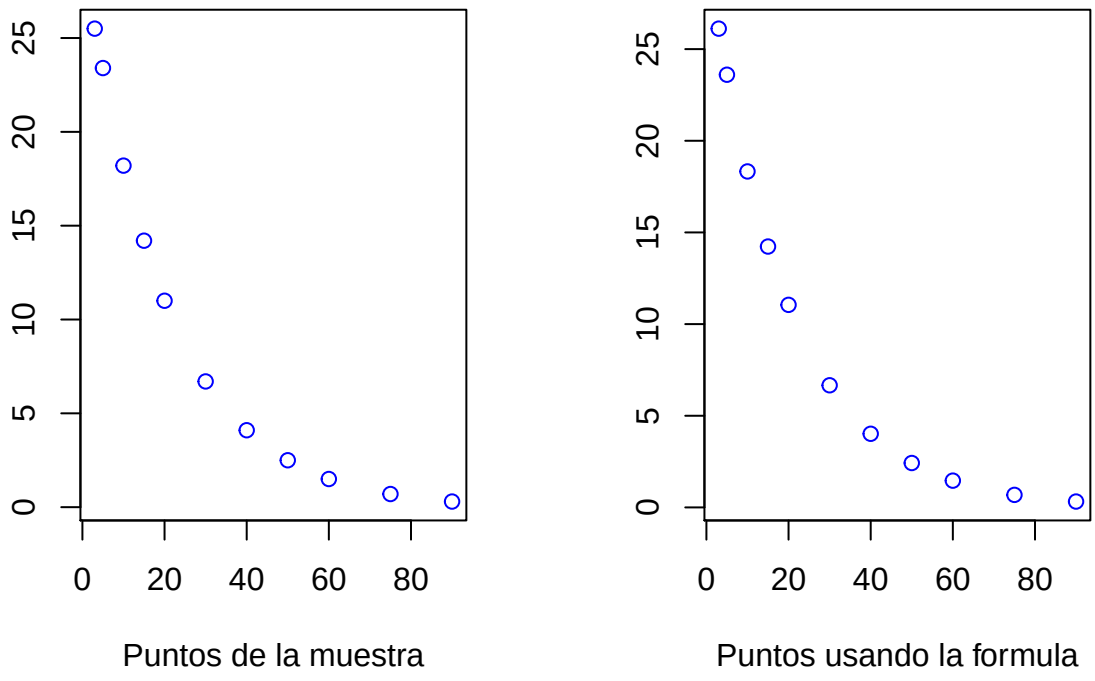
$$C_0 = e^{3.4144} = 30.3987$$
$$k = -(-0.0506) = 0.0506$$

Para corroborar si encontramos los valores correctos de C_0 y k, podemos comparar las graficas de la nube de puntos de los datos de la muestra contra la grafica de los puntos donde para cada valor x (tiempo), el valor y (concentración) lo calculamos con la formula dada $C_t = C_0 e^{-kt} = 30.3987e^{-0.0506t}$

```
C0 <- 30.3987
k <- 0.0506

y_from_formula <- C0 * exp(1)^(-k * x)

par(mfrow = c(1, 2))
plot(x, y, xlab = "Puntos de la muestra", ylab = "", col = "blue")
plot(x, y_from_formula, xlab = "Puntos usando la formula", ylab = "", col = "blue")
```



Dado que la nube de puntos de ambas graficas son muy parecidos podemos concluir que los valores de $C_0 \approx 30.3987$ y $k \approx 0.0506$ son correctos.

b) Suponemos ahora que nos comunican que la concentración inicial del éster es $C_0 = 3.10^{-2}$(M)

Como ya sabemos que

$$C_0 = e^{\beta_0}$$
$$k = -\beta_1$$

Entonces, podemos sustuir la concentracion inicial dada 3.10^{-2}

$$C_0 = e^{\beta_0}$$
$$3.10^{-2} = e^{\beta_0}$$

Por lo tanto

$$\beta_0 = ln(3.10^{-2})$$

Y como sabemos que la relacion entre β_0 y β_1 es:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Entonces

$$\beta_1 = \frac{\bar{y} - \beta_0}{\bar{x}}$$

```
b0 <- log(3.10^(-2))
b1 <- (mean(y) - b0) / mean(x)
round(b1, 4)
```

```
## [1] 0.3341
```

Por lo tanto el nuevo valor de k es:

$$k = -\beta_1 \approx -0.3341$$

Ejercicio 6

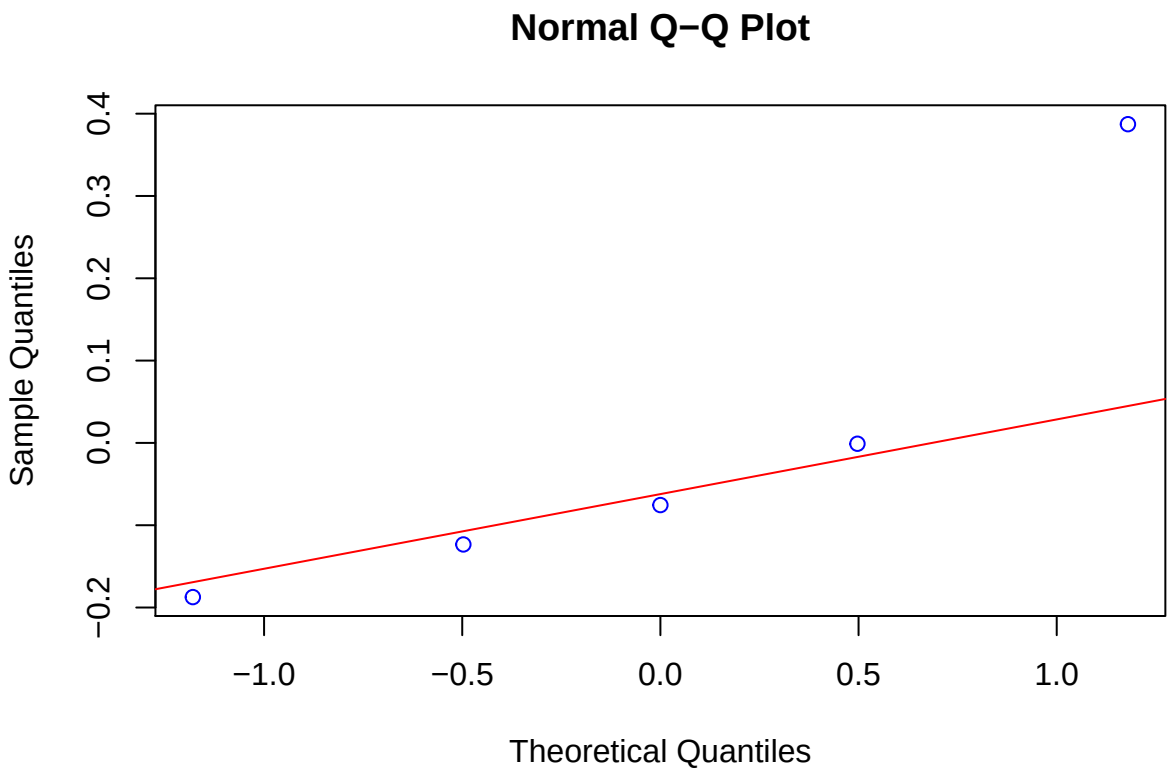
De los ejercicios anteriores donde respondiste que es adecuado realizar regresión lineal para analizar la asociación entre las dos variables, revisa cada uno de los supuestos que cumple, tanto de forma gráfica como por medio de pruebas.

Solución.

Supuestos para ejercicio 1.

Supuesto de normalidad (forma gráfica). Visualizamos el gráfico.

```r
# Se ingresan los datos.
x <- c(1.0, 1.3, 2, 2.5, 3.5)
y <- c(3, 3.5, 5, 6.5,  8)
# Modelo de regresión lineal en R.
reg <- lm(y~x)
# Residuales.
e <- reg$residuals
# Gráfico.
qqnorm(e, col = "blue")
qqline(e, col = "red")
```



Como se puede observar gran parte de los puntos se encuentran sobre la linea roja a excepción de uno, por ello se esperaría que se cumpla el supuesto de normalidad.

Supuesto de normalidad (por medio de pruebas). Hipotesis:

$$H_0 : \text{Los residuales son normales}$$

$$H_1 : \text{Los residuales no son normales}$$

Prueba de Shapiro-Wilk.

```r
shapiro.test(e)
```

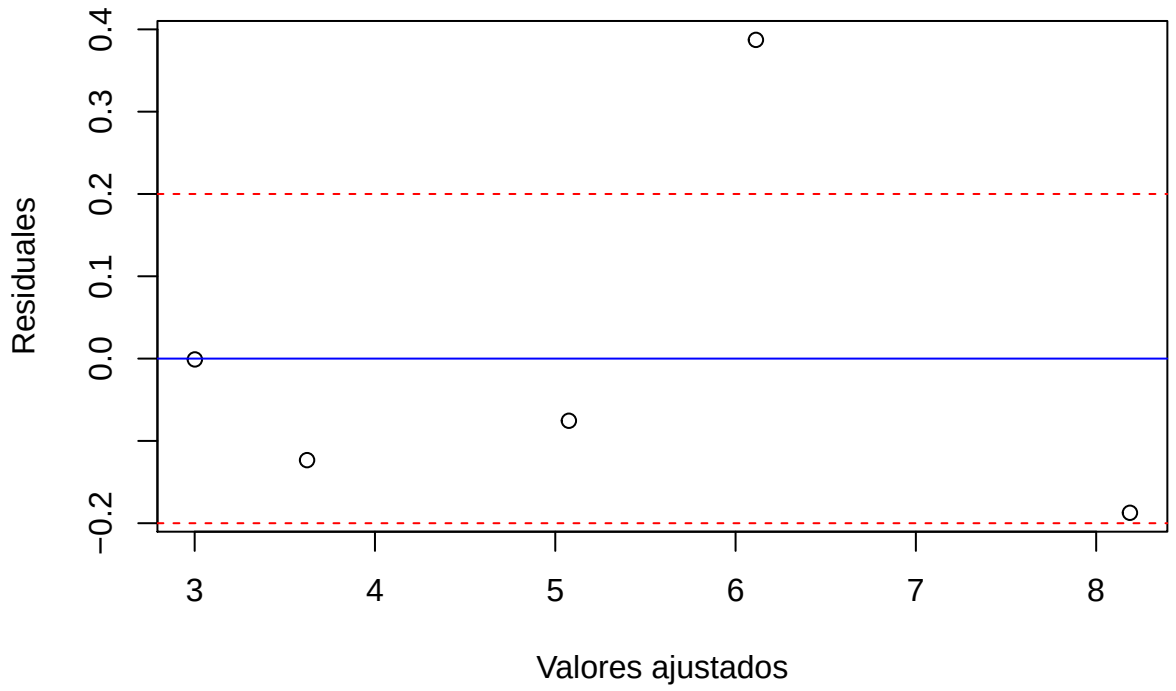
```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  e
## W = 0.82176, p-value = 0.1205
```

Como $p - value = 0.1205$ se puede concluir que a partir de un nivel $\alpha = 0.13$ es decir con un máximo de confianza del 87% se puede rechazar H_0 lo que da lugar a que los residuales no sean normales. Por otra parte solo hasta un $\alpha = 0.12$ con un minimo de confianza de 88% se puede asumir que los residuales son normales.

Supuesto de varianza constante (forma gráfica). Visualizamos el gráfico.

```r
# Datos ajustados.
aj <- reg$fitted.values
# Grafico de ajustados vs residuales.
plot(aj, e, xlab = "Valores ajustados", ylab = "Residuales")
abline(0, 0, col = "blue")
abline(-0.2, 0, col = "red", lty = 2)
abline(0.2, 0, col = "red", lty = 2)
```



Se cumple el supuesto de varianza constante ya que hay puntos dispersos de forma aleatoria dentro de una banda centrada en cero y con bandas colocadas en -0.2 y 0.2.

Supuesto de varianza constante (por medio de pruebas). Hipótesis:

$$H_0: \text{ La varianza de los errores es constante}$$

H_1 : La varianza de los errores no es constante

Prueba de Breusch-Pagan

```
bptest(y~x)
```

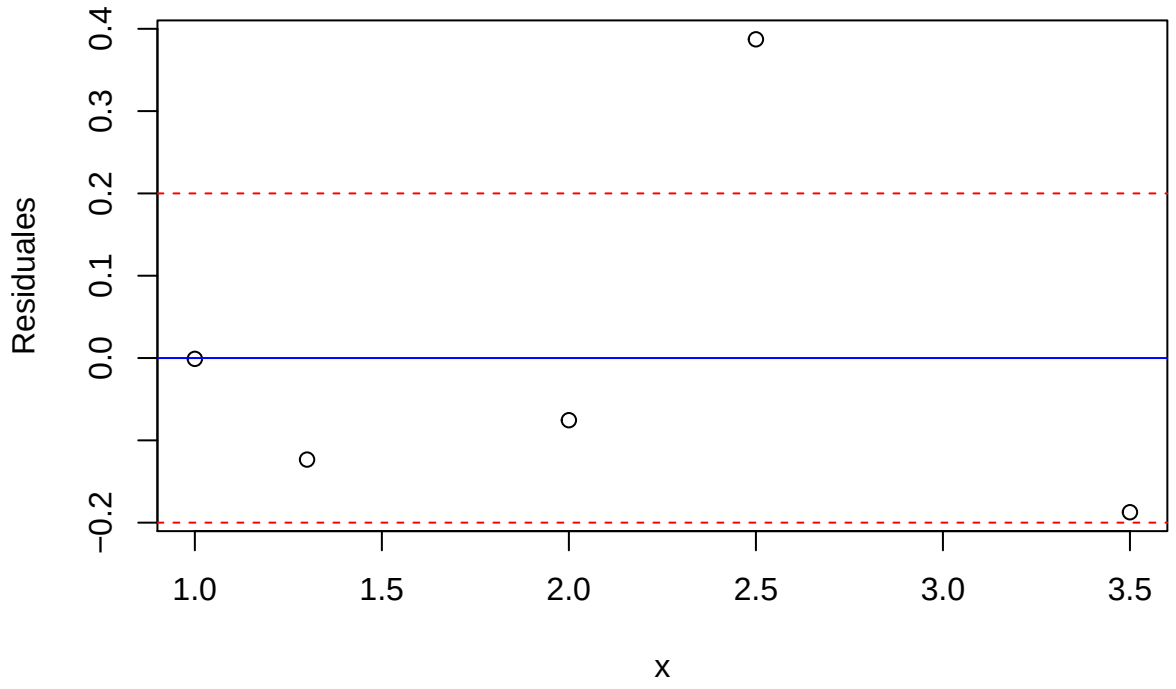
```
##
##  studentized Breusch-Pagan test
##
## data:  y ~ x
## BP = 0.88873, df = 1, p-value = 0.3458
```

Como $p - value = 0.3458$ lo cual no es tan pequeño, entonces no hay evidencia contundente para rechazar H_0 por que se se puede asumir que la varianza de los errores es constante.

Supuesto de independencia (forma gráfica). Visualizamos el gráfico.

```
e <- reg$residuals

plot(x, e, xlab = "x", ylab = "Residuales")
abline(0, 0, col = "blue")
abline(-0.2, 0, col = "red", lty = 2)
abline(0.2, 0, col = "red", lty = 2)
```



Como los puntos se encuentran dispersos de forma aleatoria dentro de una banda centrada en cero y con bandas colocadas en -0.2 y 0.2, por lo tanto se cumple el supuesto de independencia.

Supuesto de independencia (por medio de pruebas). Hipótesis:

$$H_0: \text{ Existe independencia en las observaciones}$$
$$H_1: \text{ No existe independencia en las observaciones}$$

Prueba de Durbin-Watson.

```
dwtest(y~x)
```

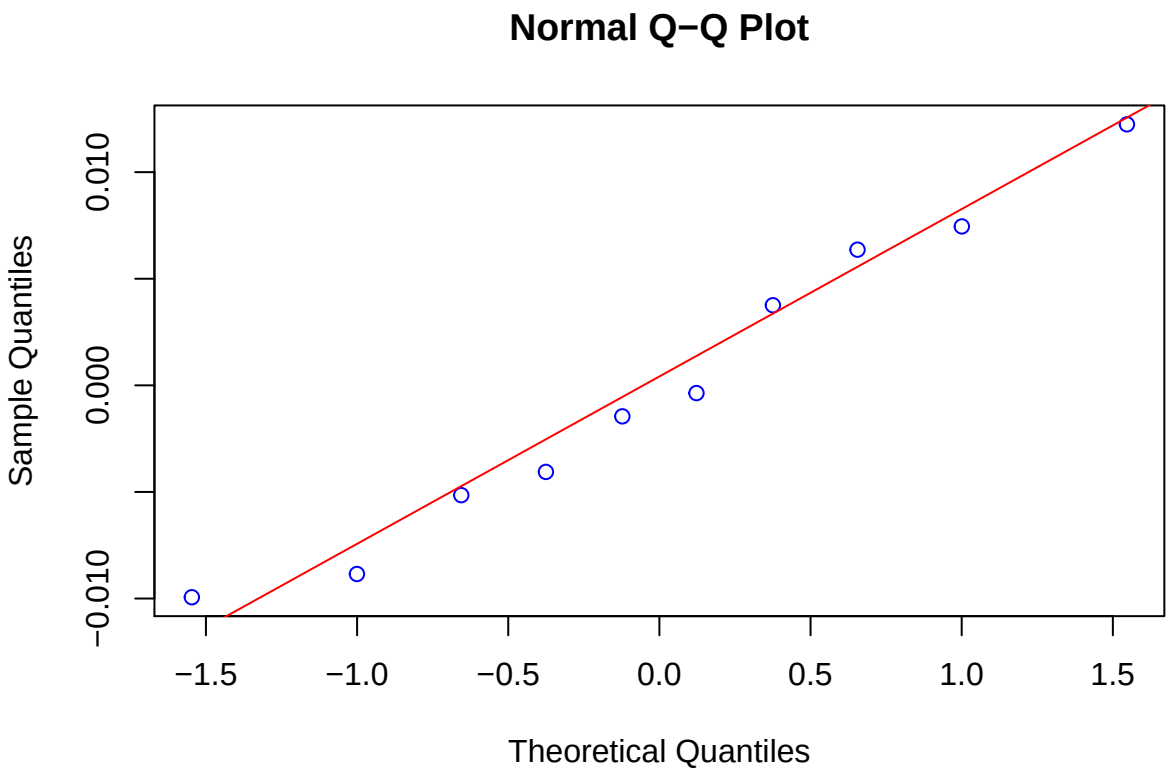
```
##
##  Durbin-Watson test
##
## data:  y ~ x
## DW = 2.7263, p-value = 0.6339
## alternative hypothesis: true autocorrelation is greater than 0
```

Como $p - value = 0.6339$ lo cual es relativamente grande, entonces no hay evidencia para rechazar H_0 por lo es válido suponer que existe independencia en las observaciones.

Supuestos para ejercicio 2.

Supuesto de normalidad (forma gráfica). Visualizamos el gráfico.

```
# Se ingresan los datos.
x <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
y <- c(0, 0, 0.02, 0.01, 0.01, 0.01, 0.03, 0.03, 0.02, 0.04)
# Modelo de regresión lineal en R.
reg <- lm(y~x)
# Residuales.
e <- reg$residuals
# Gráfico.
qqnorm(e, col = "blue")
qqline(e, col = "red")
```



Como se puede observar todos los puntos se encuentran cerca de la linea roja, por ello se esperaría que se cumpla el supuesto de normalidad.

Supuesto de normalidad (por medio de pruebas). Hipotesis:

$$H_0 : \text{Los residuales son normales}$$

$$H_1 : \text{Los residuales no son normales}$$

Prueba de Shapiro-Wilk.

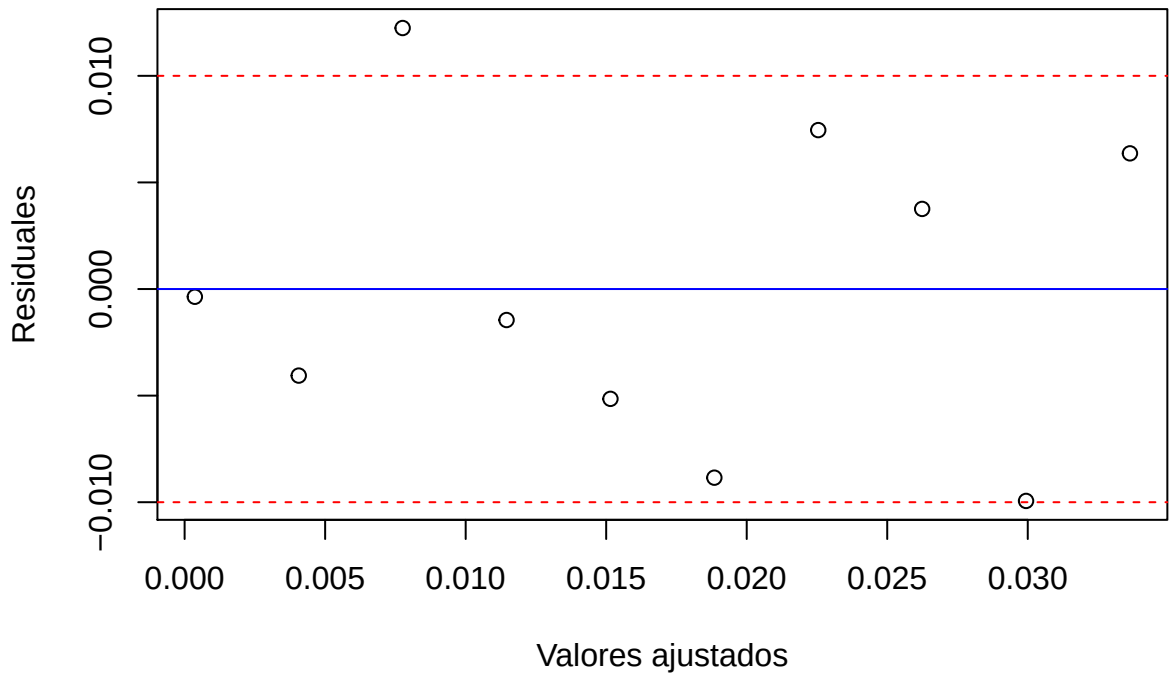
```
shapiro.test(e)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  e
## W = 0.96365, p-value = 0.8265
```

Como el $p - value = 0.8265$ es un valor grande es decir que la probabilidad de que se cumpla H_0 es alta, por lo tanto esto hace que sea válido suponer que los residuales son normales.

Supuesto de varianza constante (forma gráfica). Visualizamos el gráfico.

```r
# Datos ajustados.
aj <- reg$fitted.values
# Grafico de ajustados vs residuales.
plot(aj, e, xlab = "Valores ajustados", ylab = "Residuales")
abline(0, 0, col = "blue")
abline(-0.01, 0, col = "red", lty = 2)
abline(0.01, 0, col = "red", lty = 2)
```



Se cumple el supuesto de varianza constante ya que hay puntos dispersos de forma aleatoria dentro de una banda centrada en cero y con bandas colocadas en -0.01 y 0.01.

Supuesto de varianza constante (por medio de pruebas). Hipótesis:

$$H_0: \text{ La varianza de los errores es constante}$$
$$H_1: \text{ La varianza de los errores no es constante}$$

Prueba de Breusch-Pagan

```r
bptest(y~x)
```

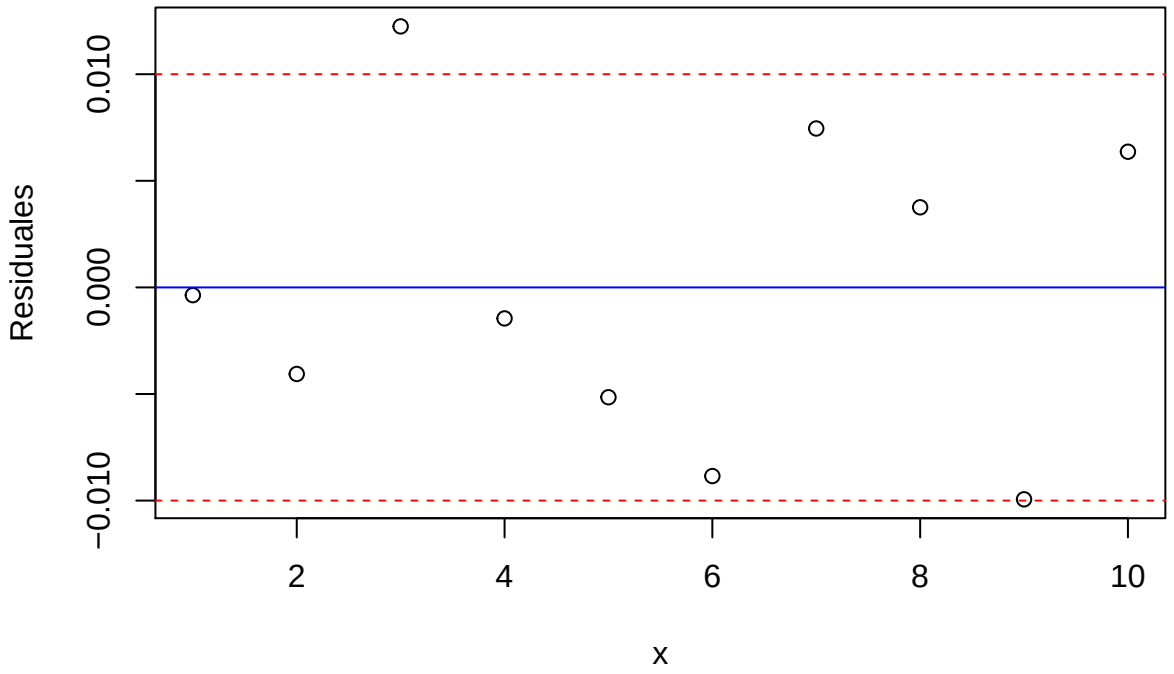
```
##
##  studentized Breusch-Pagan test
##
## data:  y ~ x
## BP = 0.32266, df = 1, p-value = 0.57
```

Como el $p - value = 0.57$ lo cual es un valor relativamente grande es decir que la probabilidad de que se cumpla H_0 es relativamente alta, por lo tanto esto hace que sea válido suponer que la varianza de los errores es constante.

Supuesto de independencia (forma gráfica). Visualizamos el gráfico.

```
e <- reg$residuals

plot(x, e, xlab = "x", ylab = "Residuales")
abline(0, 0, col = "blue")
abline(-0.01, 0, col = "red", lty = 2)
abline(0.01, 0, col = "red", lty = 2)
```



Se cumple el supuesto de independencia para las observaciones ya que hay puntos dispersos de forma aleatoria dentro de una banda centrada en cero y con bandas colocadas en -0.01 y 0.01.

Supuesto de independencia (por medio de pruebas). Hipótesis:

$$H_0: \text{ Existe independencia en las observaciones}$$
$$H_1: \text{ No existe independencia en las observaciones}$$

Prueba de Durbin-Watson.

```
dwtest(y~x)
```

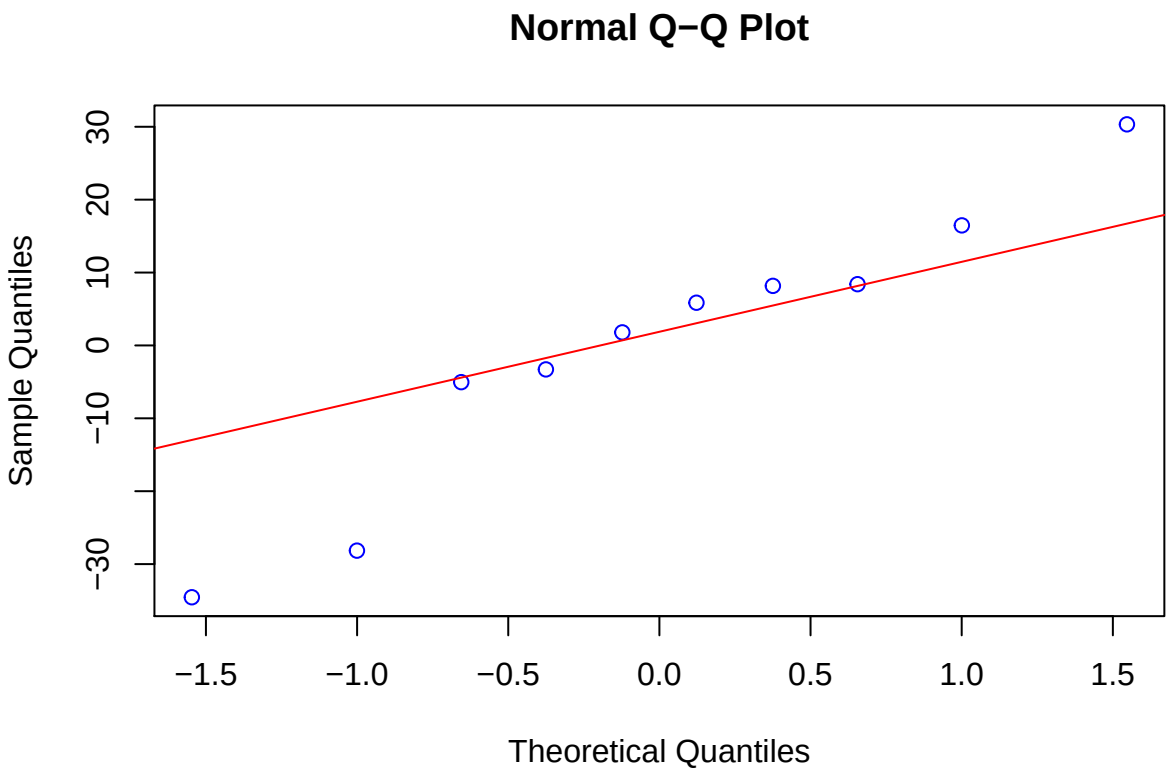
```
##
##  Durbin-Watson test
##
## data:  y ~ x
## DW = 2.5439, p-value = 0.6975
## alternative hypothesis: true autocorrelation is greater than 0
```

Como $p - value = 0.6975$ lo cual es un valor relativamente grande es decir que la probabilidad de que se cumpla H_0 es relativamente alta, por lo tanto esto hace que sea válido suponer que existe independencia en las observaciones.

Supuestos para ejercicio 4.

Supuesto de normalidad (forma gráfica). Visualizamos el gráfico.

```
# Se ingresan los datos.
x <- c(182, 232, 191, 200, 148, 249, 276, 213, 241, 480)
y <- c(198, 210, 194, 220, 138, 220, 219, 161, 210, 313)
# Modelo de regresión lineal en R.
reg <- lm(y~x)
# Residuales.
e <- reg$residuals
# Gráfico.
qqnorm(e, col = "blue")
qqline(e, col = "red")
```



Como se puede observar casi todos los puntos se encuentran cerca de la linea roja, por ello se esperaría que se cumpla el supuesto de normalidad.

Supuesto de normalidad (por medio de pruebas). Hipotesis:

$$H_0 : \text{Los residuales son normales}$$

$$H_1 : \text{Los residuales no son normales}$$

Prueba de Shapiro-Wilk.

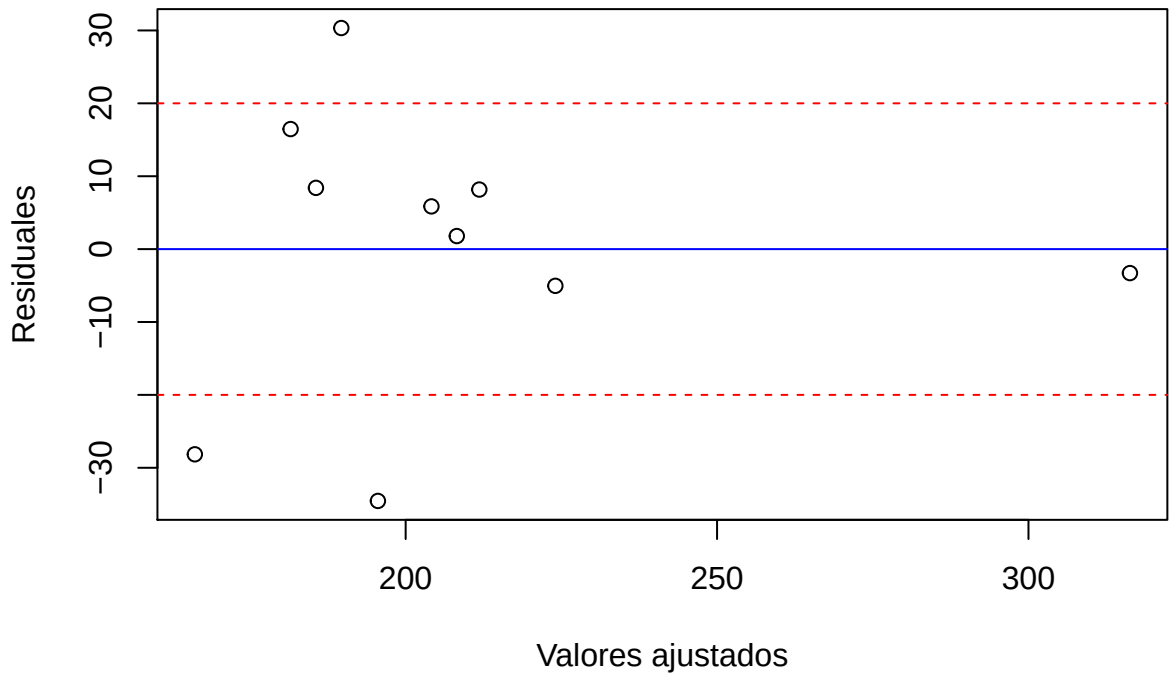
```
shapiro.test(e)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  e
## W = 0.93719, p-value = 0.5222
```

Como $p - value = 0.5222$ no se puede rechazar H_0 por lo que se asume que los datos son normales.

Supuesto de varianza constante (forma gráfica). Visualizamos el gráfico.

```r
# Datos ajustados.
aj <- reg$fitted.values
# Grafico de ajustados vs residuales.
plot(aj, e, xlab = "Valores ajustados", ylab = "Residuales")
abline(0, 0, col = "blue")
abline(-20, 0, col = "red", lty = 2)
abline(20, 0, col = "red", lty = 2)
```



Se cumple el supuesto de varianza constante ya que gran parte de los puntos estan dispersos de forma aleatoria dentro de una banda centrada en cero y con bandas colocadas en -20 y 0.20.

Supuesto de varianza constante (por medio de pruebas). Hipótesis:

$$H_0: \text{ La varianza de los errores es constante}$$
$$H_1: \text{ La varianza de los errores no es constante}$$

Prueba de Breusch-Pagan

```r
bptest(y~x)
```

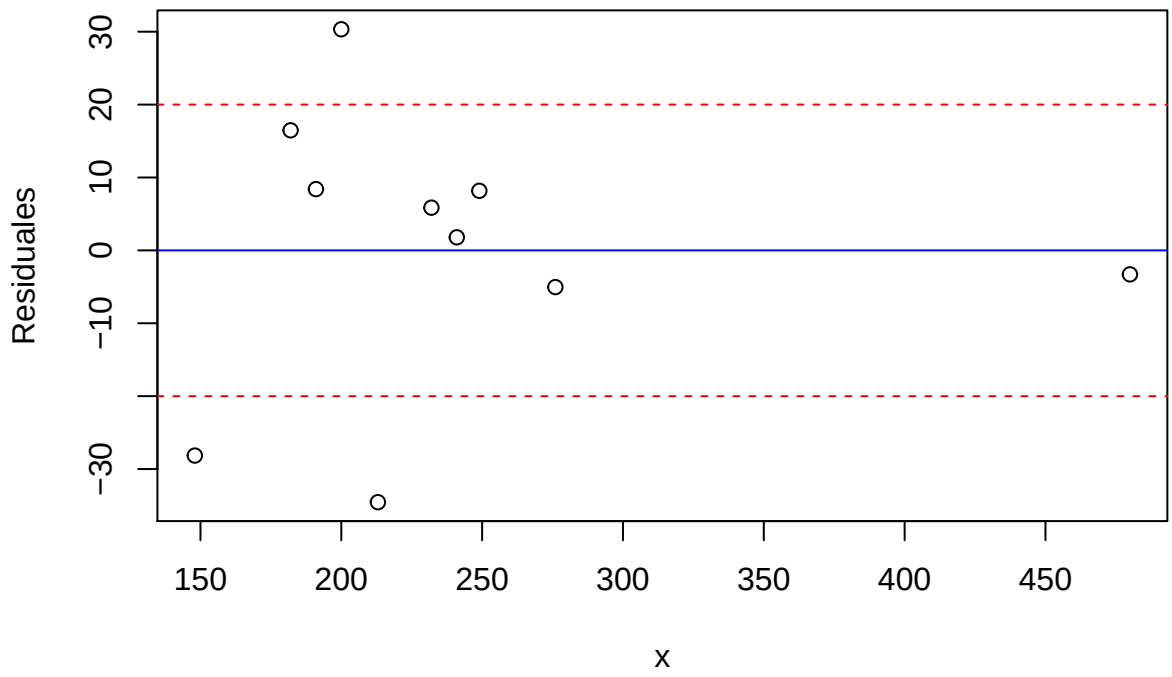
```
##
##  studentized Breusch-Pagan test
##
## data:  y ~ x
## BP = 1.8764, df = 1, p-value = 0.1707
```

Como $p - value = 0.1707$ lo cual no es significativamente pequeño entonces no se puede rechazar H_0 por lo que se asume que se cumple el supuesto de varianza constante. Aunque con un nivel se significancia minimo de $\alpha = 0.18$ (82% de confianza máximo) se podria rechazar H_0 concluyendo que la varianza de los errores no es constante.

Supuesto de independencia (forma gráfica). Visualizamos el gráfico.

```
e <- reg$residuals

plot(x, e, xlab = "x", ylab = "Residuales")
abline(0, 0, col = "blue")
abline(-20, 0, col = "red", lty = 2)
abline(20, 0, col = "red", lty = 2)
```



Se cumple el supuesto de independencia para las observaciones ya que hay puntos dispersos de forma aleatoria y gran parte de estos estan dentro de una banda centrada en cero y con bandas colocadas en -20 y 20.

Supuesto de independencia (por medio de pruebas). Hipótesis:

$$H_0: \text{ Existe independencia en las observaciones}$$

$$H_1: \text{ No existe independencia en las observaciones}$$

Prueba de Durbin-Watson.

```
dwtest(y~x)
```

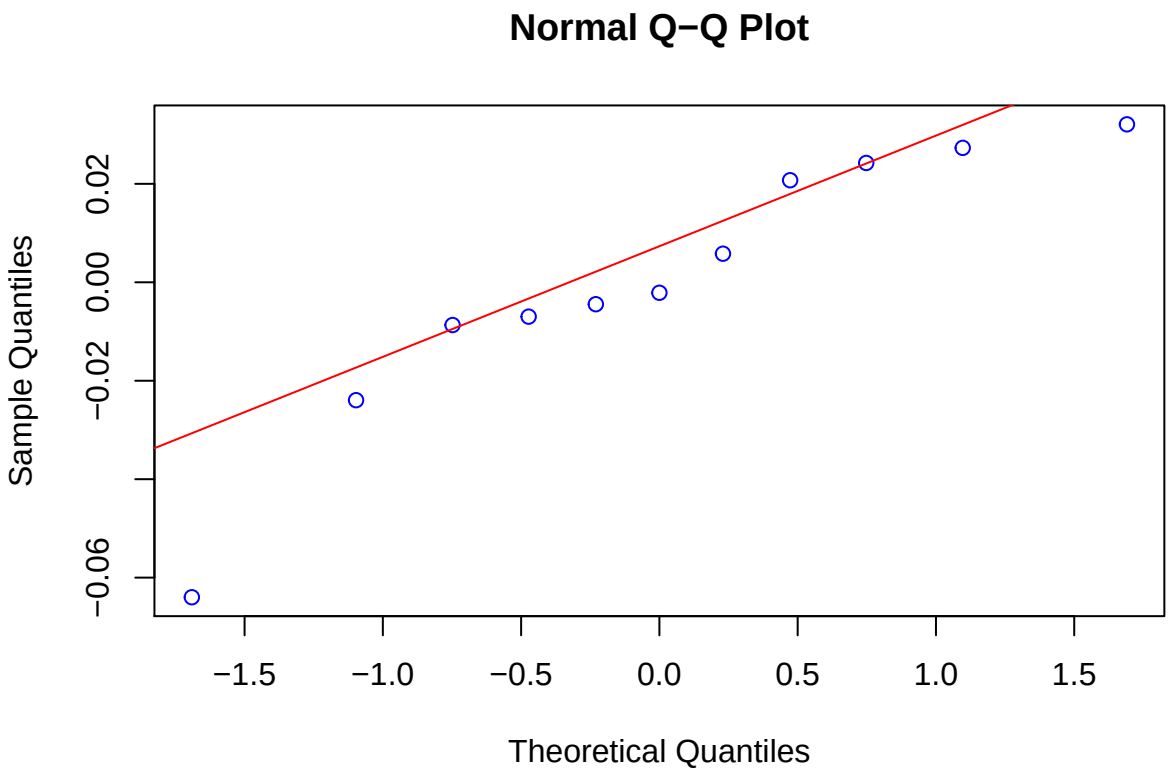
```
##
##  Durbin-Watson test
##
## data:  y ~ x
## DW = 2.2817, p-value = 0.6046
## alternative hypothesis: true autocorrelation is greater than 0
```

Como $p - value = 0.6046$ no se puede rechazar H_0 por lo que se asume que se cumple el supuesto de independencia.

Supuestos para ejercicio 5.

Supuesto de normalidad (forma gráfica). Visualizamos el gráfico.

```
# Se ingresan los datos.
x <- c(3, 5, 10, 15, 20, 30, 40, 50, 60, 75, 90)
y <- c(25.5, 23.4, 18.2, 14.2, 11, 6.7, 4.1, 2.5, 1.5, 0.7, 0.3)
y <- log(y)
# Modelo de regresión lineal en R.
reg <- lm(y~x)
# Residuales.
e <- reg$residuals
# Gráfico.
qqnorm(e, col = "blue")
qqline(e, col = "red")
```



Como se puede observar gran parte de los puntos se encuentran cerca de la linea roja a excepción de dos, por ello se esperaría que se cumpla el supuesto de normalidad.

Supuesto de normalidad (por medio de pruebas). Hipotesis:

$$H_0 : \text{Los residuales son normales}$$

$$H_1 : \text{Los residuales no son normales}$$

Prueba de Shapiro-Wilk.

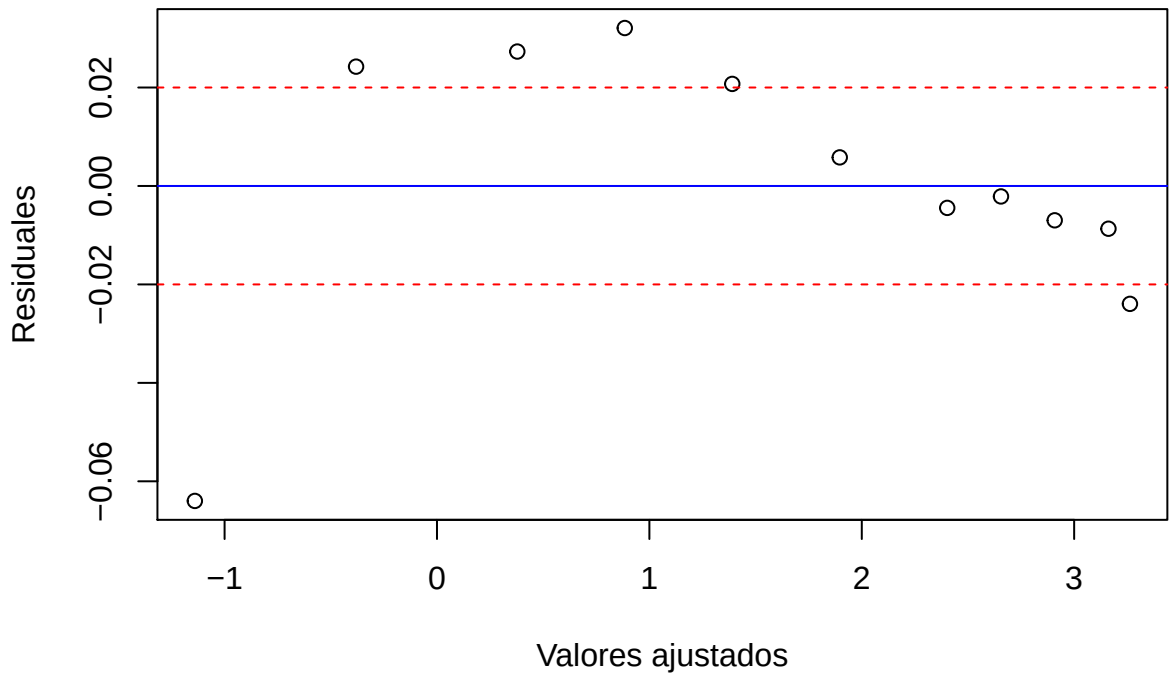
```
shapiro.test(e)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  e
## W = 0.89971, p-value = 0.1833
```

Como $p - value = 0.1833$ lo cual no es significativamente pequeño entonces no se puede rechazar H_0 por lo que se asume que se cumple el supuesto de normalidad. Aunque con un nivel se significancia minimo de $\alpha = 0.19$ (81% de confianza máximo) se podria rechazar H_0 concluyendo que los residuales no son normales.

Supuesto de varianza constante (forma gráfica). Visualizamos el gráfico.

```r
# Datos ajustados.
aj <- reg$fitted.values
# Grafico de ajustados vs residuales.
plot(aj, e, xlab = "Valores ajustados", ylab = "Residuales")
abline(0, 0, col = "blue")
abline(-0.02, 0, col = "red", lty = 2)
abline(0.02, 0, col = "red", lty = 2)
```



No se cumple el supuesto de varianza constante ya que se nota en el gráfico anterior que los datos no estan dispersos aleatoriamente sino que siguen un patrón.

Supuesto de varianza constante (por medio de pruebas). Hipótesis:

$$H_0: \text{ La varianza de los errores es constante}$$
$$H_1: \text{ La varianza de los errores no es constante}$$

Prueba de Breusch-Pagan

```r
bptest(y~x)
```

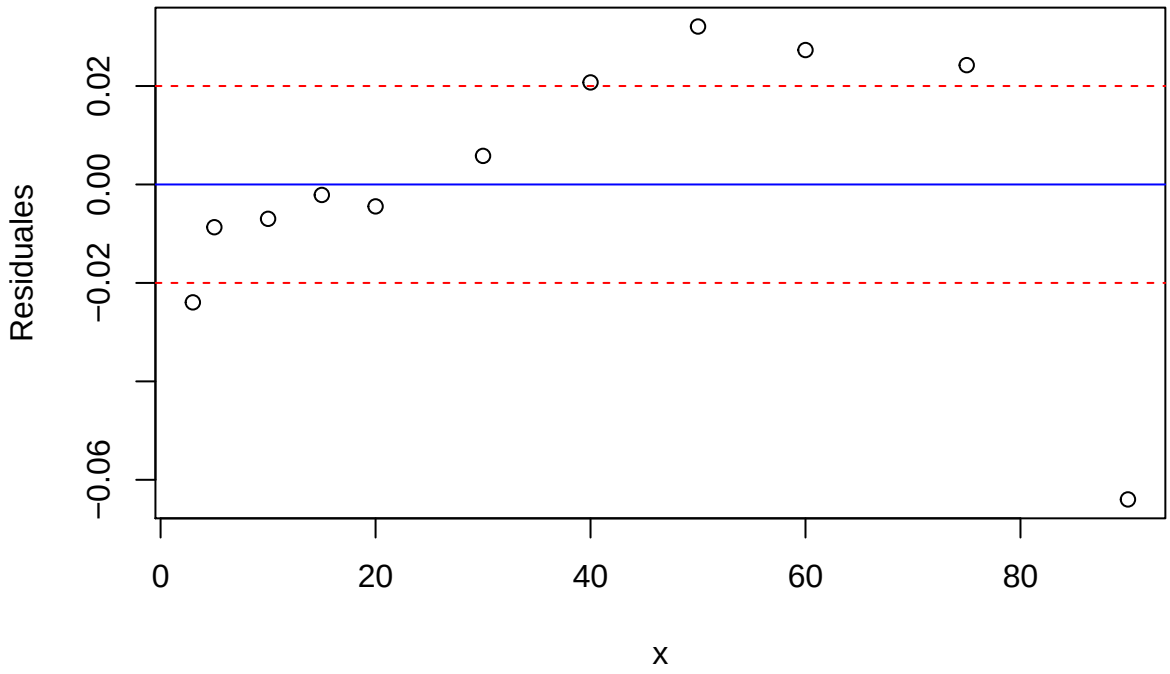
```
##
##  studentized Breusch-Pagan test
##
## data:  y ~ x
## BP = 5.8408, df = 1, p-value = 0.01566
```

Como $p - value = 0.01566$ significa que la probabilidad de que se cumpla H_0 es muy pequeña por lo que se rechaza, y se concluye que la varianza de los errores no es constante.

Supuesto de independencia (forma gráfica). Visualizamos el gráfico.

```
e <- reg$residuals

plot(x, e, xlab = "x", ylab = "Residuales")
abline(0, 0, col = "blue")
abline(-0.02, 0, col = "red", lty = 2)
abline(0.02, 0, col = "red", lty = 2)
```



No se cumple el supuesto de independencia ya que se nota en el gráfico anterior que los datos no estan dispersos aleatoriamente sino que siguen un patrón.

Supuesto de independencia (por medio de pruebas). Hipótesis:

$$H_0: \text{ Existe independencia en las observaciones}$$
$$H_1: \text{ No existe independencia en las observaciones}$$

Prueba de Durbin-Watson.

```
dwtest(y~x)
```

```
##
##  Durbin-Watson test
##
## data:  y ~ x
## DW = 1.1171, p-value = 0.01934
## alternative hypothesis: true autocorrelation is greater than 0
```

Como $p - value = 0.01934$ significa que la probabilidad de que se cumpla H_0 es muy pequeña por lo que se rechaza, y se concluye que no existe independencia en las observaciones.